Examination of Risk Factors for Stroke Survival in the presence of Missing Data and Non-Proportional Hazards.

Anna France, MSci (Hons.)

Department of Mathematics and Statistics

Lancaster University

This thesis is submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

December 2019

Abstract

We are provided with stroke audit data containing patient baseline measures and 5-year follow-up data for patients admitted to two Liverpool based hospitals with acute stroke between January and June 1996. Motivated by this data, we overview previous research on risk factors for survival post-stroke, and review methods for survival analysis and handling missing data.

Multiple imputation accounts for the additional uncertainty when handling missing data, however following analysis of multiple imputed data, assessment of model fit is complicated. We derive and justify formal and visual assessment techniques for the proportional hazards assumption of a Cox regression model fitted to multiply imputed data.

Multiple imputation using chained equations is a flexible approach for handling missing data, however misspecification of imputation model form can lead to biased and restricted analyses. There is minimal research on handling non-proportional hazards within an imputation framework. We derive suitable imputation model forms to incorporate survival outcomes appropriately in the presence of non-proportional hazards, and ensure approximate compatibility with the analysis model.

On correcting analyses to account for non-proportional hazards, model fit is rarely re-assessed in practice, with standard techniques inappropriate for non-standard models. We develop formal and visual assessment techniques of the proportional hazards assumption for a survival model with a time-split, extending the work of Grambsch and Therneau (1994) and Winnett and Sasieni (2001).

Finally, we illustrate the methodological developments achieved within this thesis through application to the stroke audit data. Our analyses identify important risk factors for time to death following stroke, aiding in identification of stroke patients most at risk of death, both in the acute phase and long-term.

Acknowledgements

Here I express my gratitude to everyone who has supported me along this journey. First and foremost, thanks go to my supervisors, Dr. Deborah Costain and Dr. Andrew Titman. Your expertise has been invaluable and I'm incredibly grateful for all the support, encouragement and patience you have shown me throughout. Extra thanks goes to Debbie for bringing this project to my attention and encouraging me to apply. My sincere gratitude goes to the Economic and Social Research Council for their financial support of this project. To all my friends, thank you for sticking by me whilst I've been in PhD mode, and continually putting up with my estrangement, but always being there too; especially Kara. Thank you to all my friends and colleagues in B18, in particular, Laura, Jessica, Abbie and Callum. You have made this journey bearable in the hard times, and given me so many good memories along the way. To my family, your love and support has kept me going. Mum, thank you so much for always pushing me, and also looking after me, particularly when the dreaded tonsillitis struck. To my dad, you may no longer be with us, but the belief you had in me has continually inspired me to achieve my potential, and for that I am eternally grateful. I dedicate this thesis to you. Finally, to my partner and rock, Ben Norwood, I can't put into words how grateful I am to have had you by my side throughout this journey. You continually challenge me to be the best I can be and lift me up when I really need it. You have helped me so much along the way, and I could not have done this without your love and support. I'm so glad we got to complete our work together.

Declaration

I declare that this thesis is my own work, and has not been submitted by myself in substantially the same form for the award of a higher degree elsewhere.

Anna France

Contents

List of Figures ix

List of Tables xii

1 Introduction 1

1.1 Thesis Aims 4

1.2 Thesis Structure 4

2 Introduction to Stroke 7

2.1 Introduction 7

2.2 Introduction to Stroke 7

2.3 Overview of Previous Stroke Research 9

2.4 Data Overview 13

2.4.1 Baseline Measures 14

2.5 Conclusion 18

3 Preliminaries **20**

3.1 Introduction 20

3.2 Survival Analysis 21

3.2.1 Background 21

3.2.2 Notation and Functions of Interest 22

3.2.3 Exploration Techniques 23

3.2.4 Cox Proportional Hazards Model 27

3.2.5 Model Selection 31

3.2.6 Model Validation 33

3.3 Missing Data 40

3.3.1 Classification of Missing Data 40

3.3.2 Methods for Handling Missing Data 44

3.3.3 Multiple Imputation Using Chained Equations 52

3.3.4 Model Selection after Multiple Imputation 60

3.4 Conclusion 62

4 Application to Stroke: Part 1 63

4.1 Introduction 63

4.2 Exploration of Data 64

4.3 Exploration of Missing Data 92

4.4 Multiple Imputation 109

4.4.1 Imputation Procedure 109

4.4.2 Imputation Diagnostics 112

4.5 Model Building 119

4.6 Model Validation 125

4.7 Conclusion 129

5 Model Validation after Multiple Imputation 132

5.1 Introduction 132

5.2 Review of Methods to Assess the Proportional Hazards Assumption 133

5.3 Assessing Proportional Hazards Assumption in Multiply Imputed Data Setting 139

5.3.1 Formal Test 140

5.3.2 Visual Examination 147

5.4 Simulation Study 151

5.4.1 Type I Error Results 153

5.5 Application to Stroke Data 155

5.6 Conclusion 163

6 Handling Non-Proportional Hazards in MICE 165

6.1 Introduction 165

6.2 Review of Current Methods 166

6.3 Imputation Model Theory 170

6.3.1 Background 170

6.3.2 The Piecewise-Proportional Hazards Model 171

6.3.3 Imputation of Binary Variables 173

6.3.4 Imputation of Categorical Variables 179

6.3.5 Imputation of Continuous Variables 185

6.3.6 Further Considerations 191

6.4 Simulation Study 192

6.4.1 Bias Results 194

6.5 Conclusion 196

7 Piecewise-Proportional Hazards Model Validation 198

7.1 Introduction 198

7.2 Piecewise-Proportional Hazards Model 199

7.3 Overview and Issues of Current Methods 201

7.4 Proposed Test of Proportional Hazards Assumption 205

7.4.1 Scaling Factor for Schoenfeld Residuals 206

7.4.2 Visualisation 211

7.4.3 Consideration of Number of Changepoints 213

7.5 Extension to Multiply Imputed Data 214

7.6 Simulation Study 218

7.6.1 Type 1 Error Results 219

7.7 Conclusion 220

8 Application to Stroke: Part 2 222

8.1 Introduction . . . 222

8.2 Summary of Application to Stroke: Part 1 . . . 223

8.3 Handling issues found during model validation . . . 223

8.3.1 Interaction between hospital and no CT scan . . . 224

8.3.2 Functional form of non-linear covariate effects . . . 224

8.3.3 Change point of time-dependent covariate effects . . . 225

8.4 Multiple Imputation . . . 227

8.4.1 Imputation Procedure . . . 227

8.4.2 Imputation Diagnostics . . . 230

8.5 Model Building . . . 235

8.6 Model Validation . . . 241

8.7 Conclusion . . . 249

9 Conclusion 251

9.1 Further Work . . . 253

List of Figures

3.1 Multiple Imputation Process 53

4.1 Overall Kaplan-Meier Survival Curve 64

4.2 Kaplan-Meier Survival Curves by Patient Characteristics 75

4.3 Kaplan-Meier Survival Curves by Patient Medical History 76

4.4 Kaplan-Meier Survival Curves by Admission Details 77

4.5 Kaplan-Meier Survival Curves by Stroke Symptoms & Classification 78

4.6 Baseline Measure Association Tests 82

4.7 Aggregation Plot of Overall Missing Data Patterns 96

4.8 Aggregation Plot of Patient Characteristics Missingness Patterns . . 97

4.9 Aggregation Plot of Facial & Limb Weakness Missingness Patterns . 98

4.10 Aggregation Plot of Patient Medical History Missingness Patterns . 99

4.11 Aggregation Plot of Stroke Symptoms Missingness Patterns 100

4.12 Matrix Plot sorted by Lesion Type in CT Scan 102

4.13 Matrix Plot sorted by Hospital 104

4.14 Matrix Plot sorted by Worst Consciousness Level 104

4.15 Spine Plots showing Missing and Available Data 106

4.16 Spine Plots showing Missing and Available Data 107

4.17 Strip Plot for Systolic BP 112

4.18 Strip Plot for Diastolic BP 113

4.19 Alcohol Consumption Imputation Histogram 115

4.20 Diabetes Mellitus Imputation Histogram 115

4.21 Facial Weakness Imputation Histogram 116

4.22 Pre-Stroke Rankin Imputation Histogram 116

4.23 Histograms of Observed & Imputed Data for Pre-Stroke Rankin . . 117

4.24 Trace Plots for Stroke Event Assessments 118

4.25 Trace Plots for Facial and Limb Weakness 119

4.26 Hazard Ratio Visualisation 122

4.27 Visualisation of Functional Form of Systolic BP 129

5.1 Schoenfeld Residual Plots for Age, Diabetes & Pre-stroke Mobility . 158

5.2 Schoenfeld Residual Plots for Systolic BP & Lesion Side 159

5.3 Schoenfeld Residual Plots for Hospital & Worst Consciousness Level 161

5.4 Schoenfeld Residual Plots for Lesion Type in CT Scan 162

7.1 Example of Schoenfeld Residual Plot for Piecewise-Constant Covariate Effect 205

7.2 Example of Schoenfeld Residual Plots for Piecewise-PH Model . . . 206

8.1 Spline and Quadratic Fit of BP 226

8.2 AIC Results for Changepoint Identification 227

8.3 Strip Plot for Systolic BP 231

8.4 Strip Plot for Diastolic BP 232

8.5 Histograms of Imputed Values of Pre-stroke Mobility 233

8.6 Histograms of Imputed Values, Imputed Data and Observed Data for Pre-stroke Mobility 234

8.7 Trace Plots for Patient Characteristics 234

8.8 Trace Plots for BP Measurements 235

8.9 Visualisation of Hazard Ratios and Confidence Intervals 240

8.10 Schoenfeld Residual Plots for Hospital, Diabetes and Systolic BP . 242

8.11 Schoenfeld Residual Plots for Age 243

8.12 Schoenfeld Residual Plots for Pre-stroke Mobility 244

8.13 Schoenfeld Residual Plots for Lesion Type in CT Scan 245

8.14 Schoenfeld Residual Plots for No CT Scan 246

8.15 Schoenfeld Residual Plots for Worst Consciousness Level 247

List of Tables

2.1 Modified Rankin Scale 18

2.2 Oxford Community Stroke Project Classification 19

4.1 Patient Characteristics Frequency Table 66

4.2 Patient Characteristics Frequency Table 67

4.3 Patient Medical History Frequency Table 68

4.4 Patient Medications Frequency Table 70

4.5 Patient Admission Details Frequency Table 71

4.6 Stroke Classification Frequency Table 72

4.7 Stroke Symptoms Frequency Table 73

4.8 Stroke Symptoms Frequency Table 74

4.9 Log-Rank Test Results of Patient Characteristics & Medical History 79

4.10 Log-Rank Test Results of Stroke Event Assessments 80

4.11 Age and Sex adjusted Cox Models for Patient Characteristics 87

4.12 Age and Sex adjusted Cox Models for Patient Medical History . . . 88

4.13 Age and Sex adjusted Cox Models for Admission Details 89

4.14 Age and Sex adjusted Cox Models for Stroke Symptoms 90

4.15 Age and Sex adjusted Cox Models for Stroke Classification 91

4.16 Missing Data Frequency Table 94

4.17 Pooled Cox PH Model Results 121

4.18 Assessment of Proportional Hazards Assumption 127

4.19 No CT scan frequency by hospital and consciousness 128

4.20 CT Scan: Lesion Type Frequency Table 128

4.21 Table of p-values for PH test for all data sets. 131

5.1 Simulation Results: Type 1 Error of PH Tests 154

5.2 Assessment of Proportional Hazards Assumption 156

6.1 Simulation Results: Bias of Imputation Approaches 195

6.2 Simulation Results: Percentage Bias of Imputation Approaches . . . 197

7.1 Simulation Results: Type 1 Error of PH Tests 220

8.1 CT Scan: Lesion Type Cox model with hospital interaction on 'no scan' 225

8.2 Pooled Piecewise-PH Model Results 237

8.3 Pooled Piecewise-PH Model Results 239

8.4 Assessment of Proportional Hazards Assumption 248

Chapter 1

Introduction

Stroke is a serious and complex condition, for which many patient and medical characteristics can affect post-stroke survival. Stroke is a prominent cause of death worldwide and in the UK, and thus it is important to understand the differences between stroke patients in terms of survival, and the risk factors that affect time to death post-stroke.

Examining factors related to both early and late death times, and understanding when a patient is at a higher risk of death, can enable appropriate interventions to be put in place. Knowledge of when a patient is at a higher risk of death, and the reasons why this is the case, is key to help identify appropriate interventions. Such interventions may take the form of treatment or end of life care, where Kendall et al. (2018) highlight the importance of palliative, or end of life care as an intervention.

There has been a large amount of research into risk factors for survival post-stroke, however, given the complexities of stroke, and difficulties in collecting observational data, inappropriate treatment of missing data has potentially compromised the reliability of previous research relating to survival post-stroke. Many studies which have researched survival post-stroke have focussed upon a single, or small subset, of potential risk factors, due to large proportions of missing information, and further, those which attempted to handle missing data have used naïve approaches (Crichton et al., 2016).

Additionally to missing data complicating survival analysis of stroke patient data, it is known that risk factors for survival post-stroke can have different effects on risk of death dependent upon the time post-stroke, where early and long term survival have previously been considered separately (Petty et al., 2000). The issues around missing data and time-dependent effects on survival post-stroke mean that further research is needed to gain a greater understanding of the risk factors for survival of stroke patients, where reliable inferences have yet to be established.

We aim to examine risk factors for survival post-stroke to establish an important set of risk factors, for both short and long term survival, which can be used together to predict which patients are most at risk of death following stroke. In order to do this, there are several methodological developments needed to allow for the complexities of stroke patient data.

Firstly, appropriate methods need to be developed which can be used to analyse survival data, alongside suitable methods for handling missing data. The survival analysis methods used need to appropriately describe the relationship between potential risk factors and survival, and it is important to handle missing data using appropriate methods in order to avoid introducing bias, whilst also minimising loss of information (Molenberghs and Kenward, 2007).

A flexible approach to missing data is multiple imputation using chained equations (van Buuren, 2012), which involves making multiple predictions of the missing values dependent upon the observed data. However, given the outcome of interest is survival, extra consideration is needed to incorporate this outcome when imputing the missing values (White et al., 2011).

Moreover, the complexity of stroke, and the crucial nature of early treatment following stroke onset, results in risk factors having different effects on survival dependent upon the time post-stroke, where previous research has identified differences in risk factors, and their impact, between early and late survival times following stroke (Andersen and Olsen, 2011). This further complicates analyses for survival post-stroke, and also methods for handling missing data, motivating the methodological developments made within this thesis.

Firstly, there is currently little established methodology for assessing model fit for models fitted to multiply imputed data. In particular, where the Cox proportional hazards model has been fitted to multiply imputed data sets to indicate the hazard of death following stroke for various risk factors, it is important to assess the assumption of proportional hazards, specifically to establish if any of the risk factors have time-dependent effects on survival post-stroke.

Furthermore, as it is already known that risk factors are likely to have time-dependent effects on survival post-stroke, methodological developments are needed to incorporate these effects into a multiple imputation procedure to avoid introducing bias into the imputations, and in turn, the inferences made following analysis. This issue is addressed within this thesis by extending the work of White and Royston (2009) to propose what information should be incorporated into the imputation models in order to accommodate the time-dependent effects, and theoretical rationale will be presented.

As an additional methodological development, this thesis derives techniques for assessment of model fit for the piecewise-proportional hazards model, an extension of the Cox regression model used to incorporate time-dependent effects for survival, where standard validation techniques are inappropriate for this model form.

Returning to the purpose of this research in relation to stroke, the methodological developments made throughout this thesis are applied to a stroke audit data set, where data was collected on patients with acute stroke admitted to two Liverpool based hospitals in 1996, and patients were followed up for five years following stroke onset. The analyses are interpreted in the context of stroke, identifying which patients are most at risk of death following stroke.

With regards to the aforementioned points, the subsequent sections of this chapter identify the aims of this research and outline the structure of the remainder of this thesis.

1.1 Thesis Aims

In terms of stroke, the aim of the thesis is to gain a greater understanding of the differences between stroke patients in terms of survival, through producing a good survival analysis of the stroke audit data, to identify important risk factors for time to death post-stroke, and produce a fully adjusted analysis of these risk factors for early and long-term survival.

In order to produce a good survival analysis of the stroke audit data, the methodological aims of the thesis are as follows:

- To develop methodology for the multiple imputation of survival data in the presence of non-proportional hazards;
- To develop methodology for assessing the proportional hazards assumption of survival models after multiple imputation;
- To develop methodology for assessing the model fit of a survival model with a time-split, in which piecewise-constant effects account for non-proportional hazards.

1.2 Thesis Structure

Here we provide an overview of the structure of the remainder of this thesis, giving a brief description of the contents of each chapter:

Chapter 2: Introduction to Stroke. This chapter gives the definition of stroke and provides a review of the literature relating to survival post-stroke. Further, this chapter gives an overview of the stroke audit data used within this thesis.

Chapter 3: Preliminaries. This chapter will overview methods for the analyses of time-to-event data and handling missing data. Survival analysis will be introduced, where methods for data exploration, fitting a Cox proportional hazards model, and model validation will be outlined. Further, missing data will be defined, highlighting issues it can cause and reviewing methods for handling missing data, with multiple imputation using chained equations discussed in depth.

Chapter 4: Application to Stroke: Part 1. Here the initial analyses will be presented. The results of the data exploration will be shown, alongside the imputation process and initial modelling results. Issues around model validation will be discussed as motivation for the next 3 chapters.

Chapter 5: Model Validation after Multiple Imputation. This chapter will give an overview of methods used for model validation of the Cox proportional hazards model, in particular the assessment of the proportional hazards assumption. Extensions of these methods will be outlined for application to multiply imputed data, with a simulation study and application of these extensions presented.

Chapter 6: Handling Non-Proportional Hazards in MICE. Motivated by the model validation results in Chapters 4 and 5, this chapter will present theoretical developments for handling non-proportional hazards within the MICE framework, with a piecewise-proportional hazards model used as the analysis model to incorporate the time-dependent covariate effects.

Chapter 7: Piecewise-Proportional Hazards Model Validation. Motivated by Chapters 4, 5 and 6, this chapter presents the development of model validation techniques for the piecewise-proportional hazards model, and provides extensions for application to multiply imputed data.

Chapter 8: Application to Stroke: Part 2. Extending upon Chapter 4, here the methods outlined in Chapters 6 and 7 are applied to the stroke audit data to produce a parsimonious model for the survival of stroke patients, interpreted in the context of stroke survival.

Chapter 9: Conclusion. This concluding chapter provides a summary of the thesis, with the potential for future work identified, alongside further discussion and conclusions about the findings in this thesis.

Chapter 2

Introduction to Stroke

2.1 Introduction

This chapter provides a review of previous research around survival post-stroke and describes the data. Firstly stroke is introduced, giving the definition of stroke and outlining the national and global burden caused by the condition.

Previous research on survival post-stroke is then discussed, giving details on the findings of previous studies and highlighting their limitations.

Finally this chapter overviews the stroke audit data analysed throughout this thesis, defining each of the baseline covariates included as potential risk factors for survival post-stroke.

2.2 Introduction to Stroke

A stroke is a serious, life-threatening medical condition that occurs when the blood supply to part of the brain is interrupted by either a blockage or rupture of an artery to the brain (Johnson et al., 2016). Stroke was the defined by the World Health Organisation (Aho et al., 1980) as a syndrome of vascular origin, characterised by rapidly developing clinical symptoms and a focal loss of cerebral function, in which symptoms last for more than 24 hours or lead to death (Rudd and Wolfe, 2002). Strokes are classed as medical emergencies which require urgent treatment, where early treatment helps to reduce the amount of damage caused by stroke (ISWP, 2012).

To reflect the improved understanding of stroke and its subtypes due to scientific advancements, the American Stroke Association (Sacco et al., 2013) and World Health Organisation (Norrving et al., 2013) provide updated definitions of stroke, defining the different subtypes separately. There are two key types of stroke; ischemic stroke where a blood clot blocks blood flow to part of the brain and haemorrhagic stroke which refers to a burst blood vessel in the brain. These two types can be further categorised into subtypes.

Considering the global burden of stroke, it is the second most common cause of death worldwide (WHO, 2018; Feigin et al., 2017), and was reported by GBD (2015) to be the third leading cause of years life lost in 2013 globally with 6.5 million stroke deaths worldwide (Feigin et al., 2015). Mozaffarian et al. (2016) reported that stroke deaths accounted for 11.8% of total deaths worldwide in 2013, where the prevalence in 2010 was 33 million, with 16.9 million first strokes.

Stroke is also a major cause of morbidity and mortality in the UK, where stroke is reported to be the fourth largest cause of death in the UK (Stroke Association, 2018). Over 100,000 strokes occur in England each year (Lee et al., 2011), and 1 in 8 are fatal within the first 30 days post-stroke (Bray et al., 2016). Andrews et al. (2016) reported that, in 2016, around 38,000 people died of stroke in the UK, however, Goldacre et al. (2008) suggest that stroke is underestimated as a cause of death due to reporting on death certificates.

Stroke is not only a prominent cause of death, but is the largest cause of adult disability in England (NAO, 2010). If a patient survives, the injury caused to the brain through stroke can lead to long-lasting problems for the patients, leaving them needing rehabilitation and long-term care. Around half of surviving stroke patients are left with a significant long term disability (Stroke Association, 2016; Rudd and Wolfe, 2002) and more than half are left dependent on others for everyday activities (Wolfe, 2000). Stroke requires long-term follow up and appropriate care, resulting in costs to the NHS of over £3 billion every year (NAO, 2010), with treatment costs accounting for approximately 5% of the total UK NHS costs (Saka et al., 2009; Mant et al., 2004).

Stroke remains underfunded in terms of research, where in 2012, stroke received 7% of the overall UK research funding for the four major causes of death. This equated to £56 million, the equivalent of 2% of the overall health and social care costs of stroke in 2012 (Luengo-Fernandez et al., 2015).

2.3 Overview of Previous Stroke Research

The impact stroke has on both short-term and long-term survival has been studied previously. A study on early outcomes post-stroke by Nakibuuka et al. (2015) found 6% of patients to have died within a week post-stroke, 27% within a month, and found two thirds of deaths to be within their hospital stay. This study had a small cohort, however, and thus cannot be generalised.

Bray et al. (2016) reported that within the first 30 days post-stroke, 1 in 8 strokes were fatal, and also found that patients admitted to hospital overnight on a weekday had an increased hazard of death compared to other times of admission, noting this was likely due to differences in early care. Parry-Jones et al. (2016) compared differences between haemorrhagic and ischemic stroke, and found that 69.4% of patients with haemorrhagic stroke had died or were moderately-severely disabled at the end of their hospital stay, compared to 45.4% of ischemic stroke patients. These studies were focused upon differences in early care, however, and did not consider long-term survival.

Considering the survival of stroke patients against the general population, Hardie et al. (2003) and Gresham et al. (1998) found stroke sufferers to have an increased hazard of death for up to 20 years following stroke compared to the age and sex adjusted general population. These studies are limited by small cohorts but highlight the long-term effect stroke has on mortality. A more recent study conducted by Crichton et al. (2016), which looked at long-term survival and other health related outcomes post-stroke, found that 80% of patients had died by 15 years and poor outcomes were common among survivors. The focus of these studies was on survival and other outcomes post-stroke, with minimal consideration of the impact of possible prognostic factors on these outcomes and survival.

Previous stroke research has also focused upon the differences between stroke patients in terms of survival, examining the risk factors that affect time to death post-stroke. This is an important area of research for stroke; knowledge of factors related to both early and late death times, and understanding when a patient is at a higher risk of death, can enable appropriate interventions to be put in place. This could be treatment, or, as considered by Kendall et al. (2018), palliative care.

There have been many previous studies which have conducted research to identify the risk factors for death post-stroke. In particular, on conducting a review of studies investigating survival post-stroke over the last 20 years, we found there are many risk factors considered to be related to survival. Risk factors can be categorised as either non-modifiable, which cannot be changed, or modifiable risk factors, which can be controlled.

Winovich et al. (2017) identified two key non-modifiable risk factors; age and sex. These have been identified as risk factors for survival post-stroke by several studies. Khosravi et al. (2017) and Weimar et al. (2002) determine age to be important, suggesting increased age increases risk of death post-stroke. Andersen and Olsen (2011) suggest sex to be an important risk factor, however there is a lack of clarity regarding the direction of this effect, with Olsen et al. (2007) suggesting the effect of sex on stroke-survival differs over time. Further Khosravi et al. (2017) and Di Carlo et al. (2018) established consciousness level post-stroke to be a non-modifiable risk factor for survival, suggesting worse consciousness increased risk of death. Stroke severity was identified as another risk factor by several studies, which suggested increased severity of stroke was associated with poorer outcomes (Andersen and Olsen, 2011; Appelros et al., 2003; Weimar et al., 2002).

Di Carlo et al. (2018) identified stroke subtype to be an important risk factor for survival post-stroke, where stroke subtype can guve an indication of stroke severity. This was reiterated by several other studies including Mudzi et al. (2012) and Petty et al. (2000), however, each of these studies used varying definitions for grouping stroke into subtypes. As a further non-modifiable risk factor, history of heart failure was shown to be associated with increased risk of death post-stroke by Di Carlo et al. (2018). Further, Doehner et al. (2012) identified living arrangements at time of stroke to be a risk factor for survival, where patients living in a institution prior to stroke were shown to be at increased risk of death.

Modifiable risk factors are those which could be changed or controlled for by things such as lifestyle changes or medication. Diabetes is shown to be a modifiable risk factor associated with increased risk of death by Ma et al. (2018), alongside several other studies; Heldner et al. (2018), Khosravi et al. (2017) and Andersen and Olsen (2011). Further, each of these studies also identify smoking status to be a risk factor for survival and suggest smoking increases risk of death post-stroke.

Heldner et al. (2018) and Khosravi et al. (2017) additionally identify hypertension to be a modifiable risk factor associated with an increased risk of death post-stroke. Likewise, Di Carlo et al. (2018) reiterate the importance of hypertension on survival post-stroke, but also further identify atrial fibrillation to be a risk factor, along side Andersen and Olsen (2011) and Appelros et al. (2003). Atrial fibrillation is suggested to increase risk of death post-stroke. Additional modifiable risk factors identified by previous studies are body mass index (Doehner et al., 2012) and depression (Robinson and Jorge, 2015). These suggested that underweight patients were most at risk of poor outcomes, as were patients who were untreated for their post-stroke depression.

The focus of many of these studies has been on a on a single risk factor or a subset previously considered to be important. Ma et al. (2018) focussed on the effect of diabetes on survival post-stroke, whereas studies such as Doehner et al. (2012) and Heldner et al. (2018) focussed on a subset of risk factors previously ascertained to be important for survival. However, as stroke is complex and there are many risk factors considered to influence survival post-stroke, it is important to examine potential risk factors jointly.

Winovich et al. (2017), Khosravi et al. (2017) and Weimar et al. (2002) considered a large set of risk factors and conducted multivariate survival analysis, however these studies encountered issues around missing data. Within these studies, naïve approaches were commonly used to handle missing data, including the exclusion of patients with incomplete observations. This complete-case analysis approach can result in a loss of information and biased inferences, and Crichton et al. (2016) highlighted this as a limitation of their work as it can result in misleading findings (Little and Rubin, 2002).

Mogensen et al. (2013) made an attempt to handle missing data more appropriately by using multiple imputation, however, they did not include information on the survival outcome when imputing the values of the missing observations. Omitting the response in the imputation process can introduce bias into the inferences (Kontopantelis et al., 2017; Moons et al., 2006), where White and Royston (2009) demonstrated bias towards the null when ignoring the survival outcome. Missing data is a common issue for observational studies, particularly for a condition as complex as stroke, however recent studies still fail to mention missing data as a potential limitation (Di Carlo et al., 2018).

A further focus of previous studies has been the difference in risk factors for survival dependent upon time post-stroke (Ma et al., 2018; Collins et al., 2003). Andersen and Olsen (2011) constructed separate stopped Cox models to consider how associations between risk factors and survival differ for 1-month, 1-year, 5-year and 10-year maximum follow-up times. The authors, however, did not perform model selection and hence non-significant effects were included within the inferences.

Petty et al. (2000) also looked at 30-day and 1-year survival separately, indicating differences between risk factors for early and late survival, however this study focussed on stroke subtypes. A change in the effect of age for early and late survival has also been noted by Easton et al. (2014).

In general, previous research has shown that there are many risk factors for survival post-stroke, with potential differences in risk factors dependent upon time post-stroke. However, previous research has often been limited by the issue of missing data, with a lack of studies appropriately considering risk factors jointly.

2.4 Data Overview

The data used throughout this thesis are from an on-going stroke audit programme which began in 1996 at University Hospital Aintree (UHA) and the Royal Liverpool University Hospitals, Broadgreen (RLBUH) in Merseyside, UK. For the audit, all hospitalised patients with acute stroke during the period January - June 1996 were identified prospectively from stroke registers, casualty registers and relevant wards, and retrospectively from hospital discharge lists. The data included 538 patients who were identified as having had an acute stroke and entered onto stroke registers between January 1st and June 11th 1996. The follow-up period was 5 years, recorded between 1996 and 2001.

Within the audit, a minimum data set was recorded according to the European Stroke Database (ESDB), and included information on patient demographics, history of known risk factors for stroke, characteristics of stroke and resource use in hospital. Information on mortality was obtained from hospital information systems and FHSA/GP registers. To ensure the audit covered all patients, not just those who could give consent or communicate, all patients had in-hospital data collected. Provided consent was given, follow-up data was also collected at discharge and at 3, 6, 12 ,24, 36, 48 and 60 months post-stroke to assess general stroke recovery, such as functionality and mood, however this project focuses on the ESDB baseline data in relation to mortality post-stroke. Details of the baseline measures are itemised below in section 2.4.1.

2.4.1 Baseline Measures

The general patient characteristics included are:

- Age - age of the patient in years at time of stroke;
- Sex - male or female;
- Smoking status - smoker, ex-smoker or non-smoker;
- Alcohol consumption - excessive, regular, occasional or non-drinker;
- Pre-stroke mobility - 200 metres outdoors, indoors or needs help;
- Pre-stroke Rankin - participation rated using the modified Rankin scale, this measures independence, incorporating mental and physical adaptations to the neurological defects, see Table 2.1;
- Pre-stroke living conditions - home alone, home with companion or in an institution.

The baseline measures also included details of patients' medical history and previous medications:

- Previous stroke - whether or not the patient had suffered from a stroke previously;

- Previous transient ischaemic attack (TIA) - whether or not the patient previously had a TIA, a temporary disruption to the blood supply to part of the brain;

- Diabetes Mellitus - whether or not the patient has diabetes, a lifelong condition that causes a person's blood sugar level to become too high;

- Hypertension - whether or not the patient has previously been diagnosed with hypertension, or high blood pressure;

- Angina - whether or not the patient has previously suffered from angina, chest pain caused by restricted blood flow to the muscles of the heart;

- Atrial Fibrillation - whether or not the patient has previously suffered from atrial fibrillation, a heart condition that causes an irregular, often abnormally fast, heart rate;

- Peripheral Vascular Disease (PVD) - whether or not the patient has PVD, a common condition in which build up of fatty deposits in the arteries causes restricted blood flow to the leg muscles;

- Myocardial Infarction - whether or not the patient has previously had a myocardial infarction, more commonly known as a heart attack, where the blood supply to the heart is suddenly blocked;

- Previous anti-hypertensive medication - whether or not the patient has previously taken medication to reduce blood pressure;

- Previous anti-platelet treatment - whether or not the patient has previously taken medication to reduce platelet aggregation and prevent formation of blood clots;

- Previous anti-coagulants - whether not the patient has previously taken anti-coagulant medications to prevent blood clots forming.

Details of the stroke event were also recorded, including symptoms shown by the patient at admission to hospital and within the first 24 hours post-stroke, along side other important diagnostic details regarding stroke. The stroke event assessments recorded are:

- Onset date - date of the onset of stroke;
- Admission date - date the patient was admitted to hospital;
- Hospital - specification of whether the patient was admitted to UHA or RLUBH;
- OCSP classification of stroke - classification of stroke: TACS, PACS, LACS, POCS, unconscious or unclassified, where definitions of the stroke subtypes are given in Table 2.2;
- Side of lesion - if the patient had a lesion, this identified the position of the lesion in the brain: left, right, both sides, or no lesion;
- CT scan results - it was recorded whether or not each patient had a CT scan, and if so, what type of lesion was shown in the CT scan: no lesion, cerebral infarction (CI), haemorrhagic cerebral infarction (HCI), or primary intra-cerebral haemorrhage (PICH);

At admission to hospital, information was recorded about the following:

- Blood pressure (BP) - systolic and diastolic BP measurements (mmHg) and whether or not the patient had hypertension;
- Arm weakness - recorded as no movement, weakness or no deficit;
- Leg weakness - recorded as no movement, weakness or no deficit.

Information was also recorded regarding symptoms occurring within the first 24 hours after onset of stroke. These include:

- Worst consciousness level within the first 24 hours after stroke onset - coma, stupor, drowsy or alert;
- Arm weakness - whether or not the patient had arm weakness;
- Leg weakness - whether or not the patient had leg weakness;
- Facial weakness - whether or not the patient experienced facial weakness;
- Dysphasia - whether or not the patient had dysphasia, an impairment in a person's ability to communicate;
- Dysarthria - whether or not the patient had dysarthria, a motor speech disorder;
- Confusion - whether or not the patient experienced confusion;
- Conjugate Gaze Paresis (CGP) - whether or not the patient had CGP, a neurological disorder affecting a person's ability to move both eyes in the same direction;
- Hemianopia - whether or not the patient had hemianopia, decreased vision or blindness in one half of the visual field;
- Sensory inattention - whether or the not the patient had sensory inattention, a sensory deficit causing the inability of a person to process or perceive stimuli on one side of the body or environment;
- Brainstem or cerebellar signs - whether or not the patient was showing signs of brainstem or cerebellar damage;
- Other deficit - whether or not the patient had signs of any other deficit.

The primary outcome measure that was recorded was death. Both date and cause of death were recorded, and survival time was calculated as the number of days from the date of onset of stroke to the date of death.

Table 2.1: Modified Rankin Scale

Level	Description
0	No symptoms
1	No significant disability despite symptoms; able to perform all usual duties or activities
2	Slight disability; unable to carry out all previous activities, but able to look after own affairs without assistance
3	Moderate disability; requires some help but able to walk without assistance
4	Moderately severe disability; unable to walk without assistance and unable to attend to own bodily needs without assistance
5	Severe disability; bedridden, incontinent, and requires constant nursing care and attention
6	Dead

2.5 Conclusion

This chapter has defined stroke and highlighted the national and global burden of the condition. The review of previous research on survival post-stroke has shown that there are many risk factors for survival post-stroke, however, the issues around missing data have led to many studies only considering a small subset of risk factors, or handling missing data using naïve approaches. Further, this review has highlighted the need to consider possible time-dependencies of risk factors for survival post-stroke. The stroke audit data has been outlined in this chapter, and the issue of missing data and time-dependent effects shown in previous studies motivate the methods used and developed throughout this thesis.

Table 2.2: Oxford Community Stroke Project (OCSP) Classification

Type	**Description**
TACS	Total Anterior Circulation Stroke is a large cortical stroke in middle or anterior cerebral artery areas showing all three of: ▶ Motor and/or sensory deficit in at least two of face, arm or leg; ▶ Homonymous hemianopia, a visual field deficit; ▶ Higher cerebral dysfunction e.g. dysphasia.
PACS	Partial Anterior Circulation Stroke is a cortical stroke in middle or anterior cerebral artery areas showing only two of: ▶ Motor and/or sensory deficit in at least two of face, arm or leg; ▶ Homonymous hemianopia, a visual field deficit; ▶ Higher cerebral dysfunction e.g. dysphasia.
POCS	Posterior Circulation Stroke shows one of the following: ▶ Ipsilateral cranial nerve palsy with contralateral long tract signs; ▶ Bilateral motor and/or sensory deficit; ▶ Cerebellar dysfunction; ▶ Disorder of conjugate eye movements; ▶ Isolated hemianopia or cortical blindness.

LACS	Lacunar Stroke is an occlusion of a single deep perforating artery adhering to one of:
	▶ Pure motor stroke;
	▶ Pure sensory stroke;
	▶ Ataxic hemiparesis.

Chapter 3

Preliminaries

3.1 Introduction

This chapter sets the notation and introduces the techniques used throughout this thesis. We overview survival analysis and the issue of missing data, providing a review of the relevant methods for both these topics.

Firstly survival data is introduced, defining important functions and notation, and outlining data exploration techniques. The Cox proportional hazards model is defined, describing in depth how the model can be fitted. Additionally, methods for selecting a parsimonious model are discussed, alongside model validation techniques, where focus is upon the assessment of the proportional hazards assumption.

Further, the issue of missing data is introduced, where classifications of missing data are outlined. The potential issues arising from the presence of missing data are highlighted, alongside a review of methods for handling missing data. Particular focus is given to multiple imputation, where a detailed discussion of how this can be implemented using multiple imputation using chained equations is provided. Techniques for obtaining a pooled parsimonious model following multiple imputation are also outlined.

3.2 Survival Analysis

3.2.1 Background

Survival analysis is used to model and analyse time-to-event data, which consists of measurements of time until the occurrence of a particular event of interest for each individual. In order for comparisons to be made, it is essential to measure the time-to-event from a well-defined origin; in a medical context this could be the time when treatment began or when the disease was diagnosed. Dependent upon the type of the event of interest, time-to-event may be described as 'survival' time, 'failure' time or 'event' time. If death is the event of interest, then time-to-event is referred to as survival time. This is the case in the context of the stroke data, which includes measurements of time to death from onset of stroke for each patient, thus we will use the term 'survival' time to describe time-to-event.

Censoring is a key feature of time-to-event data, and must be considered within any analyses carried out on time-to-event data. Censoring occurs when the event of interest has not been observed for an individual. A particular type of censoring is right censoring, which commonly occurs when individuals are observed up until a maximum time point, and if an individual has not experienced the event of interest prior to this maximum time point, a censored survival time is instead recorded. This is the time the individual was last known to be alive or event free. A right censored survival time may also be recorded due to drop-out or loss to follow-up. It is common to make the assumption that that an individual's actual survival time is independent of the mechanism causing their survival time to be censored. This is known as non-informative censoring, and assumes statistical independence between the censored and actual survival time.

3.2.2 Notation and Functions of Interest

In order to define important functions of interest, denote the actual survival time of an individual to be t, and let T be a non-negative random variable associated with the survival time. Let Δt denote a small time interval. There are three main functions of interest for summarising survival data:

Survivor function: the probability that an individual survives beyond some time t, and is given by

$$S(t) = \mathbb{P}(T \geq t);$$

Hazard function: the risk of the event occurring at time t, given by the probability of an individual having the event at time t, conditional upon that individual surviving to time t. More formally,

$$h(t) = \lim_{\Delta t \to 0^+} \left\{ \frac{\mathbb{P}(t \leq T \leq t + \Delta t | T \geq t)}{\Delta t} \right\};$$

Cumulative hazard function: the risk of an event up to time t, given it has not occurred before time t. This can be obtained from either the hazard function or survivor function.

$$H(t) = \int_0^t h(u)\mathrm{d}u = -\log S(t).$$

These functions can be used to derive the probability density function, $f(t)$, and cumulative incidence function, $F(t)$, where

$$F(t) = \mathbb{P}(T < t) = \int_0^t f(u)\mathrm{d}u.$$

Now, suppose we have a set of n individuals followed up for some time interval $[0, \tau]$, and denote the true survival time for an individual i as t_i, and the right censored time as c_i, for $i = 1, ..., n$. Then the observed survival time for an individual i

is denoted t_i^*, where $t_i^* = \min(t_i, c_i)$ and $0 \leq t_i^* \leq \tau$. Let δ_i define an indicator function, where δ_i is unity if t_i^* is the true survival time t_i, and zero otherwise; this is defined as the censoring indicator. The observed response for an individual i is then represented by the observed survival time and censoring indicator, and is denoted as the iid pair (t_i^*, δ_i). The survivor function and hazard function can be estimated using the observed survival times.

3.2.3 Exploration Techniques

An essential aspect for analyses of observed data is exploration of the data. For survival data, this can require specialised techniques. Within this section we will briefly describe the Kaplan-Meier estimator for the survivor function and how it can be used to explore survival data, alongside the Neslon-Aalen estimator of the survivor function. Additionally we will overview the log-rank test, and some techniques for considering pairwise associations between explanatory variables within the data.

Kaplan-Meier Estimator

The Kaplan-Meier (KM) estimator (Kaplan and Meier, 1958) is useful for looking at overall survival, and comparing survival between groups of individuals. The KM estimator is the product-limit estimator of survival, and adjusts the empirical survivor function to account for the presence of right-censored observations. It is a non-parametric approach so no assumptions about the underlying distribution need to be made, however the estimator assumes independent censoring and considers a discrete version of the hazard function. Consider N distinct ordered event times denoted $0 < t_1 < t_2 < ... < t_N$. For some event time t_j, $j = 1, ..., N$, denote the risk set as $R(t_j)$, where the risk set is defined to be the set of individuals who are still in follow-up and remain event free, in other words, the set of individuals still at risk of the event at time t_j. Let n_j be the number of individuals in the risk set at t_j and d_j be the number of events that occur at t_j. An individual being event free just before t_j is equivalent to the individual being event free beyond t_{j-1}. This equivalence means the probability of the event at t_j can be defined as a discrete version of the hazard function

$$h(t_j) = \mathbb{P}(T = t_j | T > t_{j-1}),$$

which can be estimated by the proportion of individuals at risk at t_j,

$$\hat{h}(t_j) = \frac{d_j}{n_j}.$$

The probability of being event free up to time point t_j is given by taking the product of the probability of an individual being event free beyond t_{j-1} and the conditional probability of being event free up to t_j given event free at t_{j-1}. Algebraically, this is

$$\hat{S}(t_j) = \hat{S}(t_{j-1})(1 - \hat{h}(t_j)) = \hat{S}(t_{j-1})(1 - \frac{d_j}{n_j}).$$

Through repeated application of this product rule, the KM estimator is obtained as

$$\hat{S}(t) = \prod_{j:t_j \leq t} \left(1 - \frac{d_j}{n_j}\right)$$

The KM estimator, $\hat{S}(t)$, is a step-wise function that approaches a continuous distribution as the number of individuals, n, increases. Plotting the KM survival curves is a commonly used approach to establish any changes in the estimated survivor function as time progresses, and can be used to view differences in survival across groups of individuals.

Log-Rank Test

The log-rank test can be used to assess the difference in survival between two or more groups. The log-rank test is a non-parametric approach which tests the null hypothesis that there no is difference in survival between groups, or more precisely, levels of a particular covariate, where the null hypothesis is tested against the χ^2-distribution. As an example, the log-rank test could be used to assess differences in survival by sex, and assess the null hypothesis that there is no difference in survival between males and females.

In order to calculate the log-rank test statistic, the observed and expected number of events for each group need to be defined. Let L be the number of levels to be compared, and $l = 1, ..., L$. For the jth event time, t_j, denote the number of events at time t_j as d_j, or d_{lj} for level l, and the let n_j denote the number of patients at risk at t_j, n_{lj} for level l. Under the assumption that the null hypothesis is true and there is no difference in survival between the levels, the expected number of events in each group is given by

$$E_{lj} = \frac{d_j n_{lj}}{n_j},$$

and the observed number of events, O_{lj}, is equal to the number of events at t_j, that is

$$O_{lj} = d_{lj}.$$

The log-rank test statistic, $\mathcal{L}$, is calculated as

$$\mathcal{L} = \sum_{l=1}^{L} \frac{(\sum_j O_{lj} - \sum_j E_{lj})^2}{\sum_j E_{lj}},$$

and can be compared to the χ^2-distribution with $L - 1$ degrees of freedom.

Techniques for Exploring Pairwise Associations

When conducting data exploration, it can be important to examine relationships between covariates in addition to the exploration of covariate effects on the response. There are several methods which can be used to measure associations between covariates, and the appropriate method should be chosen dependent upon variable type.

For two categorical or binary covariates, X_1 and X_2 say, the χ^2-test can be used to assess any associations between them. The χ^2-test will assess the null hypothesis that X_1 and X_2 are independent, against the alternative that they are not independent. Suppose X_1 and X_2 have A and B total levels respectively, and level $a = 1, ..., A$ and $b = 1, ..., B$. The expected frequency count for level a of X_1 and level b of X_2 is given as

$$E_{a,b} = \frac{n_a \times n_b}{n},$$

where n_a and n_b are the total number of sample observations at level a of X_1 and level b of X_2 respectively, and n is the total sample size. The test statistic then follows the χ^2-distribution and is defined as

$$\chi^2 = \sum \frac{(O_{a,b} - E_{a,b})^2}{E_{a,b}},$$

where $O_{a,b}$ is the observed frequency count at level a of X_1 and level b of X_2. To assess the null hypothesis, this test statistic is compared to the χ^2-distribution with $(A - 1)(B - 1)$ degrees of freedom at the chosen significance level.

An alternative test for categorical covariates is Kendall's τ, however this is specifically used to assess the relationship between two ordered categorical variables. Kendall's τ considers the difference between the number of concordant pairs and the number of discordant pairs for the two covariates being considered, and gives a non-parametric measure of correlation.

To assess the association between a continuous covariate and a categorical covariate, the ANOVA F-test can be used. In order to carry out this test, a linear regression is set up between the two covariates, where one is chosen to be the dependent covariate and the other is identified as the independent covariate. The linear regression can assess if the independent covariate can explain or predict the dependent covariate. The F-statistic can be obtained through calculating the error sum of squares and regression sum of squares, which can then be compared to the F-distribution for a chosen significance level and appropriate degrees of freedom.

Pearson's correlation coefficient can be used to assess the association between two continuous covariates. Consider two continuous covariates X_1 and X_2, with sample size n, then the Pearson's correlation coefficient, ρ, is given by

$$\rho = \frac{n\sum_{i=1}^{n} X_{1i}X_{2i} - (\sum_{i=1}^{n} X_{1i})(\sum_{i=1}^{n} X_{2i})}{\sqrt{[n\sum_{i=1}^{n} X_{1i}^2 - (\sum_{i=1}^{n} X_{1i})^2][n\sum_{i=1}^{n} X_{2i}^2 - (\sum_{i=1}^{n} X_{2i})^2]}}.$$

The correlation coefficient ρ can take any value between -1 and 1, and the t-distribution can be used to assess if ρ is significantly different from zero, indicating an association between the two covariates.

3.2.4 Cox Proportional Hazards Model

After initial data exploration, in order to examine the relationship between survival and explanatory variables further, a statistical modelling approach is needed. As the hazard or risk of an event, such as death, occurring is of primary interest in the analysis of survival data, the hazard function is modelled directly. Through modelling the hazard function, we can determine which combination of explanatory variables affect the form of the hazard function, whilst examining the extent to which the explanatory variables affect the hazard function. This provides a method to obtain an estimate of the hazard function for an individual (Collett, 2015).

The Cox Proportional Hazards (PH) model is the most commonly used procedure for modelling survival data, and is the method we will focus upon. The Cox

PH model is generally given in terms of the hazard function, defined as

$$h(t|\boldsymbol{X}) = h_0(t)\exp(\boldsymbol{\beta}'\boldsymbol{X}), \tag{3.1}$$

where $h_0(t)$ is the baseline hazard function, $\boldsymbol{X}$ is the observed explanatory variables, and $\boldsymbol{\beta}$ is the vector of coefficients for the explanatory variables.

The key assumption of the Cox PH model is the proportional hazards assumption, where the hazard ratio (HR) is constant with respect to time t, that is, there are no time by predictor interactions. Another key property of the Cox PH model is that no assumptions are made regarding the shape of the underlying hazard or survivor functions, $h(t)$ and $S(t)$. This means that the β-coefficients can be estimated without requiring estimation of the baseline hazard function, $h_0(t)$. Also, in it simplest form, the model assumes linearity and additivity of the predictors with respect to the log hazard or log cumulative hazard.

In order to fit the Cox PH model, the unknown β-coefficients for the explanatory variables need to be estimated. This can be done using maximum likelihood estimation. Within the Cox PH model, this involves obtaining the joint probability of the observed data as a function of the observed survival times and the unknown β-coefficients in the linear component in the model; giving the likelihood of the sample. As the hazard is presumed to be zero in the intervals between successive event times, the construction of the Cox PH model is based on the assumption that these intervals do not convey any information about the effect of the explanatory variables on the hazard function.

Suppose we have n individuals, and N distinct event times in the data. Assume there are no ties in the data, so that only one individual has an event at each event time. Denote the N ordered event times as $t_1, t_2, ..., t_N$, where $t_1 < t_2 < ... < t_N$, with t_j denoting the jth ordered event time. Let $R(t_j)$ denote the risk set at t_j, which is the set of individuals at risk at t_j, meaning they are event free and uncensored just before t_j. The likelihood function defined by Cox (1972) for the model in Equation (3.1) is given as

$$L(\boldsymbol{\beta}) = \prod_{j=1}^{N} \frac{\exp(\boldsymbol{\beta}'\boldsymbol{x}_j)}{\sum_{l \in R(t_j)} \exp(\boldsymbol{\beta}'\boldsymbol{x}_l)}, \tag{3.2}$$

where $\boldsymbol{x}_j$ is the vector of covariates for the individual with an event at t_j. This is defined as a partial likelihood as it depends only on the ranking of the event times and does not use the censored and uncensored survival times directly.

An alternative form of the partial likelihood includes a censoring indicator. Suppose the data consists of n observed survival times, $t_1, t_2, ..., t_n$ and let δ_i denote the censoring indicator which is zero if the ith survival time is right-censored, and one otherwise. Let $R(t_i)$ denote the risk set at t_i. The partial likelihood function given in Equation (3.2) can be expressed as

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left[\frac{\exp(\boldsymbol{\beta}'\boldsymbol{x}_i)}{\sum_{l \in R(t_i)} \exp(\boldsymbol{\beta}'\boldsymbol{x}_l)} \right]^{\delta_i},$$

and the corresponding log-likelihood function is given by

$$\log L(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i \left\{ \boldsymbol{\beta}'\boldsymbol{x}_i - \log \sum_{l \in R(t_i)} \exp(\boldsymbol{\beta}'\boldsymbol{x}_l) \right\}. \tag{3.3}$$

Maximising the log-likelihood function in Equation (3.3) enables us to find the maximum likelihood estimates of the β-coefficients in the Cox PH model. This can be achieved using numerical methods, generally the Newton-Raphson procedure.

The Newton-Raphson procedure starts with an initial guess for β, denoted $\beta^{(0)}$. The algorithm then iteratively computes

$$\hat{\beta}^{(n+1)} = \hat{\beta}^{(n)} + \mathcal{I}^{-1}(\hat{\beta}^{(n)})U(\hat{\beta}^{(n)}),$$

until convergence, where U is the score vector obtained by differentiating the log partial likelihood with respect to β. Convergence is assessed by the stability of the log partial likelihood (Therneau and Grambsch, 2013).

The likelihood and log-likelihood defined above assume no tied events, or survival, times, however tied survival times can arise due to often being recorded to the nearest day or month. Multiple censored observations can also occur at an event time. If at a given time, both censored and uncensored survival times occur, the censoring is assumed to occur just after all the events to remove ambiguity about which individuals should be included in the risk set for that time.

There have been several methods proposed to handle ties in survival times, including Efron (1977), Cox (1972) and Breslow (1974), with the approximation suggested by Breslow (1974) being the simplest (Collett, 2015). For an individual with an event at time t_j, denote the vector of sums of the q explanatory variables as $\boldsymbol{w}_j$. If the number of events at t_j is d_j, then the kth element of $\boldsymbol{w}_j$ can be defined as $w_{kj} = \sum_{h=1}^{d_j} x_{kjh}$, where x_{kjh} is the value of the kth explanatory variable for the hth of d_j individuals who have an event at t_j. Breslow (1974) proposed the approximate likelihood

$$\prod_{j=1}^{N} \frac{\exp(\boldsymbol{\beta}'\boldsymbol{w}_j)}{\left\{\sum_{l\in R(t_j)} \exp(\boldsymbol{\beta}'\boldsymbol{x}_l)\right\}^{d_j}}, \tag{3.4}$$

and Efron (1977) proposed the approximation

$$\prod_{j=1}^{N} \frac{\exp(\boldsymbol{\beta}'\boldsymbol{w}_j)}{\prod_{h=1}^{d_j}\left[\sum_{l\in R(t_j)} \exp(\boldsymbol{\beta}'\boldsymbol{x}_l) - (h-1)d_j^{-1}\sum_{l\in D(t_j)} \exp(\boldsymbol{\beta}'\boldsymbol{x}_l)\right]}, \tag{3.5}$$

where $D(t_j)$ is the set of individuals with an event at t_j. Efron's approximation in Equation (3.5) gives a closer approximation, but the approximation by Breslow, Equation (3.4), is appropriate when the number of tied observations at event times is not too large, and both of these approximations often give similar results.

3.2.5 Model Selection

The model selection process is concerned with determining a parsimonious subset of variables for inclusion in the model, where the strategy taken to carry out model selection may depend upon the purpose of the study (Collett, 2015). The procedures outlined below coincide with the aim of identifying a subset of variables upon which the hazard function depends.

There are three commonly used approaches which can be automated in statistical software: forward selection, backwards selection and the stepwise approach. Forward selection starts with the null and adds variables one at a time, where the variable chosen for inclusion at each stage will be the one with the most significant association with the outcome which is not already included in the model (Simon and Altman, 1994). Addition of variables stops when the model is no longer improved by inclusion of further variables, where improvement can be assessed using various testing procedures with a predefined significance level.

Backwards selection, on the other hand, begins by fitting a model containing the maximum number of variables under consideration, and excludes variables one at a time. Again this process ceases when omission of further variable does not significantly improve the model. The stepwise approach is a combination of the two procedures, where initially like forward selection, the procedure begins with addition of variables, however, variables already included in the model can be considered for exclusion at a later stage. In other words, after each addition, the procedure checks if any previously included variables can now be removed (Collett, 2015).

Each of these procedures can be flawed in application as the resulting parsimonious model has been selected solely on statistical grounds. In application terms this may potentially result in excluding variables deemed to be relevant by experts in the topic of the application, for example, clinical relevance. Backwards selection has been suggested to be the best option for this reason, as it allows for examination of the full model (Clark et al., 2003). Further, the full model is the only fit to provide accurate statistical measures (Harrell, 2006).

There are several tests which can be used to assess the significance of any improvement from addition or exclusion of a variable. The likelihood ratio, Wald and score tests are the standard likelihood inference tests, and can be used to assess hypotheses about β for the Cox model. The global null hypotheses is $\beta = \beta^{(0)}$, where $\beta^{(0)}$ is the initial value for the estimate of $\hat{\beta}$, often defaulted as zero in statistical software (Therneau and Grambsch, 2013).

The likelihood ratio test is defined as twice the difference between the log partial likelihood at the initial and final estimates of $\hat{\beta}$, algebraically this is

$$2\{l(\hat{\beta}) - l(\beta^{(0)})\},$$

where $l(.)$ is the log partial likelihood. The Wald test is defined as

$$(\hat{\beta} - \beta^{(0)})'\mathcal{I}(\hat{\beta})((\hat{\beta} - \beta^{(0)}),$$

where $\mathcal{I}(\hat{\beta})$ is the estimated information matrix. For a single covariate, the Wald test reduces to the z-statistic, $\hat{\beta}/\text{SE}(\hat{\beta})$, where $\text{SE}(\hat{\beta})$ is the standard error of $\hat{\beta}$.

The score test statistic is calculated using the first iteration of the Newton-Raphson procedure, and is defined as

$$U'(\beta^{(0)})\mathcal{I}(\beta^{(0)})^{-1}U(\beta_0).$$

The score test is closely related to a Wald test based upon one iteration of the Newton-Raphson procedure. These tests are asymptotically equivalent, and the null hypothesis distribution for each of them is the χ^2-distribution on q degrees of freedom (Therneau and Grambsch, 2013).

3.2.6 Model Validation

Once the final model has been fitted, model diagnostics need to be carried out in order to test if the model fits to the data adequately and assess if the underlying model assumptions have been satisfied. There are several different types of residual checking that can be used to conduct model validation. The key assumption to check is the proportional hazards assumption, which can be assessed visually or in a formal test using the Schoenfeld residuals (Schoenfeld, 1982). It is also important to check the functional form of covariates; this can be done using martingale residuals or smoothing splines.

Schoenfeld Residuals

The ith Schoenfeld residual for the kth explanatory variable in the model is an estimate of the ith component of the first derivative of the partial log-likelihood function, evaluated at $\hat{\boldsymbol{\beta}}$, and is defined as

$$s_{ki} = \delta_i \left\{ x_{ki} - \frac{\sum_{l \in R(t_i)} x_{kl} \exp(\hat{\boldsymbol{\beta}}' \boldsymbol{x}_l)}{\sum_{l \in R(t_i)} \exp(\hat{\boldsymbol{\beta}}' \boldsymbol{x}_l)} \right\},$$

where x_{ki} is the value of the kth explanatory variable, X_k, for the ith individual and $R(t_i)$ is the risk set at t_i. Schoenfeld residuals are only non-zero for uncensored observations, and as the β-coefficients are estimated such that

$$\left. \frac{\partial \log L(\boldsymbol{\beta})}{\partial \hat{\beta}_k} \right|_{\hat{\boldsymbol{\beta}}} = 0,$$

we have the constraint that the Schoenfeld residuals must sum to zero. A scaled version of the Schoenfeld residuals proposed by Grambsch and Therneau (1994) is more effective in detecting a violation of the proportional hazards assumption.

Defining $\boldsymbol{s}_i$ to be the vector of Schoenfeld residuals for the ith individual, the scaled Schoenfeld residuals, s^*_{ki}, are the the components of the vector

$$\boldsymbol{s}^*_i = d\,\mathrm{var}(\hat{\boldsymbol{\beta}})\,\boldsymbol{s}_i,$$

where d is the number of events among the n individuals and $\mathrm{var}(\hat{\boldsymbol{\beta}})$ is the variance-covariance matrix of the parameter estimates in the fitted Cox regression model.

The expected value of the scaled Schoenfeld residual, s^*_{ki}, for the ith individual and kth explanatory variable is

$$\mathbb{E}(s^*_{ki}) \approx \beta_k(t_i) - \hat{\beta}_k, \tag{3.6}$$

where $\beta_k(t)$ is the time-varying coefficient of the kth explanatory variable, so that $\beta_k(t_i)$ is the value of this coefficient at the event time of individual i, and $\hat{\beta}_k$ is the estimated value of β_k.

Testing the Proportional Hazards Assumption

The proportional hazards assumption can be assessed visually by plotting the scaled Schoenfeld residuals against the observed survival times, or more formally by calculating a test statistic. Hosmer et al. (2008) recommend the assessment of the proportional hazards assumption should be a two-step procedure, where covariate specific tests are calculated as step 1, and step 2 plots the scaled and smoothed Schoenfeld residuals; the results of the two steps should support each other.

In order to understand how the proportional hazards assumption can be assessed, firstly consider a model with a time-dependent coefficient

$$h_i(t) = h_0(t)\exp(\boldsymbol{\beta}(t)'\boldsymbol{x}_i).$$

If $\beta_k(t)$ is not a constant, this implies that the impact of the kth explanatory variable varies over time, violating the proportional hazards assumption. If the proportional hazards assumption holds then a plot of $\hat{\beta}_k(t)$ against time will be a horizontal line.

As defined in Equation (3.6), the expected value of the scaled Schoenfeld residual, s^*_{ki}, is approximately the difference between the time-varying coefficient of the kth explanatory variable and the estimated value, $\hat{\beta}_k$, of β_k. The proportional hazards assumption is satisfied if $\hat{\beta}_k(t)$ is equal to $\hat{\beta}_k$, therefore, under the proportional hazards assumption, the expected value of the scaled Schoenfeld residuals should be close to zero and constant over time. This relationship means that the Schoenfeld residuals can be used to assess the proportional hazards assumption.

Plotting the scaled Schoenfeld residuals against the observed survival times, or some function of time, provides information regarding the form of the time-varying coefficient $\beta_k(t)$, presenting a way to visualise the nature and extent of the non-proportional hazards. In particular, if this plot presents a horizontal line, this would suggest the proportional hazards assumption is satisfied.

To test for proportional hazards more formally, fitting a line to the plot and testing for zero slope can be used, where a non-zero slope would give evidence of a violation of the proportional hazards assumption. A spline fit of the scaled Schoenfeld residuals against time can be used to aid in the visualisation of the slope. Further, Therneau and Grambsch (2013) suggest that an analogy to generalised least squares can be used to motivate a formal test statistic, where the linear dependence can be expressed by writing $\beta(t)$ as a regression on some function of time $G(t)$, so that the dependence can be expressed as

$$\beta_k(t) = \beta_k + \theta_k G(t).$$

The null hypothesis for proportional hazards corresponds to $\theta_k = 0$, $k = 1, ..., q$. Given the expected values of the scaled Schoenfeld residuals, as defined in Equation

(3.6), assessment of this null hypothesis, $\theta_k = 0$, can be done using a score test of the relationship between the scaled Schoenfeld residuals and time. This test is outlined in more depth in Chapter 5.

Martingale Residuals

Martingale residuals can be used to assess the functional form of covariates. Therneau and Grambsch (2013) outline martingale residuals based upon counting processes, however a simple approach to defining them is to show how they can be obtained through modification of Cox-Snell residuals (Collett, 2015).

For the ith individual, $i = 1, ..., n$, the Cox-Snell residual can be defined as

$$r_{C_i} = \hat{H}_i(t_i) = -\log \hat{S}_i(t_i), \tag{3.7}$$

where $\hat{H}_i(t_i)$ and $-\log \hat{S}_i(t_i)$ are the estimated values of the cumulative hazard and survivor functions, respectively, of the ith individual at t_i, the observed survival time of individual i. The Cox-Snell residuals, r_{C_i}, can be thought of as a censored sample with a unit exponential distribution (Hosmer et al., 2008).

Collett (2015) outlines that if the observed survival time of individual i is right-censored, then the Cox-Snell residual of individual i will also be right-censored, and hence the definition of the Cox-Snell residuals in Equation (3.7) can be modified to account for censored survival times. Let t_i^* denote the ith survival if it is censored, where t_i is the actual, but unknown, survival time. The Cox-Snell residual evaluated at the censored survival time for this individual is defined by replacing t_i with t_i^* in Equation (3.7), where $\hat{H}_i(t_i^*)$ and $-\log \hat{S}_i(t_i^*)$ again define the estimated values of the cumulative hazard and survivor functions, respectively, but now of the ith individual at the censored survival time.

Now, to take into account that the greater the survival time the greater the corresponding Cox-Snell residual will be, the Cox-Snell residuals for censored survival times can be modified by the addition of a positive constant. This constant can be taken to be one since the r_{C_i} are assumed to follow the unit exponential distribution, and the modified Cox-Snell residuals can be defined as

$$\tilde{r}_{C_i} = 1 - \delta_i + r_{C_i}, \tag{3.8}$$

where δ_i is the event indicator taking the value zero if the observed survival time is censored, and one if it is uncensored.

To define the martingale residuals, Collett (2015) suggests the modified Cox-Snell residuals in Equation (3.8) can be refined so that they have zero mean when an observation is uncensored, and further multiply through by -1. This results in the martingale residuals being defined as

$$r_{M_i} = \delta_i - r_{C_i}.$$

The martingale residuals can therefore be thought of as the difference between the observed number of events for individual i in time interval $(0, t_i)$ and their conditionally expected number of events given the fitted model (Therneau and Grambsch, 2013). Martingale residuals take values in the interval $(-\infty, 1]$, and are negative for censored survival times.

In order to assess the functional form of covariates, Therneau et al. (1990) suggested plotting the martingale residuals from the null model, where $\hat{\beta} = 0$, against each covariate X_k separately. Superimposing a scatterplot smooth can then indicate the functional form of X_k. Let f denote the smooth function. Therneau and Grambsch (2013) state that if the correct model for the kth covariate is $\exp(f(X_k)\beta_k)$, then the smooth for the kth covariate will display the form of f. This can be expressed mathematically as

$$\mathbb{E}(r_{M_i} | X_{ik} = x_k) \approx cf(x_k),$$

where c simply scales the y-axis and depends on the amount of censoring, but is roughly independent of x_k.

Splines

As flexible fitting functions, Therneau and Grambsch (2013) recommend splines as useful tools for exploring the functional form of covariates. As smoothing tools, splines are able to summarise a trend between a response and one or more predictors by producing a trend less variable than the response itself, whilst avoiding the assumption of a rigid form of the dependence of the response on the predictors (Hastie and Tibshirani, 1990). A key property of splines is the locality of influence, where a large change in one part of the curve will have minimal affect on the fit in other areas of the curve.

There are several types of splines, including regression or natural splines, and smoothing splines. To introduce how splines can be fitted, we first outline regression splines. Regression splines represent the trend between a response and predictor through piecewise polynomials, where breakpoints, or knots, separate the regions defining the pieces. Hastie and Tibshirani (1990) state that piecewise cubic polynomials are a common choice, where the polynomials are constrained to have first and second derivatives at the knot points to ensure they join smoothly at these points. A larger number of knot points gives more flexibility to the curve.

Basis functions are needed to represent the particular family of piecewise polynomials, with basis vectors being the basis functions evaluated at the observed values of the predictor. For regression splines, the smooth for any given set of knots is computed using multiple regression on the basis vectors.

Regression splines require specification of both the position and number of knot points however, and Hastie and Tibshirani (1990) highlight this as a drawback since poor choice of horizontal position of the knots can result in non-local behaviour. Smoothing splines have been shown by Hastie and Tibshirani (1990) to have better properties regarding locality of influence compared to regression splines for small degrees of freedom, or number of knots. Further, Therneau and Grambsch (2013) highlight that from a user perspective, smoothing splines are simpler as they only require pre-specification of the number of knots, or degrees of freedom.

Smoothing splines are not constructed explicitly as with regression splines, but instead are a result of an optimization problem. Denote the response measure by y and the predictor as x for a spline based on a large number of knots, denoted $f(x, \beta)$. Therneau and Grambsch (2013) state that choice of the coefficients, β, for the basis functions should be chosen to minimize the combined criterion

$$\theta \sum_{i=1}^{n} [y_i - f(x_i, \beta)]^2 + (1 - \theta) \int [f''(x, \beta)]^2 \, \mathrm{d}x,$$

where the first term is the residual sum of squares, measuring closeness to the data, and the second term penalizes curvature in the function. In order to optimize the function, θ acts as the tuning parameter, where values of θ closer to zero minimise the curvature towards the linear least squares line, and as θ approaches one, the solution converges to an interpolating curve passing through every point, giving n degrees of freedom.

Therneau and Grambsch (2013) highlight that although smoothing splines are computationally more difficult than regression splines, smoothing splines can be fitted as a special case of penalized proportional hazards models in R using the `pspline` function within the `survival` package (Therneau, 2015a). A plot of a smoothing spline for particular predictors can show their functional form in relation to the response. Superimposing simpler smooths over the plot, such as a quadratic polynomial, can indicate if this would be sufficient to represent the trend between the response and predictor in a model.

3.3 Missing Data

Missing data is a common issue in observational studies, and refers to observations we intended to make but did not (Carpenter and Bartlett, 2016). Missing data complicates the statistical analysis of data collected in almost every discipline, and generally causes two key problems; bias and loss of efficiency. Loss of efficiency, or information, is an inevitable consequence of missing data, however it is not directly related to the proportion of incomplete records. Biased inferences are caused by mishandling of the missing data, and the extent of bias depends on the statistical behaviour of the missing data, including the patterns and mechanisms of the missing data (Carpenter and Kenward, 2013).

The subset of complete records is not necessarily representative of the full study population, dependent upon the patterns and mechanisms of missingness. Restricting analysis to the complete records can therefore lead to biased inferences, and thus knowledge about the patterns and mechanisms is vital in deciding the most appropriate method for handling the missing data (Molenberghs et al., 2014).

3.3.1 Classification of Missing Data

Missing data can be categorised in two ways; by considering the patterns or by the mechanism, which considers the underlying reason why the data is missing.

There are four main types of missing data; univariate, monotone, file matching and arbitrary patterns. Univariate missing data is the simplest and refers to the case where missingness is confined to one variable. Data following a monotone missingness pattern can be sorted according to the percentage of missing data. File matching is where two sets of variables are never jointly observed, and an arbitrary pattern means missing values occur in any variable in any position.

The missing data mechanism describes the probability that a response is observed or missing, and is not in the control of the study investigator. Rubin (1976) specified a hierarchy of three different types of missing data mechanisms: Missing Completely at Random (MCAR), Missing at Random (MAR) and Not Missing at Random (NMAR).

In order to define the missing data mechanisms, suppose we have n individuals and let $\boldsymbol{x}_i = (x_{i,1}, x_{i,2}, ..., x_{i,q})'$ denote the q variables intended to be observed from the ith individual, $i = 1, ..., n$. For each individual i, let $\boldsymbol{x}_{i,obs}$ denote the subset of variables that are observed, and $\boldsymbol{x}_{i,mis}$ denote the subset that are missing. These can be different subsets of the q variables for different individuals. Define $D_{i,k}$ to be the missing data indicator for the ith individual and kth variable, where $D_{i,k}$ is zero if $x_{i,k}$ is observed and one if $x_{i,k}$ is missing. Let $\boldsymbol{D}_i = (D_{i,1}, D_{i,2}, ..., D_{i,q})'$. We can then define the missing data mechanism as the probability of $\boldsymbol{D}_i$ conditional upon $\boldsymbol{x}_i$, $\mathbb{P}(\boldsymbol{D}_i|\boldsymbol{x}_i)$ (Carpenter and Kenward, 2013).

Missing Completely at Random

Missing Completely at Random (MCAR) refers to data where the probability of a value being missing is not associated with the observed or unobserved responses for that individual. In algebraic terms,

$$\mathbb{P}(\boldsymbol{D}_i|\boldsymbol{x}_i) = \mathbb{P}(\boldsymbol{D}_i).$$

For MCAR data, as the chance of being missing is unrelated to the values, the observed data is representative of the population, however, information has been lost relative to the information that was intended to be collected.

Missing at Random

Data is classified as Missing at Random (MAR) when the missingness depends only on the observed values of the data set, and not on the components that are missing. Given the observed data, the probability of a value being missing is independent of the unobserved data, and this can be expressed algebraically as

$$\mathbb{P}(\boldsymbol{D}_i|\boldsymbol{x}_i) = \mathbb{P}(\boldsymbol{D}_i|\boldsymbol{x}_{i,obs}).$$

Under MAR, the chance of a variable being missing will depend on its value, however, when conditioned on the observed data, this dependence is broken.

Not Missing at Random

Data is Not Missing At Random (NMAR) if the probability of an observation being missing is dependent on the underlying value that should have been obtained, where this dependence remains even given the observed data. This can be expressed as

$$\mathbb{P}(\boldsymbol{D}_i|\boldsymbol{x}_i) \neq \mathbb{P}(\boldsymbol{D}_i|\boldsymbol{x}_{i,obs}).$$

Analysis is more complex under NMAR, and in general, the specification of a model for the missing data mechanism is required for any valid inferential method under NMAR.

It is important to note that it is not possible to distinguish between MAR and NMAR from the observed data alone, making NMAR un-testable. For this reason, it is essential to explore the missing data thoroughly, and also consider methods of data collection and expert opinions when attempting to identify which mechanism may be the most plausible.

Methods for Exploring Missing Data

In order to classify missing data, patterns and associations within the data, particularly related to the missingness, need exploring. Visualisation tools are particularly useful for the exploration of missing data, where the `VIM` package (Kowarik and Templ, 2016) in R contains many features designed for this purpose. The `VIM` package in R provides many visualisation tools for exploring missing data, providing useful techniques for examining the data structure and detecting the most plausible missing data mechanism. Below we provide an overview of some of the more general visualisation tools, where further discussion and additional techniques can be found in Templ and Filzmoser (2008).

When exploring missing, it is of particular interest to identify the amount of missing values within each variable, and also identify combinations of variables which have a high number of missing values (Templ and Filzmoser, 2008). *Aggregation plots* are a useful tool to explore this, where these plots show the combinations of missing and observed for the incomplete variables. Within aggregation plots, each column represents a particular covariate, and each row gives a different combination of missing and observed. The rows can be ordered so that they ascend from most common combination to least common, with a histogram or count alongside to show the frequency of each combination. The `aggr` function used to produce aggregation plots also provides histograms showing the proportion of missing data for each covariate considered within the plot. Aggregation plots are useful for identifying the missingness pattern of the data as a whole, alongside patterns within particular subsets of variables.

Matrix plots are another useful tool for examining missing data. Templ and Filzmoser (2008) describes the matrix plot as a visualisation of each cell of the data matrix, where each cell is represented by a horizontal line. Observed data are presented on a grey scale and missing values are drawn as red lines. As with aggregation plots, each column will represent a covariate, and the rows present the observed or missing values for each individual. These plots are useful for identifying the mechanism of the missing data, showing how missingness relate to the values of the observed data. Further, these plots can be ordered by the values of particular covariate to aid in understanding and classification of the missing data mechanism.

To explore the dependency of missingness on other covariate values in more depth, a *spineplot* can be used, where a spineplot shows the amount of missingness in one variable dependent upon the values of another, in a similar manner to a stacked histogram. The horizontal axis gives the categories of the explanatory covariate, scaled according to the relative frequencies within each category. The vertical axis gives the proportion of missing and observed of the dependent covariate within each category of the explanatory covariate (Templ and Filzmoser, 2008). Comparison of the area for missing across the explanatory covariate categories can indicate the missing data mechanism between the two covariates being considered. For example strictly decreasing, or increasing, areas for missing would indicate MAR.

The `marginplot` function in the `VIM` package is also a useful tool for exploring the missing data mechanism between two continuous covariate, where this function extends upon the usual two-way scatter plot. The main body of these plots gives a two-way scatter plot of the observed data between the two covariates, with univariate scatter plots of the missing data being given in the margins. Additionally, box plots are given in the margins, for missing and observed separately, where a comparison of the box plots can identify the missing data mechanism.

3.3.2 Methods for Handling Missing Data

There are many methods for handling missing data, and these can be categorised as procedures based on completely recorded data, weighting procedures, imputation-based procedures and model-based procedures (Little and Rubin, 2002). These categories are not mutually exclusive however.

Procedures based on completely recorded data simply means that individuals with incomplete data are discarded, and only individuals with complete data are included in the analysis; this method will be outlined in more detail below. Weighting procedures involve modifying the contribution, or weight, of each observation in an attempt to adjust for non-response as though it were part of the sample design. For example, a covariate pattern with larger proportions of missing observations would result in higher weights for the observed data within that covariate pattern.

Imputation-based procedures fill in the missing values so that the resultant complete data can be analysed via standard methods. There are many types of imputation-based procedures which will be outlined below. Model-based procedures broadly covers methods involving a model being defined for the observed data with inferences based on the likelihood or posterior distribution under that model.

Complete-case Analysis

Complete-case analysis is one of the simplest and most commonly used methods for handling missing data (Molenberghs et al., 2014), and may be considered advantageous due to its simplicity as standard complete-data statistical analyses can be applied without modification. However, as complete-case analysis restricts attention to individuals where all the variables are observed, it can lead to loss of information from discarding incomplete cases. This can result in loss of precision, and also the introduction of bias if the data are not MCAR. Use of complete-case analysis can be justified when the bias and loss of precision is minimal, and this is most likely when the proportion of complete cases is high. However, the loss of precision and degree of bias do not solely depend on the proportion of missingness, making it difficult to develop general rules of thumb for an acceptable proportion. The extent that complete cases differ and the parameters of interest also influence bias and loss of precision (Little and Rubin, 2002).

Complete-case analysis can be adjusted for bias using weighting methods, where estimations are based on observed responses which are weighted is some way to account for the probability of non-response. Weighting methods are generally best suited to monotone data, and can be complicated and conceptually difficult to formulate in practice for non-monotone data (Molenberghs et al., 2014).

For univariate analyses, available-case analysis is an alternative to complete case analysis. Loss of efficiency can be particularly high in complete-case analysis for data containing a large number of variables; available-case analysis attempts to reduce this loss of efficiency. The idea behind this method is to use all the available information, however due to the changing sample base between variables, this method has practical problems. As sophisticated optimisation techniques and special formulas are required to calculate standard errors, van Buuren (2012) recommends that this pairwise deletion method only be used if the procedure that follows it is specifically designed to take the deletion into account.

Full Likelihood and Bayesian Approach

Many methods for estimating the missing values of incomplete data can be based upon likelihood and Bayesian approaches, where estimation is based upon the likelihood function under specific modelling assumptions, and the likelihood for the parameters can be derived based on the incomplete data. Maximum likelihood estimation can be carried out by solving the likelihood equation, and Bayesian inference can be carried out by using a prior distribution to obtain the posterior distribution. Little and Rubin (2002) have highlighted complications with these methods however. Compared to likelihood or Bayesian inference of complete data, Little and Rubin (2002) note that for the incomplete data setting, the asymptotic standard errors calculated from the information matrix are more questionable as the observed data are generally not an independent and identically distributed sample. Further complications arise when dealing with the underlying reason for the occurrence of missing data, where Molenberghs et al. (2014) highlight that NMAR, a non-ignorable mechanism, can never be truly ruled out.

The methods within the likelihood and Bayesian approaches depend upon whether the missing data mechanism is ignorable. Let $\boldsymbol{X}$ denote the intended data set, so that $\boldsymbol{X} = (\boldsymbol{X}_{obs}, \boldsymbol{X}_{mis})$, where $\boldsymbol{X}_{obs}$ denotes the observed data and $\boldsymbol{X}_{mis}$ denotes the missing values. The likelihood of θ based upon the observed data, $\boldsymbol{X}_{obs}$, but ignoring the missing data mechanism, can be defined to be any function of θ proportional to the marginal probability density of $\boldsymbol{X}_{obs}$,

$$L_{ign}(\theta|\boldsymbol{X}_{obs}) \propto f(\theta|\boldsymbol{X}_{obs}). \tag{3.9}$$

If the missing data mechanism can be ignored, inference about θ can be based upon this likelihood, where ignorable maximum likelihood estimation can be obtained using Equation (3.9) with respect to θ. Bayesian inference for θ based upon $\boldsymbol{X}_{obs}$ incorporates a prior distribution to give the posterior distribution

$$\pi(\theta|\boldsymbol{X}_{obs}) \propto \pi(\theta) \times L_{ign}(\theta|\boldsymbol{X}_{obs})$$

on which inferences can be based.

Let $D_{i,k}$ be the missingness indicator as defined in Section 3.3.1, and treat D as a random variable. To handle the case where the missing data mechanism is not ignorable we need to specify the full model. This can be achieved through specification of the joint distribution of D and $\boldsymbol{X}$, where this distribution defines the missing data mechanism, subject to an unknown parameter φ. The observed data consist of the values of variables $(\boldsymbol{X}_{obs}, D)$, and the full likelihood of θ and φ is any function of θ and φ based on the observed data $(\boldsymbol{X}_{obs}, D)$, to give

$$L_{full}(\theta, \varphi|\boldsymbol{X}_{obs}, D) \propto f(\boldsymbol{X}_{obs}, D|\theta, \varphi) \tag{3.10}$$

Bayesian inference can be obtained by combining the full likelihood with a prior distribution for θ and φ to the posterior distribution

$$\pi(\theta, \varphi|\boldsymbol{X}_{obs}, D) \propto \pi(\theta, \varphi) \times L_{full}(\theta, \varphi|\boldsymbol{X}_{obs}, D), \qquad (\theta, \varphi) \in \Omega_{\theta,\varphi}, \tag{3.11}$$

where $\Omega_{\theta,\varphi}$ is the parameter space of (θ, φ).

For both likelihood and Bayesian inference, the missing data mechanism is ignorable if the missing data are MAR. For likelihood inference, θ and φ must also be distinct, $\Omega_{\theta,\varphi} = \Omega_\theta \times \Omega_\varphi$, and additionally Bayesian inference requires θ and φ to be *a priori* independent, that is $\pi(\theta, \varphi) = \pi(\theta)\pi(\varphi)$. If these conditions do not hold, then for likelihood or Bayesian approaches, respectively, inference should be based on Equation (3.10) or Equation (3.11) (Little and Rubin, 2002).

Imputation

Imputation methods assign values to the non-observed data, and has two main approaches; single imputation or multiple imputation. Single imputation involves methods that can be applied to impute a single value for each missing observation, whereas multiple imputation involves methods which impute more than one value to allow for uncertainty. Little and Rubin (2002) define imputation as means or draws from a predictive distribution of the missing data. A method is therefore required to develop the predictive distribution based upon the observed data. Methods for imputation can be categorised into two types of approaches; explicit imputation which is based on a formal statistical model and implicit imputation which are more ad hoc methods for approaching imputation.

Examples of implicit imputation methods are hot deck imputation, cold deck imputation, substitution and composite methods. Explicit imputation methods include mean imputation, regression imputation, and stochastic regression imputation. We will outline several imputation methods below, including various single imputation methods and the multiple imputation procedure.

Last observation carried forward (LOCF) is a technique used to impute longitudinal data, and replaces every missing value with the last observed value from the same subject. *Baseline observation carried forward* (BOCF) takes a similar approach. LOCF has the underlying assumption that the most recent observation is the best guess for the subsequent missing values (McKnight et al., 2007). Although LOCF is convenient in supplying a complete data set, it has several issues, as highlighted by van Buuren (2012). LOCF can yield biased estimates even under MCAR, and any statistical analyses following LOCF should distinguish between the real and imputed data.

Hot deck imputation involves substituting individual values drawn from other ‘similar’ responding units. Selection of these units that are deemed similar can involve complex elaborate schemes. Three methods to carry out hot deck imputation are the nearest neighbour technique, randomly selecting a value from the observed data with probability of selection based on rate of occurrence, and hot deck with adjustment cells, which is similar to the previous technique but also blocks on relevant covariates. Hot deck imputation is known to underestimate the standard errors as missing values are replaced with values which already exist in the observed data, decreasing variability. McKnight et al. (2007) also state that hot deck procedures can introduce biases that are unpredictable unless the data are MCAR.

Cold deck imputation is an alternative to hot deck imputation, and replaces each missing value with a single value from an external source. However, like hot deck imputation, cold deck imputation again assumes MCAR, and has the issue of introducing bias and underestimating variance sampling error.

Mean imputation is another ad hoc method used to produce a complete data set so that standard complete data methods of analysis can be used. Mean imputation has two forms; marginal and conditional. *Marginal mean imputation* imputes missing values using the average of the observed values for that variable, ignoring all other variables. This is problematic as ignoring other variables can cause misrepresentation of associations within the data set, and the precision will be overestimated. Little and Rubin (2002) express many concerns regarding marginal mean imputation including underestimating the variance, and the inability of standard complete data methods to produce consistent estimators after imputation has been carried out.

Conditional mean imputation is an improvement on marginal mean imputation as it imputes conditionally upon the observed values. One method to carry out conditional mean imputation is to classify non-respondent and respondent into adjustment classes, and impute the respondent mean for the non-respondents in the same class. Generally, conditional mean imputation imputes more plausible values than marginal mean imputation, however similar issues still arise as the values are less variable than the observed values, so standard errors are generally underestimated (Molenberghs et al., 2014).

Regression imputation replaces the missing values in the data set with predicted values from the regression of the covariate with missing values on covariates observed for the individual. A regression model can be developed based upon the variables associated with the variable containing missing values, and predicted values are usually calculated from individuals with both missing and observed values present. Multiple variable types can be incorporated into the regression model, as can less restrictive parameter forms such as splines. As with mean imputation, underestimation of the variance remains a downfall of regression imputation, unless steps are taken with the analyses of the 'completed' data to account for this. Allison (2001) recommends the addition of an error term to reduce the underestimation of the variance, a method known as *stochastic regression imputation*.

All of the single imputation methods have a common downfall in that inference about parameters based upon the imputed data do not account for imputation uncertainty. Unlike multiple imputation, single imputation methods are unable to capture the between-imputation variability, a reflection of the uncertainty due to missing information, thus resulting in standard errors which are too small (White et al., 2011).

Multiple Imputation

Multiple imputation (MI) involves imputing more than one value for each missing value, and is a method of handling missing data which provides a computationally feasible approach to wide range of problems under a wide range of missingness mechanisms (Carpenter and Kenward, 2013). There are three distinct steps involved within multiple imputation. Firstly, an appropriate imputation model needs to be specified and fitted to fill in the missing values M times, to create a series of M imputed data sets. Note that the observed data values remain the same across the imputed data sets, and only the originally missing values will differ (Azur et al., 2011). Next, the M completed data sets can be analysed using standard, complete data procedures; referred to as the substantive, or analysis, model. The final step is to combine the results from the M analyses to produce a single estimator and draw inferences. The results of the M analyses can be combined using Rubin's rules (Rubin, 1987), which are a general procedure for summarising M results in order to obtain point estimates and associated estimates of variance, and to carry out statistical tests.

Rubin's rules are based on asymptotic theory in Bayesian framework, and ensure the combined variance-covariance matrix incorporates both the within-imputation variability and the between-imputation variability. Within-imputation variability is the uncertainty about the results from an imputed data set, denoted W, and between-imputation variability, denoted B, reflects the uncertainty due to the missing information. Let $\boldsymbol{\beta}$ denote the vectors of parameters in the substantive model, where in general Rubin's rules will be applied to all or part of this parameter vector. Suppose $\hat{\beta}_m$ is an estimate of a univariate or multivariate quantity of interest obtained from the mth imputed data set, and W_m is the estimated variance of $\hat{\beta}_m$. The combined estimate $\hat{\beta}$ is the average of the individual estimates:

$$\tilde{\beta} = \frac{1}{M}\sum_{m=1}^{M}\hat{\beta}_m. \tag{3.12}$$

The variance estimator, $\text{var}(\tilde{\beta})$, is the total variance of $\tilde{\beta}$, and is formed from the within-imputation variance $W = (1/M)\sum_{m=1}^{M} W_m$ and the between-imputation variance $B = (1/(M-1))\sum_{m=1}^{M}(\hat{\beta}_m - \tilde{\beta})^2$. The variance estimator is defined as

$$\text{var}(\tilde{\beta}) = W + \left(1 + \frac{1}{M}\right)B. \tag{3.13}$$

Figure 3.1 gives a visualisation of the MI process, where the general MI procedure is outlined as follows: For a general data matrix $\boldsymbol{X}$, let $\boldsymbol{X}_{obs}$ be the observed data and $\boldsymbol{X}_{mis}$ be the missing data, then:

1. For $m = 1, ..., M$, impute the missing data from the distribution of the missing data given the observed data, $f(\boldsymbol{X}_{mis}|\boldsymbol{X}_{obs})$, ensuring to take full account of the uncertainty, to obtain M ‘complete’ data sets.

2. Fit the substantive model to each of the M imputed data sets, $m = 1, ..., M$, to obtain M estimates of the parameters of the substantive model, say $\hat{\beta}_m$, and also the M estimates of their variance, $\text{var}(\hat{\beta}_m)$.

3. Combine the parameter and variance estimates for inference using Rubin’s rules to obtain the pooled results, $\tilde{\beta}$ and $\text{var}(\tilde{\beta})$.

Section 3.3.3 below outlines in depth a procedure for carrying out Step 1, and Section 3.3.4 outlines how Step 3 can be conducted to achieve a parsimonious pooled model.

3.3.3 Multiple Imputation Using Chained Equations

Multiple imputation using chained equations (MICE) is an approach for carrying out Step 1 of the multiple imputation procedure outlined above. MICE is recommended as the method of choice for handling incomplete data problems by van

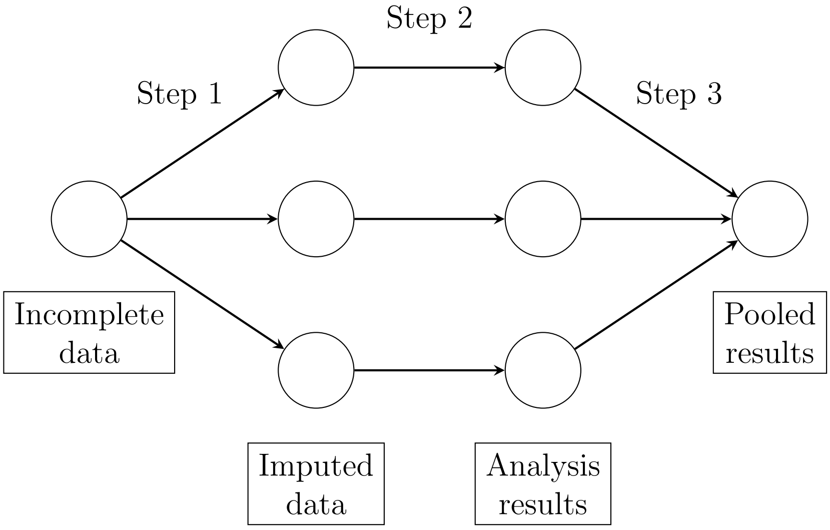


Figure 3.1: Diagram of the multiple imputation process; showing a single incomplete data, multiple imputed data sets, multiple analyses results and the pooled analyses results.

Buuren and Groothuis-Oudshoorn (2011). It is a practical approach to generating imputations based on a set of imputation models; one for each variable with missing values. An important feature of MICE is that it has the ability to handle different variable types (White et al., 2011). Due to each variable having its own imputation model, MICE can handle multiple variable types including continuous, binary, ordered categorical and unordered categorical.

The MICE procedure involves initially filling in all the missing values using simple random sampling with replacement from the observed values. Denote the first variable with missing values as x_1. Regress x_1 on all the other variables $x_1, ..., x_K$, restricted to individuals with observed x_1. The missing values in x_1 are then replaced by simulated draws from the corresponding predictive distribution of x_1. The next variable with missingness, x_2 say, is then regressed on all the other variables $x_1, x_3, ..., x_K$, restricted to individuals with observed x_2, and using imputed values of x_1. This process is repeated with all other variables with missing values in turn; this is called a cycle. The MICE procedure is then repeated, or iterated, for several cycles in order to stabilize the results and produce a single imputed data set. The whole procedure is then repeated M times to produce M

imputed data sets.

There are many important considerations to be made during the MICE procedure. These include specifying the appropriate form of the imputation models for different variable types, and inclusion of appropriate variables as predictors for the missing values; application to survival data further complicates specification of the imputation models within the MICE procedure. Two other important considerations are the choice of the number of cycles to iterate over, and choice of the number of imputed data sets to produce. These considerations will be discussed in depth throughout the remainder of this section.

Specification of the Imputation Models

The main aim in multiple imputation is to draw valid and efficient inferences by fitting analysis models to multiply imputed data. In order to avoid bias and gain precision, it is important to satisfy some key requirements when developing the imputation models. An imputation model should account for the process that created the missing data, preserve the relations in the data and preserve the uncertainty in the relations (van Buuren and Groothuis-Oudshoorn, 2011). Further, variables selected for inclusion in imputation models should be included in the appropriate way, ensuring correct functional form and including any required interactions. It is also necessary to ensure imputation models are of the correct form, where we need to ensure approximate compatibility with the analysis model, and that the imputation model is appropriate for the type of variable with missing values we wish to impute.

It can be easy to mis-specify the imputation model when aiming to find the true imputation model as this has no standard form. An alternative approach is to find an imputation model that is approximately compatible with the analysis model but is not necessarily correctly specified, where for two conditional models to be compatible, there should exist a joint model in which the conditionals for the relevant variables equal these conditional models. If imputations are drawn from a model incompatible with the analysis model, this can lead to biased estimates of parameters in the analysis model (Bartlett et al., 2015).

In terms of variable selection, the imputation model needs to include all the variables that are in the the analysis model, in particular, the imputation model must include the outcome from the analysis model. Further, Bartlett et al. (2015) recommend that the imputation model needs to account for any non-linear or interaction terms within the analysis model. This is to ensure approximate compatibility between the imputation model and analysis model, and avoid bias within the analysis after imputation.

Considering survival data, the outcome of the analysis model comprises of survival time t and the censoring indicator δ. Research has been conducted into the best approach for including this outcome, some possibilities are to include all of t, $\log(t)$ and δ, δ and $\log(t)$, or δ and t. In the case where the analysis model is a proportional hazards model, White and Royston (2009) found that inclusion of $\log(t)$ in the imputation model can bias associations towards the null. An alternative is to include δ and $H_0(t)$, where $H_0(t)$ is the cumulative baseline hazard function. White et al. (2011) state this gives the correct imputation model for a single binary covariate and is approximately correct for more complex situations. In general, $H_0(t)$ is not known, however the Nelson-Aalen estimate, $\hat{H}(t)$, (Aalen, 1978) of the cumulative hazard function provides an adequate approximation. It is recommended by White and Royston (2009) that the imputation model should be based on the Nelson-Aalen estimate of the cumulative hazard to the survival time.

Any predictors of the incomplete variable should also be included in the imputation model. This makes the MAR assumption more plausible, reducing bias, and can also help improve the imputations, in turn reducing the standard errors of the estimates in the analysis model. As it is not possible to distinguish between MAR and NMAR from the observed data alone, inclusion of more explanatory variables can help make the MAR assumption more plausible. The MAR condition for valid inferences is that, conditional on the observed data, the probability of data being missing does not depend on the unobserved data. Due to this, it is recommended by White et al. (2011) that the imputation model should included every variable that both predicts the incomplete variable and predicts whether the incomplete variable is missing.

Imputation Models for Different Variable Types

In order to determine the appropriate form of imputation models for different variable types in a survival data setting, we must first identify appropriate general approaches for each of the variable types. We consider approaches for continuous, binary and both unordered and ordered categorical variables.

A logistic regression model is usually chosen to impute the missing values in a binary variable. For imputing a Normally distributed continuous variable, a linear regression model is the most suitable choice. For both unordered and ordered categorical variables with more than two levels, multinomial logistic regression can be used to impute the missing values, and the proportional odds model can also be used for ordered categorical variables.

Now in the case of survival data, it can be assumed that the data follows a Cox proportional hazards model defined as

$$h(t|X,Z) = h_0(t)\exp(\beta_X X + \beta_Z Z),$$

where X is an incomplete variable and Z is complete. An exposure model is needed for incomplete X, that is $p(X|Z;\zeta)$. The imputation model is then defined as $p(X|t,\delta,Z)$, and a number of exact and approximate results have been proved by White and Royston (2009) regarding the imputation model in terms of the model parameters $\theta = (\zeta, \beta_X, \beta_Z, h_0(.))$. These results can be used to determine the regression model $p(X|t, \delta, Z; \alpha)$, where α is some function of θ. Although in practice θ is generally unknown, parameters α can be estimated directly from the complete cases.

The formulation of each of the imputation models for each of the variable types can be shown by considering the log-likelihood, given the complete data, for the survival data outcomes and applying Bayes' theorem to get the conditional distribution of X given the observed data. This will be shown in more depth in Chapter 6 where a deeper understanding is needed, but here we will give a brief overview of the model forms, as outlined by White and Royston (2009).

Lets first consider the case of binary X, and look at the simplest case where there is no Z. The missing X can be imputed by fitting a logistic regression of X on the censoring indicator, δ and the baseline cumulative hazard, $H_0(t)$, giving the model

$$\operatorname{logit} p(X = 1|t, \delta) = \alpha_0 + \alpha_1\delta + \alpha_2 H_0(t).$$

Considering a further case of binary X with binary or categorical Z, we take the most general exposure model $\operatorname{logit} p(X = 1|Z) = \zeta_Z$, and get a logistic regression on δ, Z, $H_0(T)$ and the interaction between $H_0(t)$ and Z

$$\text{logit} p(X = 1|t, \delta, Z) = \alpha_0 + \alpha_1\delta + \alpha_2 H_0(t) + \alpha_{3Z} + \alpha_{4Z} H_0(t),$$

where terms such as α_{3Z} represent a set of dummy variables with their coefficients. For the most general case of binary X with general Z there are no exact results. Assuming the exposure model $\text{logit} p(X = 1|Z) = \zeta_0 + \zeta_1 Z$, and taking the Taylor series approximation for $\exp(\beta_Z Z)$ for small variance of $\beta_Z Z$, we get a logistic regression on δ, $H_0(t)$ and Z:

$$\text{logit} p(X = 1|t, \delta, Z) = \alpha_0 + \alpha_1\delta + \alpha_2 H_0(t) + \alpha_3 Z.$$

Addition of an interaction term $\alpha_4 H_0(t) \times Z$ improves the accuracy of the approximation.

For continuous X and general Z, we take the exposure model $X|Z \sim N(\zeta_0 + \zeta_1 Z, \sigma^2)$ and make a fuller Taylor series approximation of $\exp(\beta_X X + \beta_Z Z)$ that is valid for small variance of $\beta_X X + \beta_Z Z$. This gives a regression model on Z, δ and $H_0(t)$, which can be written as

$$X|t, \delta, Z \sim N(\alpha_0 + \alpha_1 \delta + \alpha_2 H_0(t) + \alpha_3 Z, \sigma^2).$$

Again, the addition of an interaction term $\alpha_4 H_0(t) \times Z$ improves the accuracy of the approximation.

For the case of categorical X, where X has levels $l = 1, ..., L$, we can take the general exposure model $\operatorname{logit} p(X = l|Z) = \zeta_{l0} + \zeta_{l1} Z$, and again take an approximation for $\exp(\beta_Z Z)$. This give a multinomial logistic regression on Z, δ, and $H_0(t)$, written as

$$\log\left(\frac{p(X = l|t, \delta, Z)}{p(X = 1|t, \delta, Z)}\right) = \alpha_0 + \alpha_1 \delta + \alpha_2 H_0(t) + \alpha_3 Z.$$

As in the previous cases, addition of the interaction term $\alpha_4 H_0(t) \times Z$ improves the accuracy of the approximation.

Further Considerations

Another important aspect to consider within the MICE procedure is choosing the appropriate number of imputations. There are several arguments outlined by White et al. (2011) for choosing the number imputations, including the efficiency and reproducibility arguments, and a rule of thumb suggested by Bodner (2008).

The efficiency argument bases choice of the number of imputations on statistical efficiency of the estimates. Letting W and B be the within-imputation and between-imputation variance respectively, as defined in Section 3.3.2, the true variance of a parameter estimate based on M imputations is $W + (1 + 1/M)B$. Comparing infinitely many imputations to M imputations, the relative efficiency is

$$\frac{W + (1 + 1/M)B}{W + B} = 1 + \frac{F}{M},$$

where $F = B/(B + W)$, the fraction of missing information. Using this, M can be found to satisfy the condition of F/M being less than the accepted loss of efficiency. Another argument is that limiting loss of power should be considered when choosing the number of imputations (Graham et al., 2007).

The reproducibility argument stems from the need to be confident that a repeat analysis of the same data would produce essentially the same results, and would suggest that statistical efficiency and power are not sufficient considerations. This means that the Monte Carlo error of the results should be considered, where Monte Carlo errors are defined as the standard deviation across repeated runs of the same imputation procedure of the same data. This tends to zero as M increases.

Bodner (2008) proposed that F can be estimated as the fraction of incomplete cases, leading to a rule of thumb that the number of imputations should be similar to the percentage of cases that are incomplete. This rule of thumb is understandably not universally appropriate and individual settings need to be considered before deciding how to choose the number of imputations. A suggestion by White et al. (2011) was to impute a larger number of datasets but only use a portion of them in analyses, however, for $F < 0.5$, Bodner's rule of thumb may be appropriate. In practice, choice of number of imputations needs to be determined by what is feasible and practical based upon the size of the data set, the amount of missing information and the computational resources available (Azur et al., 2011).

Another consideration is the number of iterations to cycle over within the MICE procedure. This needs to be sufficient such that it can be assumed the algorithm has converged to a stationary distribution, where Bartlett et al. (2015) highlight that a larger number of covariates with missing values results in the need to run the procedure over a higher number of iterations to stabilise the results. Assessment of convergence can be done through examination of plots of the means, by iteration number, of the variables which were partially observed (Bartlett et al., 2015).

3.3.4 Model Selection after Multiple Imputation

Model selection after multiple imputation can be carried out using backwards elimination as outlined in Section 3.2.5, with the multivariable Wald test used to assess the significance of inclusion or exclusion of variables.

Let $\tilde{\beta}$ denote the average of the M estimates, $\hat{\beta}_m$, as given in Equation (3.12). Denote the variance estimate, $\text{Var}(\tilde{\beta})$, given in Equation (3.13), as $\tilde{V}$, where $\tilde{V}$ incorporates both the within-imputation variance, W, and the between-imputation variance, B. It has been suggested that if the number of imputations, M, is small, then the estimates of the between-imputation variance, B, may be unstable thus making $\tilde{V}$ unreliable. It was proposed by Li et al. (1991b) that, under the assumption of B and W being proportional to each other, a more stable estimate of $\tilde{V}$ can be given as

$$\tilde{V} = (1 + \bar{r})W,$$

where $\bar{r} = (1 + \frac{1}{M})\text{tr}(BW)/q$. This bypasses the need for B, as $\bar{r}$ is considered a good overall measure since the assumption of proportionality between B and W is equivalent to assuming equal fractions of missing data. The Wald test statistic is then defined as

$$T_\omega = \frac{(\tilde{\beta} - \beta^{(0)})'\tilde{V}^{-1}(\tilde{\beta} - \beta^{(0)})}{q},$$

and the p-value for T_ω is given as

$$P_\omega = 1 - F^{-1}_{q,\nu_\omega}(T_\omega),$$

where F^{-1}_{q,ν_ω} is the inverse cumulative distribution function of the F-distribution with q and ν_ω degrees of freedom (van Buuren, 2012).

Considering the degrees of freedom, based on large samples Li et al. (1991b) suggested the degrees of freedom ν_ω be given as

$$\nu_\omega = \begin{cases} 4 + (\eta - 4)[1 + (1 - 2\eta^{-1})\bar{r}^{-1}]^2, & \text{if } \eta = q(M-1) > 4 \\ \eta(1 + q^{-1})(1 + \bar{r}^{-1})^2/2, & \text{otherwise.} \end{cases}$$

An alternative degrees of freedom can be found in Reiter (2007), which was developed for use in smaller samples, and used similar ideas to Barnard and Rubin (1999).

For completeness, note that it is also possible to use the likelihood ratio test, however, as long as W and B are available, van Buuren (2012) suggests it is often more convenient to perform the Wald test. Further, it is possible to pool χ^2-statistics and associated p-values using a procedure outlined by Li et al. (1991a) and Rubin (1987), however results from pooling χ^2-statistics are considerably less reliable and thus should only be used if B and W are unattainable, or only χ^2 statistics are available. Details of these tests can be found in van Buuren (2012).

Wood et al. (2008) have also suggested a less computationally intensive approach to model selection after multiple imputation. This alternative approach involves stacking the imputed data sets to obtain one large data set. Weighted regression can then be applied to the single stacked data set to obtain valid parameter estimates. The weighting is important to correct for standard errors which would otherwise be too small. Although this method can provide a simpler approach compared to the use of Rubin's rules, Wood et al. (2008) highlight several situations under which this approach can have substantially inflated Type 1 error. The Rubin's rules approach remains the gold standard under most circumstances.

3.4 Conclusion

This chapter has provided an outline of methods which are applied to the stroke audit data in Chapter 4 as an intitial analysis, and are the basis for the methodological developments made in subsequent chapters. Notation and techniques for the analysis of survival data have been introduced, outlining data exploration techniques alongside use of the Cox proportional hazards model and assessment of model fit. Further, this chapter has introduced the issue of missing data and reviewed methods for exploring and handling missing information. Multiple imputation using chained equations was discussed in depth, outlining how the imputation procedure can be carried out, and how survival analysis techniques can be applied to multiply imputed data.

Chapter 4

Application to Stroke: Part 1

4.1 Introduction

This chapter presents an initial analysis of the stroke audit data, where the methods outlined in Chapter 3 are applied to the stroke audit data described in Section 2.4.

Firstly this chapter presents the data exploration, where the initial assessment of covariate effects on survival post-stroke is given, alongside examination of associations between the baseline covariates. The missing data is examined in depth, where patterns and potential reasons for missingness are explored to conclude an appropriate classification of MAR.

Further, this chapter outlines the application of the MICE imputation procedure, giving the results of the validation of the imputed values. The building of the analysis model is presented, where the results of the modelling procedure are displayed alongside interpretation of the parameter estimates in the context of the survival of stroke patients.

Finally, this chapter presents validation of the modelling procedure, where issues are highlighted to motivate the remainder of this thesis.

4.2 Exploration of Data

The first steps in the exploration of the data were to examine the percentage and frequency of patient deaths for each of the baseline measures, explore the amount of missing data within each variable and plot the KM estimates for the survivor function for each of the variables. Using the KM survival curves, alongside tables showing the incidence of death by variable, we can gain an initial understanding of potential relationships between the baseline measures and mortality post-stroke.

Figure 4.1 gives the overall KM survival curve, and shows that only 30% of patients survive beyond 5 years post-stroke, where 50% of patients died within the first 500 days post-stroke, and around 40% had died within the first 100 days post-stroke. The KM curve in Figure 4.1 drops very steeply initially, with around 30% of patients dead within the first month post-stroke; the curve becomes less steep between 100 and 1000 days post-stroke, but the survival rate drops from 0.6 to 0.4 within this interval. The curve then levels slightly after 1000 days, with only a drop of 0.1 in survival rate between 1000 and 2000 days post-stroke.

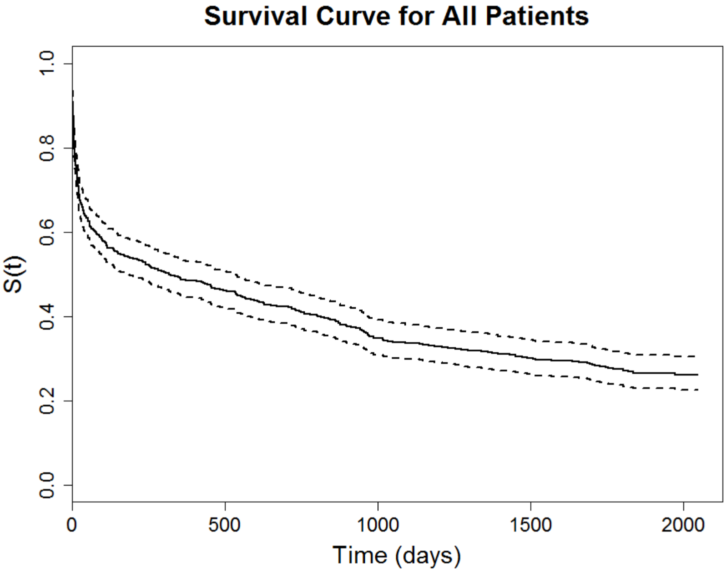


Figure 4.1: Overall Kaplan-Meier survival curve for all patients, showing the probability of survival over 5 years of follow-up. Dotted lines indicate the 95% confidence interval.

Figures 4.2, 4.3, 4.4 and 4.5 show the KM survival curves for a selection of the baseline measures. Tables 4.1, 4.2, 4.3, 4.4, 4.5, 4.6, 4.7 and 4.8 give the spread of patients across the levels of each variable, giving the number (n) and percentage (%) of patients of the total number of patients. These tables also show the incidence of death for each variable, giving the amount (n) of patients who died within each level, and the percentage (%) who died out of the total number of patients within that level. For any variables with missing data, 'Missing' is included as an extra level in the tables. These tables give an initial indication of possible relationships between the baseline measures and incidence of death, alongside the amount of missing observations.

We can see in Table 4.1 that the older patient groups have higher proportions of death, with the exception of the 60-70 group which had the lowest percentage of deaths of all the age groups. Figure 4.2(a) further highlights this, where the curves show that increasing age gives lower survival rates, however initially the survival rates for the 50-60 (red) age group drops more rapidly compared to the 60-70 (green) and 70-80 (dark blue) age groups. Considering sex, Table 4.1 shows that females had a higher rate of incidence of death compared to males, further shown in Figure 4.2(b), where the survival curve for females (red) is consistently lower.

Both smoking status and alcohol consumption gave an unexpected result, where smokers and patients which regularly consume alcohol had the lowest rate of incidence of death compared to the other groups for these variables. Table 4.1 shows however that those missing information about these behaviours have the highest death rate. Figures 4.2(c) and 4.2(d) further highlight this, and also reiterate, respectively, that non-smokers and non-drinkers have worse survival than those known to smoke or drink. It is well known that smoking and excessive consumption of alcohol can lead to many health problems and thus would be expected to increase the incidence of death. This indicates that smoking status and alcohol consumption may be associated with other risk factors for survival post-stroke and further exploration of this is needed.

Table 4.1: Patient characteristics: Number and percentage of patients within each level of the patient characteristics variables and the number and percentage of patients who died within each level.

	Spread		**Died**	
Variable	n	%	n	%
Sex				
Male	249	46.3	160	64.3
Female	289	53.7	219	75.8
Age (Years)				
0-50	18	3.3	9	50.0
50-60	36	5.7	21	58.3
60-70	118	21.9	57	48.3
70-80	196	36.4	146	74.5
80-90	147	27.3	124	84.3
90-100	23	4.3	22	95.7
Smoker				
Yes	112	20.8	68	60.7
No	170	31.6	118	69.4
Ex-smoker	133	24.7	88	66.2
Missing	123	22.9	105	85.4
Alcohol Consumption				
Excessive	18	3.3	10	55.6
Regular	46	8.6	24	52.2
Occasional	87	16.2	51	58.6
None	246	45.7	175	71.1
Missing	141	26.2	119	84.4

Table 4.2 shows that the proportion of deaths is higher for worsened pre-stroke measures; living conditions, mobility and modified Rankin score. In the worst cases for each of these, over 90% of patients died. Those living at home with a companion had a slightly lower rate of incidence of death compared to those living home alone prior to stroke, but there was a clear ordering of survival rates for pre-stroke Rankin and pre-stroke mobility. Figure 4.3(a) clearly demonstrates the

Table 4.2: Patient characteristics: Number and percentage of patients within each level of the patient characteristics variables and the number and percentage of patients who died within each level.

	Spread		**Died**	
Variable	*n*	%	*n*	%
Pre-Stroke Living Conditions				
Home Alone	191	35.5	132	69.1
Home with Companion	293	54.5	195	66.6
Institution	54	10.0	52	96.3
Pre-stroke Mobility				
Need Help	50	9.3	49	98.0
Indoors	203	37.7	158	77.8
200m Outside	246	45.7	141	57.3
Missing	39	7.3	31	79.5
Pre-stroke Rankin				
No symptoms at all	213	39.6	126	59.2
No significant disability despite symptoms	103	19.1	67	65.0
Slight disability	93	17.3	67	72.0
Moderate disability	63	11.7	58	92.1
Moderately severe disability	34	6.3	33	97.1
Severe disability	10	1.9	10	100.0
Missing	22	4.1	18	81.8

difference in survival between the levels of pre-stroke mobility, where those capable of walking 200 metres outdoors (black) have the highest survival rates. Between each of the survival curves in Figure 4.3(a), there is around a 20% difference in survival rates, with the worst survival rate for patients that need help moving around.

Table 4.3 shows the incidence of death related to the patients' previous medical conditions, and shows that for each of the medical conditions, the percentage of deaths is higher for the patient groups who have suffered from the medical condition previously. The incidence of death is around 75-80% for patients who have previously suffered from a stroke or myocardial infarction, or have diabetes mellitus, atrial fibrillation or angina, compared to a 62-65% incidence of death for those

Table 4.3: Medical history: Number and percentage of patients within each level of the previous health conditions variables and the number and percentage of patients who died within each level.

	Spread		Died	
Variable	n	%	n	%
Diabetes Mellitus				
Yes	68	12.6	52	76.5
No	307	57.1	197	64.2
Missing	163	30.3	130	79.8
Previous Stroke				
Yes	152	28.3	115	75.7
No	213	39.6	138	64.8
Missing	173	32.1	126	72.8
Previous TIA				
Yes	72	13.4	51	70.8
No	156	29.0	97	62.2
Missing	310	57.6	231	74.5
Atrial Fibrillation				
Yes	101	18.8	78	77.2
No	224	41.6	141	62.9
Missing	213	39.6	160	75.1
Angina				
Yes	91	16.9	74	81.3
No	142	26.4	88	62.0
Missing	305	56.6	217	71.1
Hypertension				
Yes	193	35.9	130	67.4
No	168	31.2	110	65.5
Missing	177	32.9	139	78.5
Peripheral Vascular Disease				
Yes	12	2.2	8	66.7
No	38	7.1	25	65.8
Missing	488	90.7	346	70.9
Myocardial Infarction				
Yes	67	12.5	52	77.6
No	302	56.1	193	63.9
Missing	169	31.4	134	79.3

that do not have a history of these medical conditions. The KM survival curve by previous stroke is shown in Figure 4.3(b), and shows that patients who have previously suffered from a stroke have lower survival rates overall, but this difference in survival rates increases at around 1000 days post-stroke. Figures 4.3(c) and 4.3(d) show the KM survival curves for diabetes and atrial fibrillation respectively. These survival curves show that patients who have diabetes or atrial fibrillation have lower survival rates overall. The difference in the rate of incidence of death between whether or not the patients have these conditions is smaller for hypertension, PVD and previous TIA, however Table 4.3 highlights that there is a high percentage of missingness in the variables PVD and previous TIA. All the variables in Table 4.3 have over 30% missing, and there is generally a higher proportion of deaths within the groups of patients missing information about these medical conditions. Table 4.4 suggests little difference in incidence of death between patients who have previously taken anti-hypertensives, anti-platelets and anti-coagulants, compared to those who have not, however the proportion of deaths is higher in the groups missing information about previous treatments.

In terms of admission details, Table 4.5 shows that Hospital 1 had a slightly lower rate of death compared to Hospital 2. Figure 4.4(a) also shows a slightly lower rate of survival for patients admitted to Hospital 2. Arm and leg weakness at admission gave similar results in Table 4.5, where patients with no movement had the highest rate of death for both these measures. Figure 4.4(b) further highlights the difference in survival for patients with no leg movement on admission to hospital, where the KM survival curve for patients with no leg movement (green) is lower, and drops more steeply initially compared to patients with some or no leg weakness. Patients with hypertension at admission to hospital had a lower incidence of death compared to those that did not, 62% compared to 76% respectively as stated in Table 4.5. Figure 4.4(c) further highlights this difference in survival, where the survival curve for patients with hypertension at admission

Table 4.4: Patient medication: Number and percentage of patients within each level of the previous medications variables and the number and percentage of patients who died within each level.

	Spread		**Died**	
Variable	*n*	%	*n*	%
Previous anti-hypertensives				
Yes	161	29.9	110	68.3
No	323	60.1	225	69.7
Missing	54	10.0	44	81.5
Previous anti-platelets				
Yes	159	29.6	111	69.8
No	334	62.1	230	68.9
Missing	45	8.4	38	84.4
Previous anti-coagulants				
Yes	28	5.2	18	64.3
No	465	86.4	324	69.7
Missing	45	8.4	37	82.2

(red) is consistently above the curve for no hypertension. Considering systolic BP at admission, we can see in Figure 4.4(d) that again patients with a higher blood pressure at admission to hospital seem to have better survival overall, where the survival curve for systolic BP in the range 170 to 200mmHg is highest overall. Patients with highest systolic BP, of the range 200 to 260mmHg, however, had worse survival than those in the 170 to 200mmHg range, suggesting the possibility of a non-linear effect of systolic BP on survival.

Considering the characteristics of stroke shown in Table 4.6, we can see that unconscious and unclassified patients had the highest incidence of deaths compared to the remaining OCSP classifications. Figure 4.5(b) shows the KM survival curves for the OCSP classification of stroke. We can see in Figure 4.5(b) that the most distinct and lowest survival curve is for patients that were unconscious. The survival curve for unclassified patients is initially close to the survival curve for those classified as TACS, however at 800 days post-stroke, the curve for unclassified patients drops steeply and becomes closer to the survival curve for unconscious

Table 4.5: Admission details: Number and percentage of patients within each level of the admission details variables and the number and percentage of patients who died within each level.

	Spread		**Died**	
Variable	n	%	n	%
Hospital				
Hospital 1	270	50.2	184	68.1
Hospital 2	268	49.8	195	72.8
Hypertension				
Yes	220	40.9	137	62.3
No	314	58.4	239	76.1
Missing	4	0.7	3	75.0
Arm Weakness				
No deficit	101	18.8	64	63.4
Weakness	285	53.0	193	67.7
No movement	125	23.2	99	79.2
Missing	27	5.0	23	85.2
Leg Weakness				
No deficit	111	20.6	67	60.4
Weakness	281	52.2	196	69.8
No movement	118	21.9	92	78.0
Missing	28	5.2	24	85.7

patients. The survival curves for TACS, PACS, LACS and POCS are quite close together in Figure 4.5(b), with some crossing of the curves. Table 4.6 also shows that patients with a lesion on the right side of their brain had the lowest incidence of death. Of patients that had a lesion on both sides, or were missing information about side of lesion, 100% died. In Figure 4.5(d), we can see that the survival curve for patients with a lesion on both sides drops steeply, and all these patients had died by 250 days post-stroke. The survival curves in Figure 4.5(d) are close together for no lesion or a lesion on one side, either left or right, with the right lesion having the highest survival rates overall.

The incidence of death was highest for the group of patients that did not have a CT scan; Table 4.6 shows that 30% of patients did not receive a CT scan, of

Table 4.6: Classification and lesion details of stroke: Number and percentage of patients within each level of these variables and the number and percentage of patients who died within each level.

	Spread		**Died**	
Variable	n	%	n	%
Class of Stroke				
Unclassified	20	3.7	16	80.0
TACS	53	9.9	36	67.9
PACS	156	29.0	103	66.0
LACS	139	25.8	88	63.3
POCS	17	3.2	10	58.8
Unconscious	97	18.0	87	89.7
Missing	56	10.4	39	69.6
Side of Lesion				
No lesion	112	208	79	70.5
Right	198	36.8	130	65.7
Left	204	37.9	146	71.6
Both	16	3.0	16	100.0
Missing	8	1.5	8	100.0
CT Scan Results				
No lesion	108	20.1	58	53.7
CI	183	34.0	121	66.1
HCI	17	3.2	13	76.5
PICH	56	10.4	39	69.6
No scan	174	32.3	148	85.1

which, 85% died. Patients with no lesion found in the CT scan had the lowest incidence of death. Figure 4.5(c) further highlights this, and also shows that of the lesion types shown in a CT scan, patients with a CI had the highest survival rates. Initially patients with a HCI had better survival compared to those with a PICH, however at around 400 days post-stroke the survival curves for PICH and HCI cross. This means that at the end of the 5 year follow-up, PICH patients had a higher rate of survival compared to those with a HCI.

Tables 4.7 and 4.8 give details of the incidence of death dependent upon symptoms in the first 24 hours after onset of stroke. Worst consciousness level is the

Table 4.7: Symptoms in first 24 hours after stroke onset: Number and percentage of patients within each level of these variables and the number and percentage of patients who died within each level.

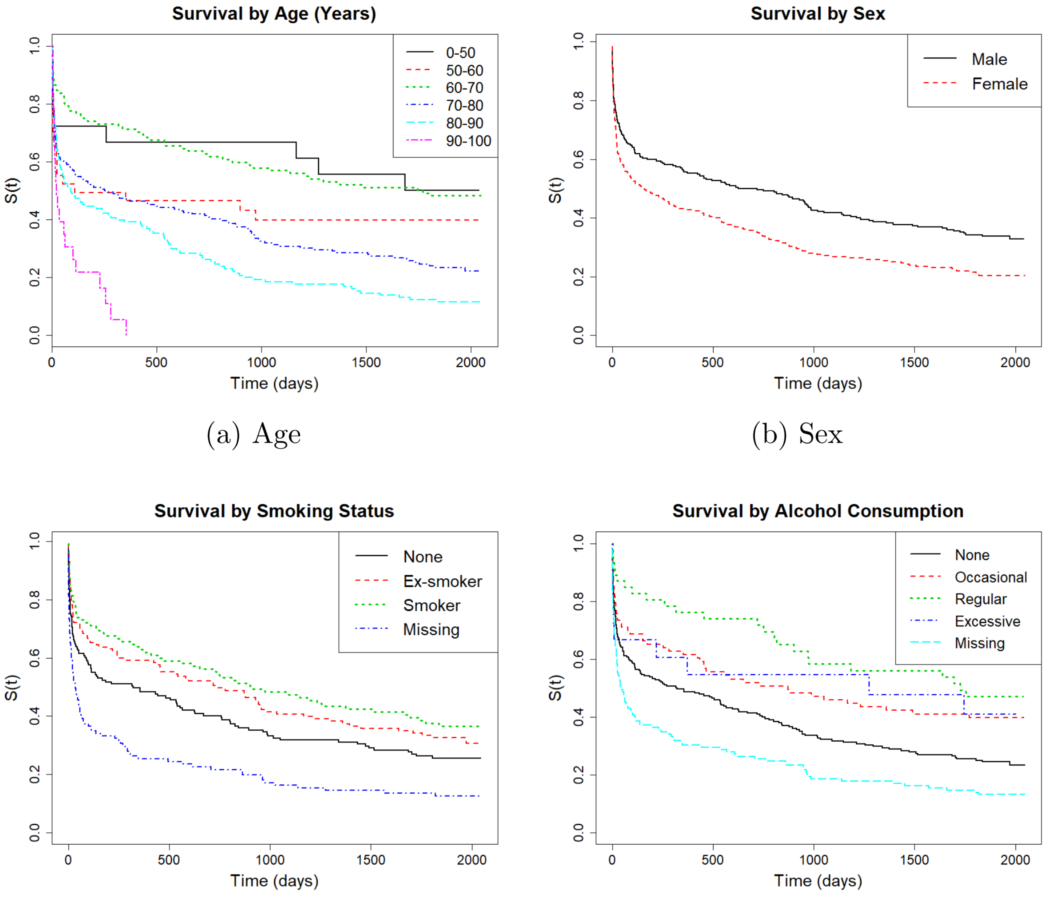
	Spread		**Died**	
Variable	*n*	%	*n*	%
Worst Consciousness Level				
Alert	368	68.4	222	60.3
Drowsy	60	11.2	49	81.7
Stupour	39	7.2	33	84.6
Coma	81	15.1	75	92.6
Facial Weakness				
Yes	190	35.3	143	75.3
No	215	40.0	129	60.0
Missing	133	24.7	107	80.5
Arm Weakness				
Yes	423	78.6	302	71.4
No	70	13.0	38	54.3
Missing	45	8.4	39	86.7
Leg Weakness				
Yes	406	75.5	293	72.2
No	85	15.8	45	52.9
Missing	47	8.7	41	87.2
Dysphasia				
Yes	164	30.5	116	70.7
No	219	40.7	140	63.9
Missing	155	28.8	123	79.4
Dysarthria				
Yes	159	29.6	107	67.3
No	215	40.0	144	67.0
Missing	164	30.5	128	78.0

only complete variable within these tables, and for the remaining variables, the groups of patients missing information about these symptoms mostly had the highest incidence of death. Table 4.7 shows that worse levels of consciousness gave a higher incidence of death, where over 90% of comatose patients died compared to 60% of alert patients. Figure 4.5(a) demonstrates the differences in survival for

Table 4.8: Symptoms in first 24 hours after stroke onset: Number and percentage of patients within each level of these variables and the number and percentage of patients who died within each level.

	Spread		**Died**	
Variable	*n*	%	*n*	%
Confusion				
Yes	81	15.1	54	66.7
No	222	41.3	131	59.0
Missing	235	43.7	194	82.6
Conjugate Gaze Paresis				
Yes	58	10.8	50	86.2
No	119	22.1	64	53.8
Missing	361	67.1	265	73.4
Hemianopia				
Yes	28	5.2	18	64.3
No	87	16.2	45	51.7
Missing	423	78.6	316	74.7
Sensory Inattention				
Yes	82	15.2	55	67.1
No	138	25.7	77	55.8
Missing	318	59.1	245	77.0
Brainstem/cerebellar signs				
Yes	34	6.3	25	73.5
No	213	39.6	141	66.2
Missing	291	54.1	213	73.2
Other deficit				
Yes	7	1.3	6	85.7
No	250	46.5	169	67.6
Missing	281	52.2	204	72.6

varying levels of worst consciousness within the first 24 hours post-stroke, where there is a clear distinction between the survival curves for each level of consciousness; lower for worsening consciousness levels. Table 4.7 shows that the groups of patients who experienced facial, arm or leg weakness had more than a 15% higher incidence of death compared to patients that did not have the specified weakness. Table 4.7 shows that patients who had dysarthria had a 7% higher incidence of death compared to those who did not, however the incidence of death was very similar regardless of whether or not patients had dysphasia. The symptoms of stroke shown in Table 4.8 all have a large proportion of missing values, varying between 43% and 78%, where patients missing information about these symptoms had the highest incidence of death. For each of these symptoms, the rate of incidence of death is lower for patients that did not experience the symptom being considered, but considering the frequencies of patients identified as having one of these symptoms, and the amount of missing information for each, these symptoms may be uncommon or difficult to identify.



(a) Age

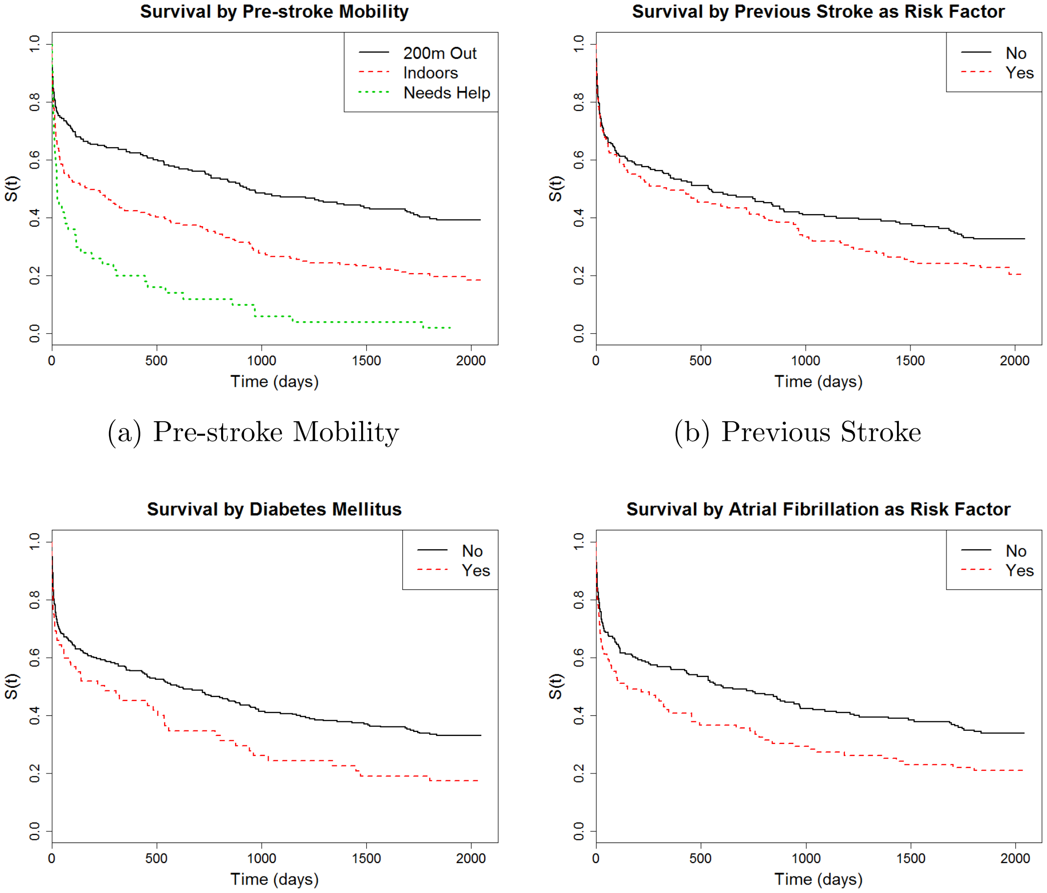
(b) Sex

(c) Smoking Status

(d) Alcohol Consumption

Figure 4.2: Plots of the Kaplan-Meier survival curves, split by patient characteristics: age, sex, smoking status and alcohol consumption.

The frequency tables and KM survival curves gave an initial indication of potential relationships between the baseline measures and survival. The next step



(a) Pre-stroke Mobility

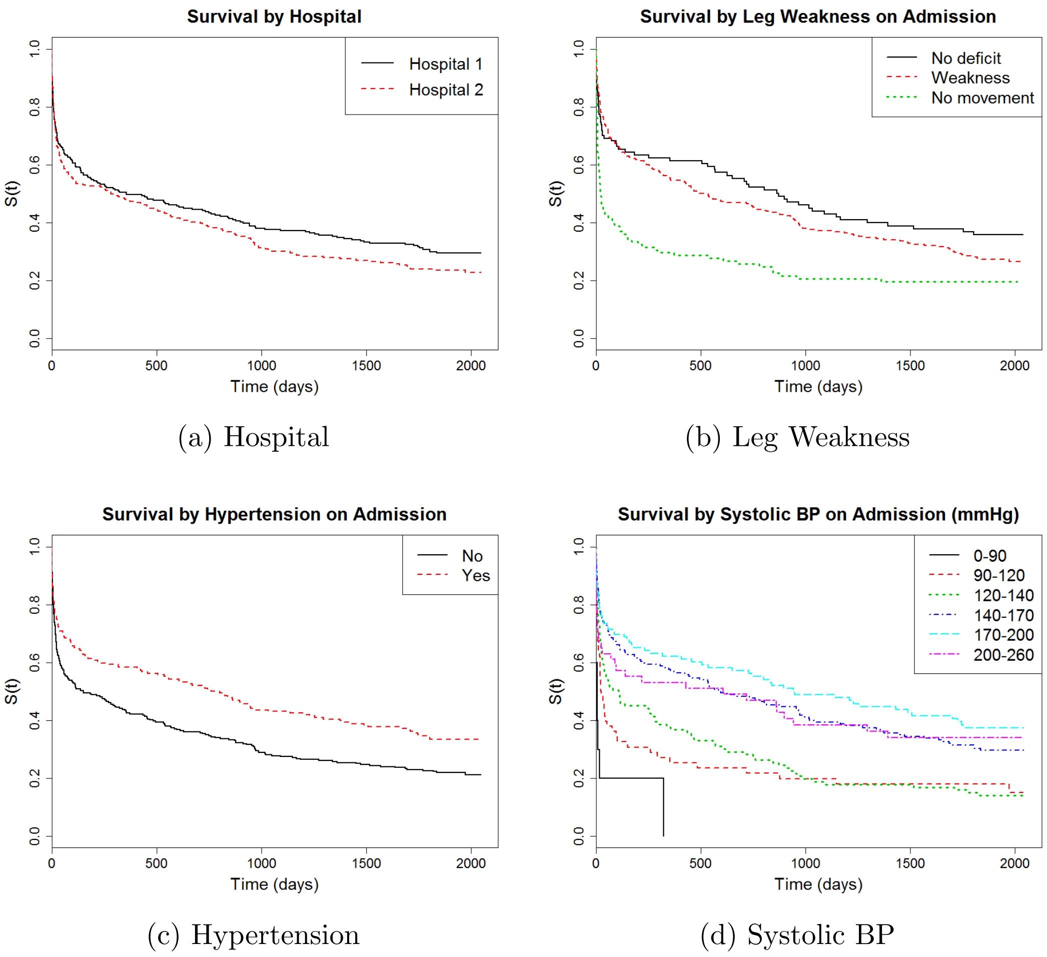
(b) Previous Stroke

(c) Diabetes Mellitus

(d) Atrial Fibrillation

Figure 4.3: Plots of the Kaplan-Meier survival curves, split by patient health prior to stroke: pre-stroke mobility, previous stroke, diabetes mellitus and atrial fibrillation.

within the data exploration was to carry out log-rank tests to establish if any of the baseline measures have a significant effect on survival. The results of the log-rank tests can be seen in Tables 4.9 and 4.10. Table 4.9 shows that age and sex both have a significant effect on survival post-stroke at the 5% level. Linking this back to the Kaplan-Meier curves shown in Figures 4.2(a) and 4.2(b) for age and sex respectively, we can see that older patients have significantly worse survival rates, as do females. The log-rank test results for smoking status and alcohol consumption, in Table 4.9, along side the KM survival curves in Figures 4.2(c) and 4.2(d), show that smoking and regular consumption of alcohol significantly reduces the incidence of death in stroke patients. Pre-stroke living conditions, mobility and



(a) Hospital

(b) Leg Weakness

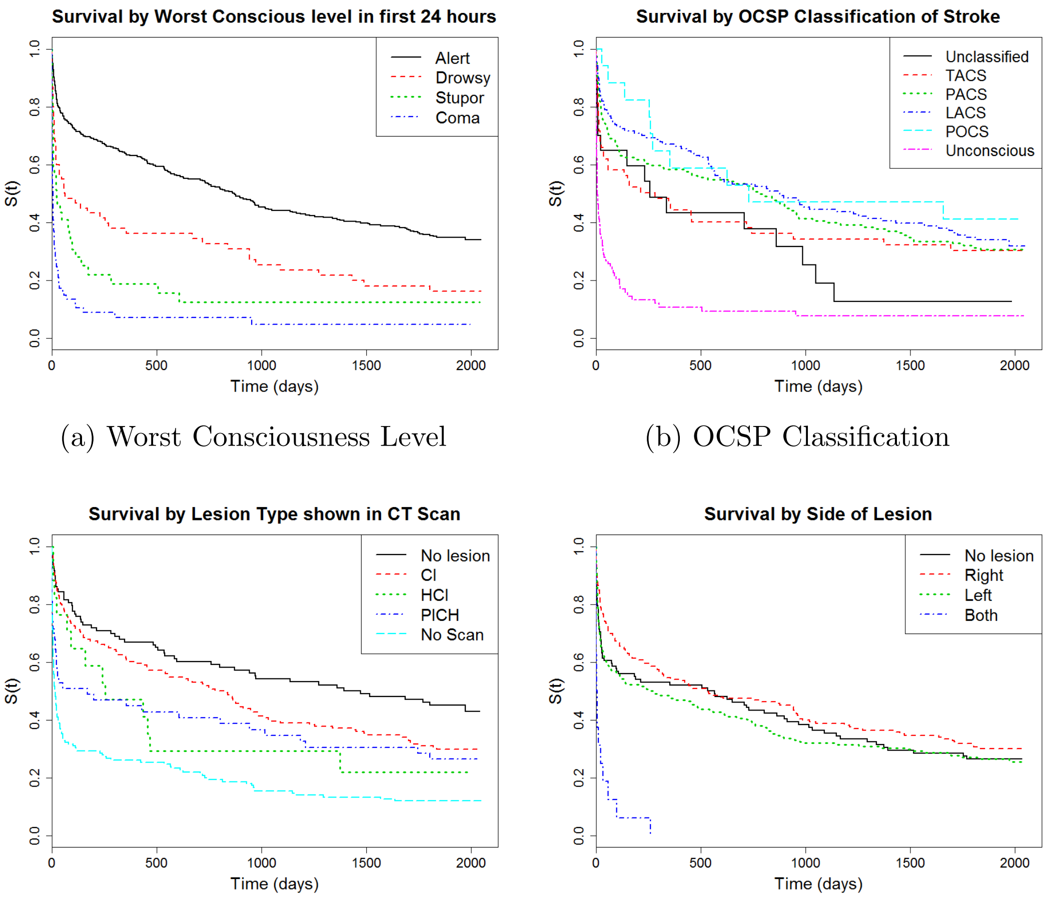
(c) Hypertension

(d) Systolic BP

Figure 4.4: Plots of the Kaplan-Meier survival curves, split by admission details and symptoms: hospital admitted to, and leg weakness, hypertension and systolic BP at admission.

modified Rankin score all had a significant effect on survival. The results for these pre-stroke measures in Table 4.9 alongside the KM survival curves suggest that survival rates are significantly reduced by worsened levels of these pre-stroke measures, at the 5% significance level, where Figure 4.3(a) shows a consistently large difference in survival rates for the levels of pre-stroke mobility.

In terms of medical history, the log-rank test results in Table 4.9 show that, at the 5% significance level, diabetes mellitus, angina and atrial fibrillation significantly reduce survival rates post-stroke. Though not significant at the 5% level, previous stroke and myocardial infarction were significant at the 10% level, suggesting history of these also reduces survival rates post-stroke. History of hypertension



(a) Worst Consciousness Level

(b) OCSP Classification

(c) CT Scan Results

(d) Side of Lesion

Figure 4.5: Plots of the Kaplan-Meier survival curves, split by characteristics of stroke: worst consciousness level in first 24 hours after stroke onset, OCSP classification of stroke, lesion type shown in CT scan and side of lesion.

or previous TIA did not give a significant effect on survival at the 5% level, neither did PVD, however PVD has a very high proportion of missing values so this result may not be representative of the true effect of PVD on survival post-stroke. Table 4.9 also shows that previous consumption of anti-hypertensive, anti-platelet or anti-coagulant medication does not significantly affect survival post-stroke.

Table 4.10 gives the log-rank test results for hospital admission details and stroke event assessments. Considering a significance level of 5%, firstly, Table 4.10 shows that the hospital patients were admitted to did not have a significant effect on survival post-stroke. Hypertension, arm weakness and leg weakness all have a significant affect on survival post-stroke. Linking this back to the KM survival

Table 4.9: Results of the log-rank tests; showing the χ^2 values, degrees of freedom (df) and p-values for each of the baseline measures related to the patient and their medical history.

Variable	χ^2	df	p**-value**
Age	64.4	5	0.0005
Sex	12.2	1	<0.0001
Smoking Status	6.5	2	0.04
Alcohol Consumption	15.6	3	0.001
Pre-stroke Living Conditions	52.8	2	<0.0001
Pre-stroke Mobility	56.8	2	<0.0001
Pre-stroke Rankin	92.4	5	<0.0001
Diabetes Mellitus	6.4	1	0.01
Previous Stroke	3.3	1	0.07
Previous TIA	0.5	1	0.5
Atrial Fibrillation	6.7	1	0.01
Angina	10.9	1	0.001
Hypertension	0.5	1	0.5
Peripheral Vascular Disease	0.1	1	0.7
Myocardial Infarction	3.3	1	0.07
Previous Anti-hypertensives	0.7	1	0.4
Previous Anti-platelets	0.3	1	0.6
Previous Anti-coagulants	0.1	1	0.8

curves, we can see in Figure 4.4(c) that not having hypertension at admission significantly reduced survival rates, and Figure 4.4(b) shows that survival rates are significantly increased for those without any leg weakness at admission. The OCSP classification of stroke, along with side of lesion, lesion type shown in CT scan and worst consciousness level were all shown to be highly significant for survival in the log-rank test results given in Table 4.10. Arm and leg weakness within the first 24 hours post-stroke were also found to be significant for survival, but facial weakness was not. Conjugate Gaze Paresis was the only other symptom from the first 24 hours post stroke to be found significant for survival at the 5% level from the log-rank tests, however this variable has a high percentage of missingness so this result may not be representative of the true effect of CGP. Dysphasia, confusion

Table 4.10: Results of the log-rank tests; showing the χ^2 values, degrees of freedom (df) and p-values for each of the baseline measures related to hospital admission and stroke event assessments.

Variable	χ^2	df	p**-value**
Hospital	2.1	1	0.1
Hypertension (Admission)	11.1	1	0.0009
Arm Weakness (Admission)	27.2	2	<0.0001
Leg Weakness (Admission)	25.8	2	<0.0001
OCSP Classfication	122	5	<0.0001
Side of Lesion	45.6	3	<0.0001
CT Scan: Lesion Type	89.9	4	<0.0001
Worst Consciousness Level (24hrs)	192	3	<0.0001
Facial Weakness (24hrs)	0.8	1	0.4
Arm Weakness (24hrs)	9.7	1	0.002
Leg Weakness (24hrs)	12.5	1	0.0004
Dysphasia (24hrs)	3.1	1	0.08
Dysarthria (24hrs)	0	1	0.9
Confusion (24hrs)	3.7	1	0.06
Congugate Gaze Paresis (24hrs)	37.1	1	<0.0001
Hemianopia (24hrs)	1.9	1	0.2
Sensory Inattention (24hrs)	3.4	1	0.07
Brainstem/Cerebellar Signs (24hrs)	1.5	1	0.2
Other Deficit (24hrs)	1.4	1	0.2

and sensory inattention were marginally not-significant at the 5% level, but would be considered significant for survival at the 10% level, where presence of these symptoms would reduce survival rates.

After consideration of individual effects on survival, the next stage within the exploration of the data was to examine the relationships between the covariates, and how their effects on survival change when adjusted for the effects of other measures. Firstly, bivariate associations were explored through looking at measures of correlation between the baseline covariates. A visualisation of these correlations can be seen in Figure 4.6, which shows the size of the p-values through a colour scale. Darker squares represent smaller p-values and show stronger associations between the baseline measures.

Figure 4.6 highlights several strong associations between some groups of baseline measures. In particular, we can see that OCSP classification of stroke is strongly associated with many other baseline measures, including symptoms and worst consciousness level in the first 24 hours post-stroke, plus the pre-stroke measures and patient characteristics. The lesion type shown in CT scan is also associated with several of the patient characteristics and their medical history, along with admission assessments, worst consciousness level and the OCSP classification. There seems to be strong associations between the patients characteristics where age, sex, smoking status, alcohol consumption, and pre-stroke measures: living conditions, mobility and modified Rankin are all associated. Arm, leg and facial weakness are all strongly associated, and side of lesion is also strongly associated with each of these.

The association between OCSP classification and worst conscious level requires further consideration as these variables have cross-over between their levels. Both of these variable categorise patients by consciousness level, where those identified as stupor or comatose in the worst consciousness level variable are categorised as unconscious in the OCSP classification variable. This is likely to result in collinearity issues during model fitting, and is discussed further in Section 4.3.

Further to looking at the bivariate associations, as age and sex are complete measures and are already known to be risk factors for survival post-stroke, we carried out Cox proportional hazards modelling adjusted for age and sex on each of the baseline measures. Age was fitted as a continuous variable and, as a binary measure, sex was included with males as the baseline reference group. The results of the age and sex adjusted modelling are given in Tables 4.11, 4.12, 4.13, 4.13 and 4.15.

Firstly considering the patient characteristics, we can see in Table 4.11 that age and sex both have a significant effect on survival. The hazard ratio for males

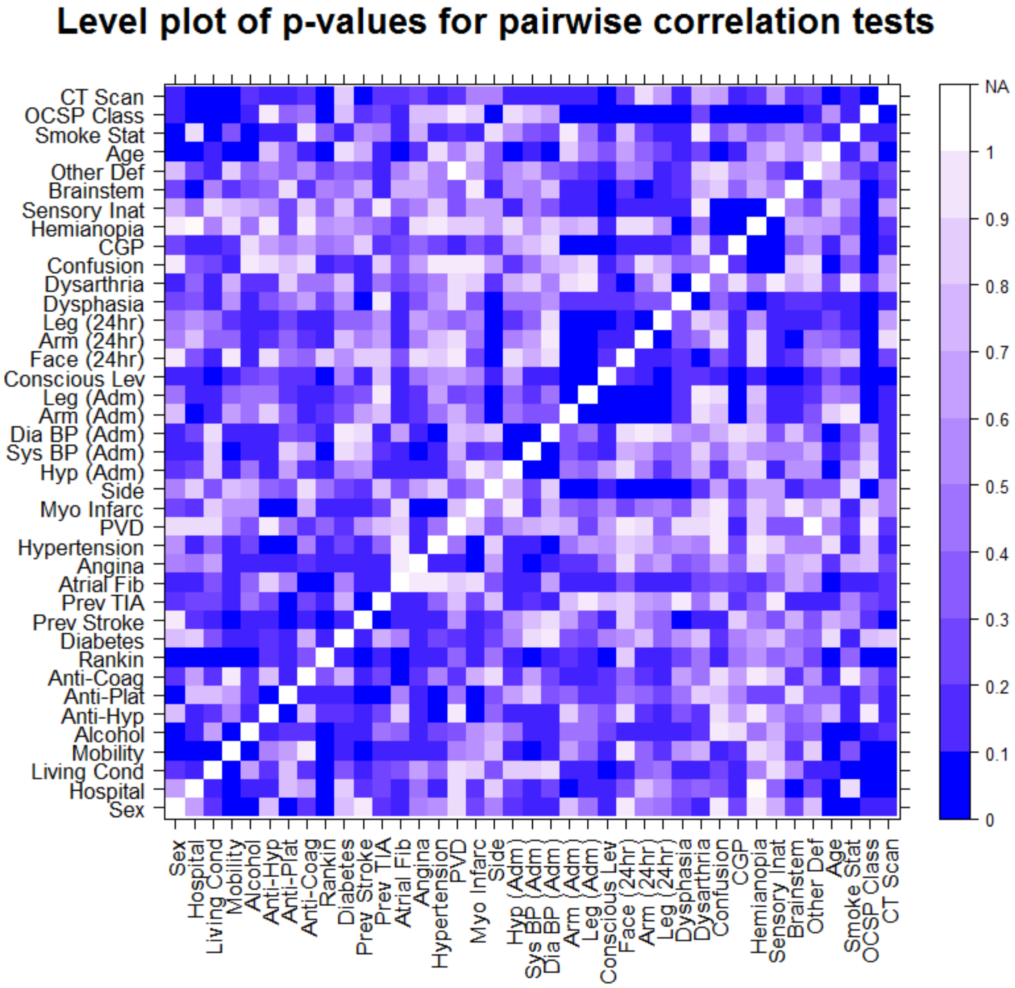


Figure 4.6: Level plot showing the p-values for each of the pairwise correlation tests.

compared to females is 1.44, however when adjusted for age the hazard ratio reduces to 1.26. Smoking status and alcohol consumption give some unexpected effects in the univariate model, where both smokers and regular consumers of alcohol have a reduced hazard of death compared to the baselines of non-smokers and non-drinkers, respectively. Table 4.11 suggests smoking significantly reduced hazard of death by around 30%, however, when smoking status is adjusted for age, this effect is no longer significant. Alcohol consumption also gave an interesting result, where regular consumption of alcohol halved the hazard of death compared to non-drinkers. Adjusting for both age and sex resulted in this effect no longer being significant however. This shows that the effect of smoking status and alcohol consumption on survival post-stroke may be confounded by an age or sex effect.

Considering pre-stroke living conditions, the univariate model in Table 4.11 shows that patients living at home with a companion have a lower hazard of death compared to the baseline of those living home alone. However, when this is adjusted for age, patients living home alone had the lowest hazard of death. The differences in effect on survival between living home alone or with a companion are not significant in any of the models shown in Table 4.11, where we can see the confidence intervals all span the reference hazard of one. Patients living in an institution prior to stroke were around 2 and a half times more likely to die post-stroke compared to those living home alone, regardless of age or sex. Poor mobility prior to stroke also increased hazard of death regardless of age or sex, although when this is adjusted for age, the hazard ratios for mobility levels of 'indoors' and 'needs help' are reduced slightly.

In terms of pre-stoke Rankin scores, we can see in Table 4.11 that a moderate to severe disability increased hazard of death significantly, irrespective of age or sex, where patients with a severe disability are over 6 times more likely to die following a stroke. The significance of the effect of a slight disability on survival post-stroke is reduced when adjusting for age or sex, or both.

Table 4.12 gives the results for the univariate, and age and sex adjusted Cox regression models for patient medical history. These results show that previous medications, such as anti-coagulants, were not important for survival post-stroke. Further, previous stroke or TIA did not have a significant effect on survival post-stroke in any of these models. Peripheral vascular disease also did not have a significant effect on survival post-stroke. Diabetes mellitus was shown to be important for survival post-stroke, where diabetic patients had almost a 50% increase in hazard of death post-stroke, regardless of age or sex. Table 4.12 also shows that atrial fibrillation increased hazard of death post-stroke, however this was not a significant increase when adjusted for age. The univariate model for myocardial infarction suggests it was not important for survival post-stroke, however when adjusted for sex, myocardial infarction was found to significantly increase hazard of death post-stroke by around 40%. Angina gave a similar result, where the univariate models suggest angina did not have a significant effect on survival post-stroke, but when adjusted for age, sex or both, angina was found to significantly increase the hazard of death in stroke patients by around 60%.

Considering hospital admission details, Table 4.13 shows that patients admitted to Hospital 2 had an increased hazard of death compared to the baseline of Hospital 1, however this was not significant. Higher blood pressure and hypertension at admission to hospital were shown to reduce hazard of death following stroke, and this effect remained significant regardless of age or sex. The results in Table 4.13 also show that patients with no arm movement or no leg movement on admission to hospital post-stroke were twice as likely to die, this is slightly reduced to a 80-90% increase in hazard of death when adjusted for age, sex or both, however the effect remained significant. Patients with a lesion on both sides of their brain had around 5 times the hazard of death post-stroke compared to those with no lesion. Adjusting for age increased this hazard, whereas adjusting for sex reduced the hazard ratio slightly.

Table 4.14 considers patient symptoms in the first 24 hours post-stroke. These results show that arm and leg weakness in the first 24 hours post-stroke both increased the hazard of death. In the univariate models the hazard of death is increased by around 70% for arm or leg weakness, when adjusting for age however this increase in hazard was reduced by around 15%. Conjugate gaze paresis (CGP) was also shown to significantly increase hazard of death, where patients experiencing CGP in the first 24 hours post-stroke were 3 times more likely to die. When adjusting for age however, this hazard ratio reduced to 2.6. The remaining symptoms considered in Table 4.14 were all shown to increase the hazard of death post-stroke, however, none of these effects were significant in any of the models.

Table 4.15 gives the results of the univariate and age and gender adjusted models for OCSP classification of stroke, worst consciousness level and lesion type shown in CT scan. The stroke classifications were compared to the baseline of LACS, and the results show that TACS and PACS increased the hazard of death, and POCS gave a lower hazard compared to LACS. However these effects were not significant within the Cox regression models given in Table 4.15. The models for classification of stroke show that unconscious patients had over 4 times the hazard of death compared to those with LACS, and this effect was significant regardless of age or sex. Considering worst consciousness level in the first 24 hour post-stroke, drowsy, stupor and comatose patients had a significantly increased hazard of death compared to alert patients, regardless of age and sex. Comatose patients had the highest hazard overall, where they were almost 6 times more likely to die post-stroke.

The model results in Table 4.15 for lesion type shown in CT scan show that, compared to no lesion, patients with a CI or HCI had an increased hazard in the univariate models, however when adjusted for age or sex, CI no longer had a significant effect on survival post-stroke, and the effect of HCI was no longer significant when adjusted for age. Patients with PICH had almost twice the hazard of those with no lesion, regardless of age and sex, and patients that did not receive a CT scan had the highest hazard overall, where they were around 3 times more likely to die post-stroke. The effect of no scan is likely to be capturing the effects of the underlying reasons as to why the patients did not receive a CT scan.

The results of the initial univariate, and age and sex adjusted Cox regression models show that many of the baseline variables had a significant effect on survival post-stroke, however, the size and importance of these effects changed on adjustment for other covariate effects. Therefore, it is important to consider the effects of these variables in a fully adjusted setting. Fitting a fully adjusted Cox model to the complete cases resulted in a model fitted to only 6 patients due to the amount of missing information. This shows that complete case analysis is inappropriate for the stroke audit data, and an alternative approach towards the missing data is needed to carry out an adjusted analysis.

Table 4.11: Results of the age and sex adjusted Cox PH modelling for patient characteristics; showing the hazard ratios (HR) and 95% confidence intervals (CI) for each model, displayed as HR (CI). (* Significant at 5% level)

Variable	Univariate	Age Adj.	Sex Adj.	Age & Sex Adj.
Age	1.038 (1.027,1.049)*		1.035 (1.024,1.047)*	
Sex	1.437 (1.171,1.764)*	1.258 (1.020,1.551)*		
Smoking Status				
Ex-smoker	0.788 (0.598,1.039)	0.847 (0.642,1.119)	0.837 (0.632,1.109)	0.881 (0.664,1.169)
Smoker	0.693 (0.514,0.934)*	0.875 (0.642,1.195)	0.720 (0.533,0.972)*	0.894 (0.654,1.222)
Alcohol Consumption				
Occasional	0.668 (0.489,0.913)*	0.831 (0.603,1.145)	0.701 (0.506,0.971)*	0.857 (0.613,1.198)
Regular	0.489 (0.319,0.750)*	0.603 (0.391,0.931)*	0.530 (0.336,0.835)	0.633 (0.399,1.003)
Excessive	0.697 (0.369,1.320)	0.897 (0.471,1.710)	0.755 (0.392,1.454)	0.942 (0.485,1.827)
Pre-stroke Living Conditions				
Home with Companion	0.911 (0.730,1.136)	1.145 (0.908,1.444)	0.975 (0.777,1.222)	1.201 (0.947,1.522)
Institution	2.737 (1.972,3.799)*	2.469 (1.775,3.435)*	2.681 (1.931,3.723)*	2.437 (1.751,3.391)*
Pre-stroke Mobility				
Indoors	1.745 (1.389,2.192)*	1.483 (1.174,1.872)*	1.691 (1.344,2.128)*	1.461 (1.155,1.847)*
Needs Help	3.209 (2.304,4.468)*	2.877 (2.060,4.019)*	3.102 (2.225,4.326)*	2.824 (2.019,3.950)*
Pre-stroke Rankin				
No Significant Disability	1.182 (0.879,1.590)	1.112 (0.826,1.496)	1.173 (0.872,1.577)	1.106 (0.822,1.488)
Slight Disability	1.391 (1.033,1.871)*	1.252 (0.929,1.686)	1.345 (0.999,1.812)	1.227 (0.910,1.655)
Moderate Disability	2.796 (2.041,3.830)*	2.492 (1.814,3.423)*	2.695 (1.965,3.697)*	2.443 (1.777,3.360)*
Moderate/Severe Disability	3.058 (2.076,4.506)*	2.943 (1.994,4.344)*	2.874 (1.945,4.246)*	2.851 (1.928,4.216)*
Severe Disability	7.072 (3.667,13.64)*	6.533 (3.382,12.62)*	6.732 (3.485,13.01)*	6.358 (3.287,12.30)*

Table 4.12: Results of the age and sex adjusted Cox PH modelling for patient medical history; showing the hazard ratios (HR) and 95% confidence intervals (CI) for each model, displayed as HR (CI). (* Significant at 5% level)

Variable	**Univariate**	**Age Adj.**	**Sex Adj.**	**Age & Sex Adj.**
Previous Anti-hypertensives	0.904 (0.720,1.135)	0.906 (0.720,1.139)	0.897 (0.714,1.126)	0.895 (0.712,1.126)
Previous Anti-platelets	0.942 (0.750,1.182)	0.981 (0.782,1.232)	1.005 (0.799,1.264)	1.020 (0.811,1.283)
Previous Anti-coagulants	0.933 (0.581,1.500)	1.091 (0.678,1.756)	0.958 (0.595,1.542)	1.094 (0.680,1.762)
Diabetes Mellitus	1.481 (1.091,2.011)*	1.441 (1.059,1.960)*	1.461 (1.076,1.983)*	1.425 (1.047,1.938)*
Previous Stroke	1.259 (0.983,1.611)	1.246 (0.973,1.595)	1.276 (0.995,1.637)	1.259 (0.981,1.614)
Previous TIA	1.131 (0.806,1.587)	1.084 (0.773,1.522)	1.092 (0.776,1.536)	1.053 (0.748,1.484)
Atrial Fibrillation	1.439 (1.092,1.897)	1.231 (0.927,1.636)	1.384 (1.046,1.832)	1.209 (0.910,1.607)
Hypertension	1.092 (0.846,1.409)	1.066 (0.826,1.375)	1.070 (0.829,1.381)	1.049 (0.813,1.354)
Myocardial Infarction	1.332 (0.981,1.809)	1.296 (0.954,1.759)	1.381 (1.015,1.879)*	1.326 (0.975,1.803)
Angina	1.170 (0.858,1.595)	1.587 (1.165,2.163)*	1.667 (1.223,2.272)*	1.568 (1.148,2.142)*
Peripheral Vascular Disease	1.160 (0.522,2.575)	1.534 (0.657,3.585)	1.121 (0.504,2.493)	1.381 (0.596,3.202)

Table 4.13: Results of the age and sex adjusted Cox PH modelling for hospital admission details; showing the hazard ratios (HR) and 95% confidence intervals (CI) for each model, displayed as HR (CI). (* Significant at 5% level)

Variable	**Univariate**	**Age Adj.**	**Sex Adj.**	**Age & Sex Adj.**
Hospital	1.161 (0.948,1.420)	1.045 (0.852,1.281)	1.169 (0.955,1.430)	1.055 (0.861,1.294)
Systolic BP	0.991 (0.987,0.995)*	0.992 (0.988,0.996)*	0.991 (0.987,0.995)*	0.992 (0.988,0.996)*
Diastolic BP	0.990 (0.984,0.996)*	0.994 (0.988,0.999)*	0.991 (0.985,0.997)*	0.994 (0.988,0.999)*
Hypertension	0.715 (0.580,0.882)*	0.763 (0.618,0.943)*	0.715 (0.579,0.884)*	0.768 (0.621,0.949)*
Arm Weakness				
Weakness	1.007 (0.759,1.335)	1.052 (0.794,1.395)	1.075 (0.870,1.328)	1.043 (0.786,1.383)
No Movement	1.950 (1.423,2.674)*	1.861 (1.357,2.551)*	1.919 (1.400,2.632)*	1.842 (1.344,2.526)*
Leg Weakness				
Weakness	1.170 (0.886,1.545)	1.119 (0.847,1.477)	1.150 (0.871,1.519)	1.104 (0.836,1.458)
No Movement	2.044 (1.491,2.803)*	1.896 (1.383,2.600)*	2.002 (1.460,2.744)*	1.874 (1.364,2.574)*
Side of Lesion				
Right	0.847 (0.640,1.121)	0.839 (0.635,1.108)	0.848 (0.641,1.122)	0.839 (0.635,1.108)
Left	1.043 (0.793,1.372)	1.092 (0.830,1.437)	1.054 (0.801,1.387)	1.092 (0.830,1.437)
Both	4.600 (2.662,7.947)*	5.145 (2.972,8.907)*	4.428 (2.558,7.666)*	4.973 (2.867,8.626)*

Table 4.14: Results of the age and sex adjusted Cox PH modelling for symptoms in the first 24 hours post-stroke; showing the hazard ratios (HR) and 95% confidence intervals (CI) for each model, displayed as HR (CI). (* Significant at 5% level)

Variable	**Univariate**	**Age Adj.**	**Sex Adj.**	**Age & Sex Adj.**
Facial Weakness	1.113 (0.878,1.411)	1.147 (0.903,1.457)	1.111 (0.876,1.408)	1.147 (0.903,1.457)
Arm Weakness	1.699 (1.213,2.380)*	1.560 (1.112,2.190)*	1.654 (1.178,2.321)*	1.537 (1.095,2.158)*
Leg Weakness	1.751 (1.279,2.396)*	1.586 (1.157,2.174)*	1.718 (1.255,2.350)*	1.570 (1.145,2.152)*
Dysphasia	1.249 (0.975,1.598)	1.186 (0.927,1.519)	1.214 (0.948,1.554)	1.171 (0.913,1.502)
Dysarthria	1.022 (0.795,1.314)	1.025 (0.798,1.318)	1.058 (0.821,1.362)	1.050 (0.816,1.352)
Confusion	1.363 (0.993,1.873)	1.129 (0.815,1.563)	1.390 (1.012,1.909)*	1.149 (0.828,1.594)
Conjugate Gaze Paresis	3.077 (2.108,4.492)*	2.593 (1.766,3.808)*	3.016 (2.062,4.412)*	2.570 (1.747,3.781)*
Hemianopia	1.462 (0.846,2.527)	1.353 (0.778,2.351)	1.426 (0.824,2.469)	1.347 (0.777,2.337)
Sensory Inattention	1.384 (0.978,1.958)	1.336 (0.945,1.891)	1.374 (0.971,1.944)	1.324 (0.936,1.874)
Brainstem/Cereballar Signs	1.310 (0.856,2.004)	1.347 (0.880,2.061)	1.366 (0.891,2.094)	1.380 (0.900,2.115)
Other Deficit	1.636 (0.724,3.696)	1.781 (0.786,4.032)	1.702 (0.752,3.855)	1.815 (0.801,4.110)

Table 4.15: Results of the age and sex adjusted Cox PH modelling for classification of stroke, worst consciousness level in the first 24 hours post-stroke and lesion type shown in CT scan; showing the hazard ratios (HR) and 95% confidence intervals (CI) for each model, displayed as HR (CI). (* Significant at 5% level)

Variable	**Univariate**	**Age Adj.**	**Sex Adj.**	**Age & Sex Adj.**
Class of Stroke				
PACS	1.131 (0.856,1.496)	1.109 (0.839,1.467)	1.101 (0.832,1.457)	1.092 (0.825,1.445)
POCS	0.811 (0.433,1.517)	0.823 (0.440,1.541)	0.826 (0.441,1.547)	0.831 (0.444,1.557)
TACS	1.390 (0.965,2.002)	1.338 (0.929,1.928)	1.338 (0.928,1.929)	1.315 (0.912,1.896)
Unclassified	1.675 (0.996,2.816)	1.497 (0.889,2.520)	1.579 (0.937,2.661)	1.447 (0.858,2.442)
Unconscious	4.528 (3.388,6.050)*	4.160 (3.109,5.567)*	4.375 (3.269,5.855)*	4.085 (3.050,5.470)*
Worst Conscious Level (24hrs)				
Drowsy	1.858 (1.363,2.533)*	1.858 (1.362,2.534)*	1.810 (1.327,2.469)*	1.832 (1.342,2.500)*
Stupor	2.860 (1.977,4.137)*	2.538 (1.752,3.675)*	2.779 (1.919,4.023)*	2.507 (1.730,3.632)*
Coma	5.752 (4.365,7.580)*	5.650 (4.285,7.450)*	5.669 (4.297,7.479)*	5.594 (4.240,7.381)*
CT Scan: Lesion Type				
CI	1.369 (1.001,1.873)*	1.305 (0.953,1.785)	1.312 (0.958,1.796)	1.273 (0.929,1.743)
HCI	1.873 (1.025,3.420)*	1.719 (0.940,3.141)	1.946 (1.065,3.556)*	1.785 (0.976,3.265)
PICH	1.918 (1.278,2.879)*	1.834 (1.221,2.754)*	1.828 (1.217,2.747)*	1.786 (1.189,2.684)*
No Scan	3.391 (2.496,4.608)*	2.875 (2.102,3.933)*	3.287 (2.417,4.469)*	2.846 (2.080,3.895)*

4.3 Exploration of Missing Data

As shown in the previous section, complete case analysis is clearly an inappropriate method of handling the missing data due to the amount of missing observations in the stroke audit data. As discussed in Section 3.3.2, multiple imputation is a flexible approach to handling missing data, and will be the method used for this data. Prior to the imputation procedure, we must first explore the missing data, considering the patterns and missing data mechanism, and possible predictors for the missing values. This exploration is needed to allow appropriate assumptions to be made about the missing data mechanism and ensure the missingness is handled appropriately.

The first step in exploring the missing data was to examine how much missing information there was in each variable. Table 4.16 shows the amount and percentage of patients missing information for each of the variables with missing values, ordered by increasing missingness. Age, sex, hospital, pre-stroke living conditions, worst consciousness level and admission date were completely recorded variables so are not shown in Table 4.16. There was one patient who was missing information regarding whether or not they had a CT scan; this patient was coded as not having had a CT scan given there was no lesion type results for them.

The missing values for onset date were also handled during this exploration stage. Onset date was missing for 12 patients, however, as survival time was calculated as the time to death from onset of stroke, missing onset dates were imputed as their admission dates. This approach was taken as the majority of patients had the same admission and onset dates, and missing survival outcomes further complicates any possible imputation procedures.

A further issue needing consideration at this stage was the OCSP classification variable. Not only did this variable have missing data, but also collinearity issues with another variable. Firstly we consider the missing data. Table 4.16 shows that 56 patients were missing OCSP classification, however, as some patients were under the group ‘unclassified’ for this variable, the difference between ‘unclassified’ and ‘missing’ was examined. Further exploration of these patients showed that all the patients missing OCSP classification were diagnosed with a primary intra-cerebral haemorrhage (PICH) in their CT scan. The OCSP classification groups are not applicable to classify PICH patients, therefore imputation would not be a feasible approach for these patients. Exclusion of these patients was also an undesirable option, giving limited options for handling these patients.

The collinearity issues related to consciousness level, where patients classified as ‘unconscious’ within the OCSP classification variable directly corresponded to patients who were categorised as as ‘stupor’ or ‘coma’ within the variable for worst consciousness level in the first 24 hours post-stroke. The collinearity between these variables meant both could not be included together in any models; imputation or analysis. This gave the options to either exclude one, or combine them into one variable. Considering the issues around missing classifications for PICH patients, and recalling the results of the Cox regression models in Table 4.15, it was decided to exclude OCSP classification from the imputation and analysis procedures, maintaining inclusion of worst consciousness level instead. Table 4.15 showed consciousness level to be more informative for survival over the differences between LACS, PACS, TACS and POCS, and this decision also enabled inclusion of patients with PICH.

Referring back to Table 4.16 to examine amount of missing data in the rest of the baseline variables, it can be seen that there are eight variables for which over 50% of patients had missing observations, all of which fall into two categories of variables: the symptoms present in the first 24 hours post-stroke and history of medical conditions known to be risk factors for stroke. Whether or not patients had a history of peripheral vascular disease had the highest amount of missing values, with data missing for over 90% of patients. As 50% missingness is a sensible cut off for imputation, the eight variables with over 50% missingness are to be excluded

Table 4.16: Amount of missingness in each incomplete baseline variable; showing number and percentage missing.

Variable	Number	Percentage (%)
Hypertension (Admission)	4	0.7
Systolic BP (Admission)	4	0.7
Diastolic BP (Admission)	4	0.7
Side of Lesion	8	1.5
Onset Date	12	2.2
Pre-stroke Rankin	22	4.1
Arm Weakness (Admission)	27	5.0
Leg Weakness (Admission)	28	5.2
Pre-stroke Mobility	39	7.2
Previous Anti-platelets	45	8.4
Previous Anti-coagulants	45	8.4
Arm Weakness (24hrs)	45	8.4
Leg Weakness (24hrs)	47	8.7
Previous Anti-hypertensives	54	10.0
OCSP Class of Stroke	56	10.4
Smoking Status	123	22.9
Facial Weakness (24hrs)	133	24.7
Alcohol Consumption	141	26.2
Dysphasia (24hrs)	155	28.8
Diabetes Mellitus	163	30.3
Dysarthria (24hrs)	164	30.5
Myocardial Infarction	169	31.4
Previous Stroke	173	32.2
Hypertension	177	32.9
Atrial Fibrillation	213	39.6
Confusion (24hrs)	235	43.7
Other Deficit (24hrs)	281	52.2
Brainstem/Cerebellar Signs (24hrs)	291	54.1
Angina	305	56.7
Previous TIA	310	57.6
Sensory Inattention (24hrs)	318	59.1
Congenital Gaze Paresis (24hrs)	361	67.1
Hemianopia (24hrs)	423	78.6
Peripheral Vascular Disease	488	90.7

from the imputation procedure and analysis, however they are included in our general exploration of the missing data as they have potential to be informative regarding the missing data mechanism.

All of the medical conditions considered as risk factors for stroke, and recorded as part of patient medical history, have missing values for over 30% of the patients. This highlights potential issues regarding the collection of this data. The variables recording symptoms present in the first 24 hours post-stroke also have high proportions of missing information, with six of these variables having over 40% of observations missing.

In order to explore the missing data in more depth, the VIM package in R was used to visualise the data. Initially aggregation plots were used to show the patterns and proportions of the missing data, where aggregation plots display the combinations of observed and missing data profiles and identify distinct patterns of missingness. For example, Figure 4.7 gives the aggregation plot for each of the incomplete variables with less than 50% missingness, showing the combinations of missing and observed data for these variables. The second to lowest row in Figure 4.7 shows the combination where 'confusion' is missing, but all other variables are observed. The combinations are ordered to ascend from the most common on the bottom row of the plot, to the least common on the top row, and red represents missing. In Figure 4.7, the vertical bars given above the aggregation plot present the proportions of missingness within each variable. It is clear this data has an arbitrary pattern of missingness, however, Figure 4.7 also shows that there are some clear groups of variables with similar missing data patterns.

Figure 4.7 clearly shows that systolic and diastolic BP at admission had the same missing data pattern, an expected result as these measurements are always taken together. Figure 4.7 also shows that the previous medications: anti-hypertensives, anti platelets and anti-coagulants, had similar missingness patterns, where if information regarding anti-hypertension medication is missing, informa-

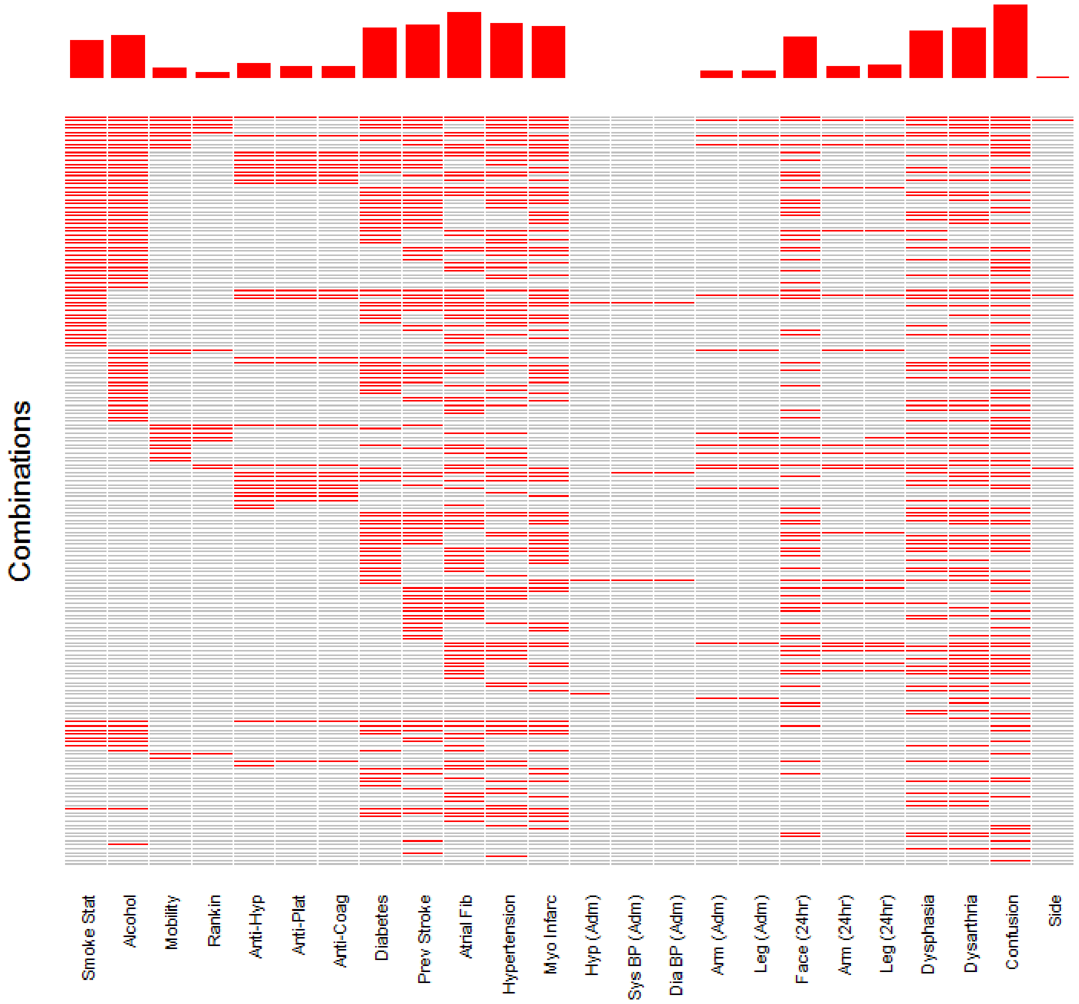


Figure 4.7: Aggregation plot showing all combinations of missing (red) and non-missing (grey) parts of the observations, for the incomplete baseline variables with less than 50% missingness. Vertical bars on top present proportions of missing observations for each variable.

tion regarding the anti-coagulants and anti-platelets were also missing.

Several other groupings can be seen in Figure 4.7, including variables such as arm and leg weakness. These groupings can be shown more clearly by fitting aggregation plots to subsets of variables. These can be seen in Figures 4.8, 4.9, 4.10, and 4.11, where a histogram is presented to the left showing the proportions of missingness for each variable, and the combinations of missingness are shown on the right, again with the most common combination of missingness being given on the lowest row, ascending to the least common combination on the top row.

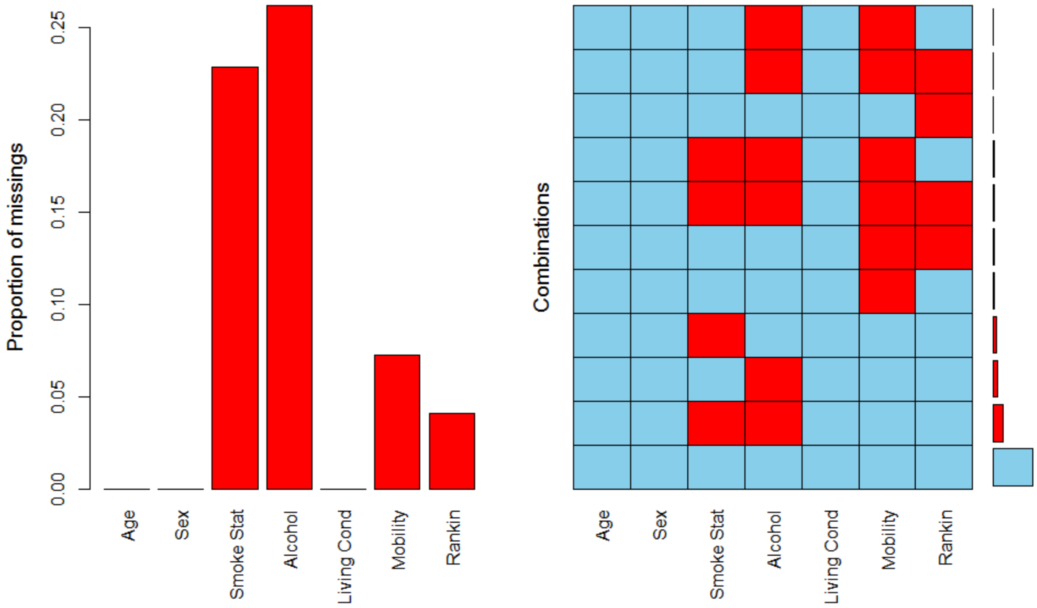


Figure 4.8: Visualisation of the patterns and proportions of missingness for patient characteristics variables. LEFT: Barplot for proportion of missing values in each variable. RIGHT: Aggregation plot showing all combinations of missing (red) and non-missing (blue) parts in the observations, with horizontal bars showing corresponding frequencies.

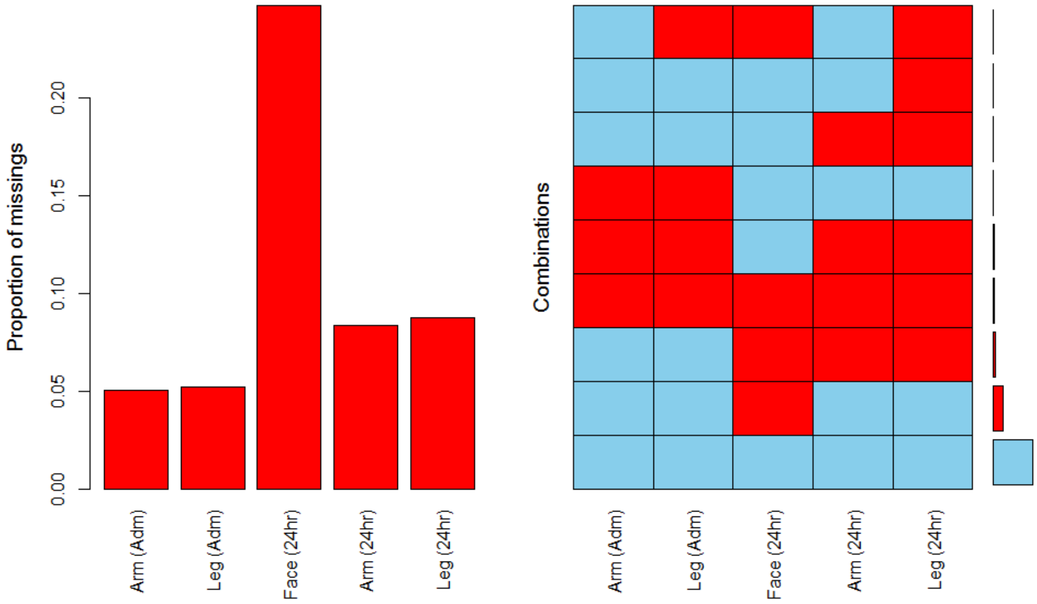


Figure 4.9: Visualisation of the patterns and proportions of missingness for variables relating to facial and limb weakness. LEFT: Barplot for proportion of missing values in each variable. RIGHT: Aggregation plot showing all combinations of missing (red) and non-missing (blue) parts in the observations, with horizontal bars showing corresponding frequencies.

Again, missing values are represented by red, and observed by blue.

Figure 4.8 shows that alcohol consumption and smoking status had similar missingness patterns, where smoking status was mostly missing if alcohol consumption was. Similarly, pre-stroke Rankin and mobility had similar missingness patterns, where pre-stroke Rankin was often missing if pre-stroke mobility was. Further, Figure 4.9 highlights that arm and leg weakness also had similar missing data patterns, where arm weakness was always missing if leg weakness was, at both admission and in the first 24 hours post-stroke. In both of these figures, the complete cases are the most common combination.

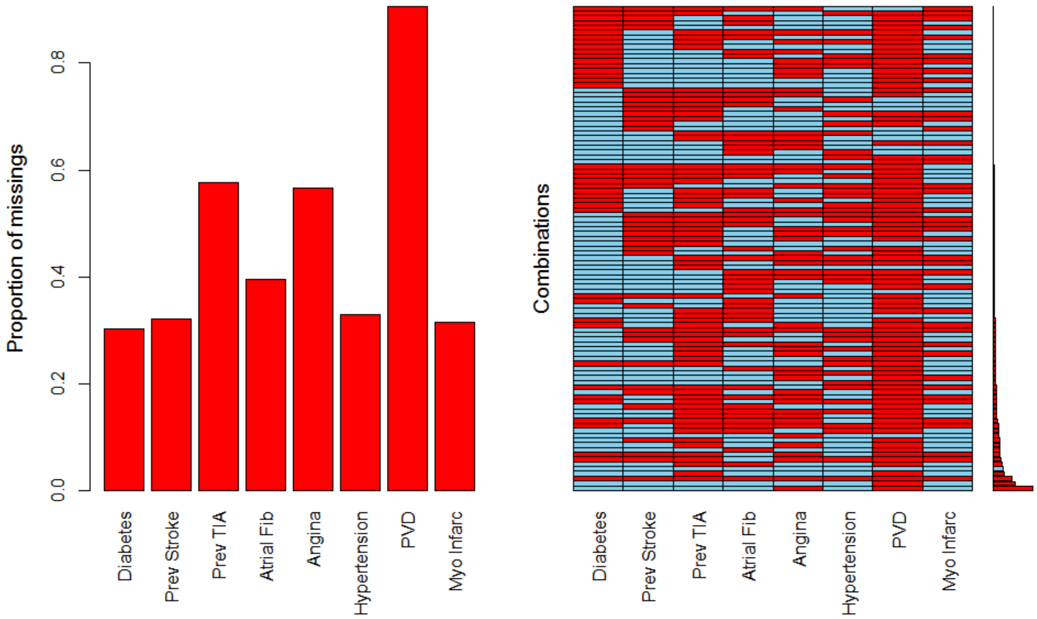


Figure 4.10: Visualisation of the patterns and proportions of missingness for variables relating to patient medical history. LEFT: Barplot for proportion of missing values in each variable. RIGHT: Aggregation plot showing all combinations of missing (red) and non-missing (blue) parts in the observations, with horizontal bars showing corresponding frequencies.

Figure 4.10 gives the aggregation plot for patient history of conditions considered as risk factors for stroke. The histogram shows that all of these variables had a high proportion of missing values, where three had over 50%. This high amount of missingness results in the combination where all of these variables are missing being more common than the combination of them all being observed. The amount of combinations in this aggregation plot is high due to the amount of missing values, and suggests an arbitrary pattern of missingness. The two variables with the most similar pattern were previous stroke and previous TIA, where they are often missing together. Clinically, this may be expected as stroke and TIA are the most similar of these medical conditions, with TIA often a pre-cursor to stroke. With similar departments treating these conditions, it makes sense that they would either know medical history regarding both or neither, but the excess missingness for TIA may relate to TIA being harder to establish.

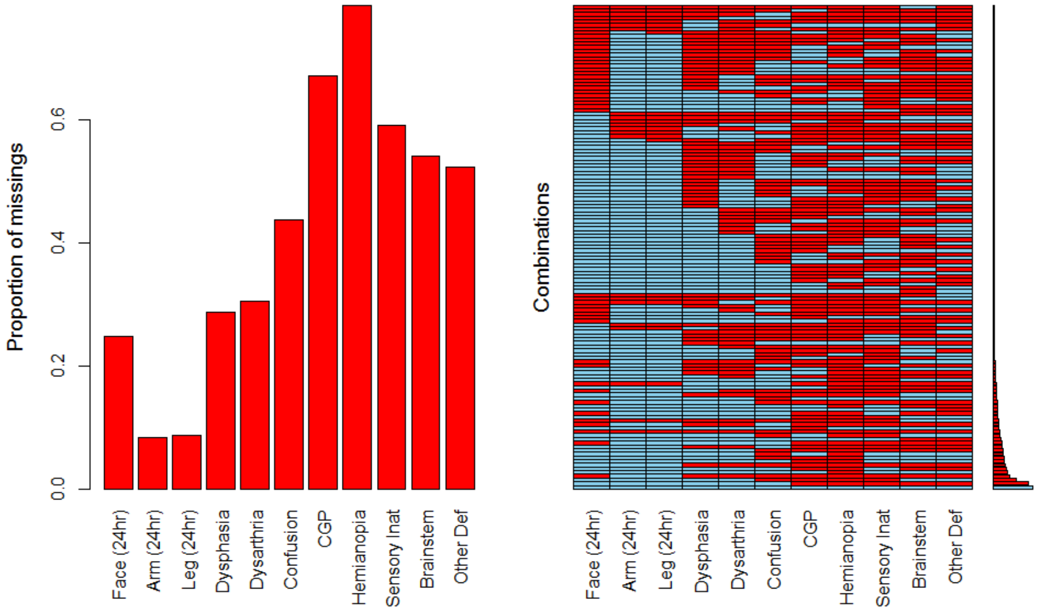


Figure 4.11: Visualisation of the patterns and proportions of missingness for variable relating to symptoms present in the first 24 hours post-stroke. LEFT: Barplot for proportion of missing values in each variable. RIGHT: Aggregation plot showing all combinations of missing (red) and non-missing (blue) parts in the observations, with horizontal bars showing corresponding frequencies.

The final subset of variables considered are the variables relating to symptoms present in the first 24 hours post-stroke. The aggregation plot for these variables is given in Figure 4.11, which shows that this group of variables generally had high proportions of missing values, with five having over 50%. The most common combination is however the case where they were all observed. Figure 4.11 shows that the missingness in arm and leg weakness could be related, where their amount and patterns of missingness were closely matched. A similar result is shown in Figure 4.11 for dysphasia and dysarthria, where they too appear to have a related amount and pattern of missing values.

Overall the aggregation plots have shown that the general missingness pattern is arbitrary, with some appearance of monotone and file matching patterns within pairs or small subsets of variables. An alternative method for visualising the missing data patterns is to use matrix plots, which also give an indication of the missing data mechanism. Matrix plots present each patients' observations as a row in the plot, where the black, white and grey scale indicates the observed values, and red represents missing observations. Matrix plots can be sorted by the values of a particular variable to highlight any patterns in the data related to that variable, although patterns can be clear regardless of sorting.

Figure 4.12 gives the matrix plot sorted by lesion type shown in CT scan. At first glance, the large amount of red in this matrix plot highlights the amount of missing data within this data set, particularly for variables regarding medical history and stroke symptoms. Looking closer at the CT scan variable itself, 'no scan' is represented in black and it appears in Figure 4.12 that patients who did not have a CT scan had more missing information within other variables; BP for example is only missing for 'no scan' patients. Another pattern shown in Figure 4.12 is that 'no scan' appeared to be more common for patients admitted to Hospital 2 (black) compared to the other hospital. There also appears to be blocks of missing data (red) within this matrix plot relating to hospital. Further, blocks of missingness can be seen to be related to patients who were unconscious within the first 24 hours post-stroke, represented as black in the matrix plot.

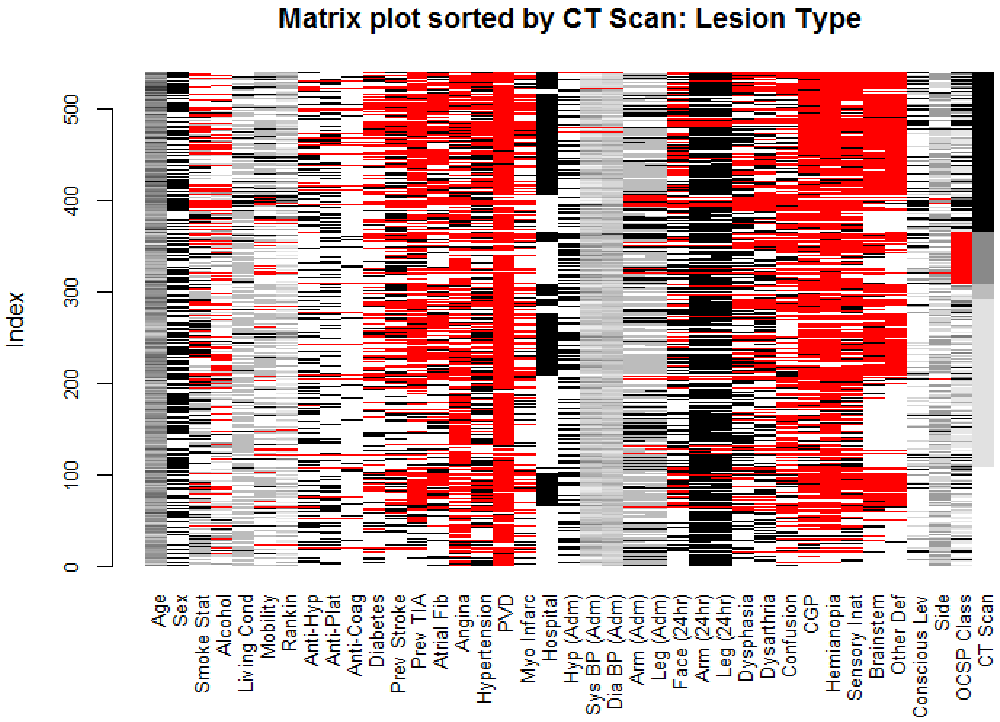


Figure 4.12: Matrix plot showing visualisation of the missing data patterns and mechanism, sorted by lesion type shown in CT Scan.

In order to explore missingness related to hospital further, Figure 4.13 gives the matrix plot sorted by hospital. Patients admitted to Hospital 2 are shown to have had more missing data overall, where there is clearly more red towards the top of the matrix plot overall. The missing data within the variables for whether patients had brainstem and cerebellar signs or other deficit within the first 24 hours post stroke appears related to hospital, where most of the missing data was for Hospital 2. Smoking status and alcohol consumption was also more commonly missing for patients admitted to Hopsital 2, as was whether or not they had facial weakness as a symptom of stroke within the first 24 hours. On the other hand, whether or not patients had arm or leg weakness as a symptom of stroke is shown in Figure 4.13 to have been more commonly missing for patients admitted to Hospital 1. Pre-stoke mobility and Rankin was also more commonly missing for patients admitted to Hospital 1.

Looking at the general blocking of missing data in Figure 4.13, it can be seen that unconscious patients commonly have missing data. Figure 4.14 sorts the matrix plot by worst consciousness level in the first 24 hours post-stroke to examine this more, where black represents ‘coma’, scaling to white for ‘alert’ patients. It can be seen in Figure 4.14 that patients missing information regarding whether or not they had arm weakness or leg weakness was missing most often for comatose patients, as was smoking status and alcohol consumption. Whether or not patients had dysarthria or dysphasia was also mostly missing for comatose patients.

More generally, the matrix plots in Figures 4.12, 4.13 and 4.14 show that there are clear patterns in the missingness between variables, but the pattern overall is clearly arbitrary with some randomness. Looking closer at some other variables, it can be seen that alcohol consumption and smoking status were often missing together, and this had some correspondence to sex, where females (black) were missing this information more often. Information regarding the previous medications was also often missing together, and the variables which had a large

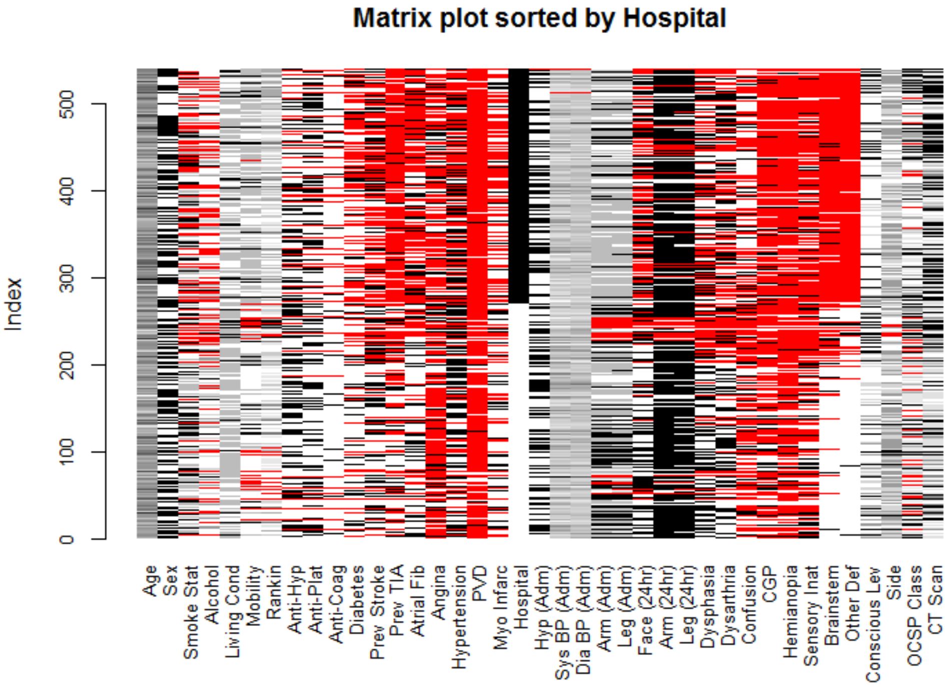


Figure 4.13: Matrix plot showing visualisation of the missing data patterns and mechanism, sorted by hospital admitted to.

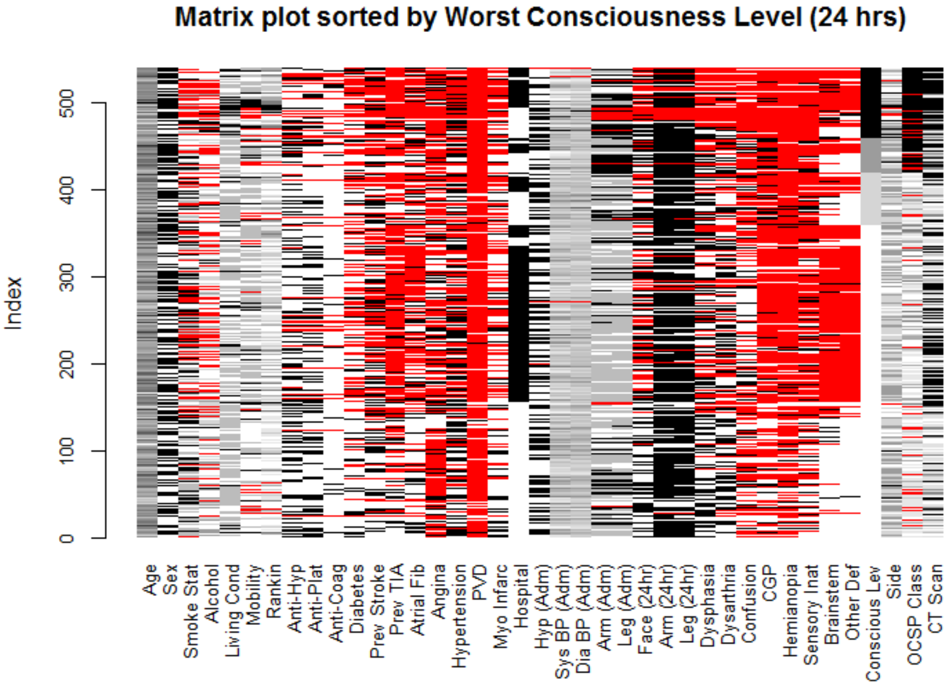


Figure 4.14: Matrix plot showing visualisation of the missing data patterns and mechanism, sorted by worst consciousness level in first 24 hours post-stroke.

proportion of missingness were mostly missing together for the same patient. The matrix plots also highlight the issues around the OCSP classification variable, where Figure 4.12 shows that patients who were diagnosed with PICH (dark grey) in the CT scan were missing OCSP classification, and Figure 4.14 shows that patients whose consciousness level was categorised as ‘coma’ (black) or ‘stupor’ (dark grey) were all categorised as ‘unconscious’ (black) in the OCSP classification variable.

An additional method of exploring missing data is to visualise the relationship between two variables using spine plots. These provide a way of visualising the amount of missing data in one variable dependent upon the value of the other variable, thus providing a method of assessing if the missing data in one variable is related to the observed values in another. Several spine plots are shown in Figures 4.15 and 4.16, where each column represents a level or set of values for the variable given on the x-axis, and the proportion of missingness for the other variable is given on the y-axis, where red in each column represents missing data and blue represents observed. The width of each column indicates the amount of patients within each level of the x-axis variable, with a column for ‘missing’ if this variable is incomplete.

Figure 4.15 gives spine plots for missingness dependent upon pre-stroke Rankin, age, sex and hospital. The spine plots in Figures 4.15(a) and 4.15(b) show how missing data may depend on pre-stroke Rankin for facial weakness and confusion, respectively, as symptoms of stroke. These show the amount of missingness to be increasing for worsening Rankin scores. This relationship was seen for many of the variables when included in a spine plot against pre-stroke Rankin. Figures 4.15(c) and 4.15(d) give the spine plots for smoking status and diabetes, respectively, against age, where a general pattern emerges that missingness was increasing with age for patients over 60 years. The youngest groups of patients had slightly higher proportions of missing data compared to the 60 to 65 age range however.

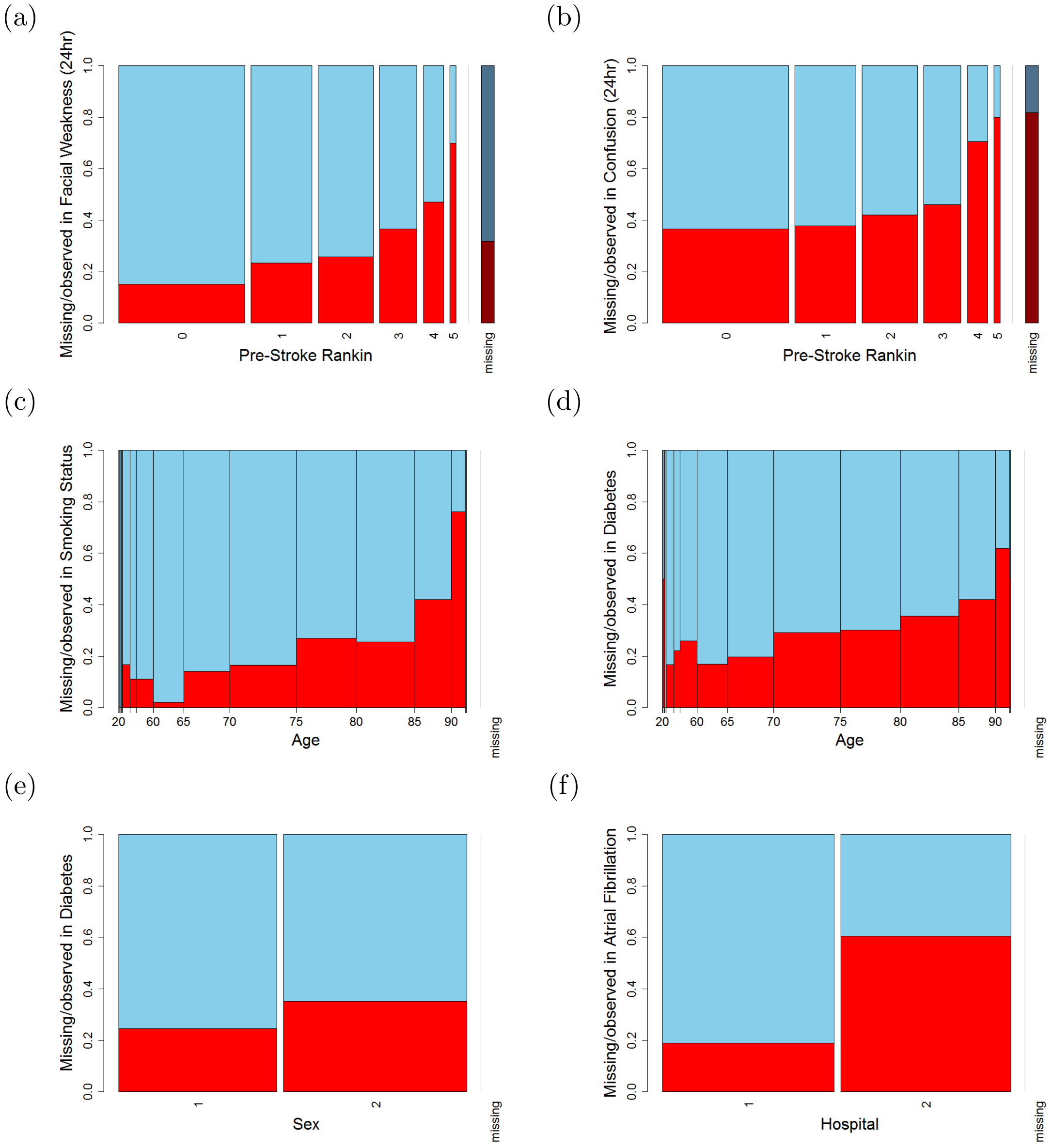


Figure 4.15: Spine plots showing colour coding of missing and available data for variables on y-axis against value of variable given on x-axis.

Figure 4.15(e) shows the amount of missing data for diabetes against sex, and shows that females had more missing data regarding diabetes. This was a common pattern for sex, where a higher proportion of the missing observations in each variable were missing for females compared to males. The spine plot in Figure 4.15(f) looks at atrial fibrillation against hospital, and shows that patients admitted to Hospital 2 had a higher proportion of missing data than those admitted to Hospital 1; again this pattern was common across the other variables.

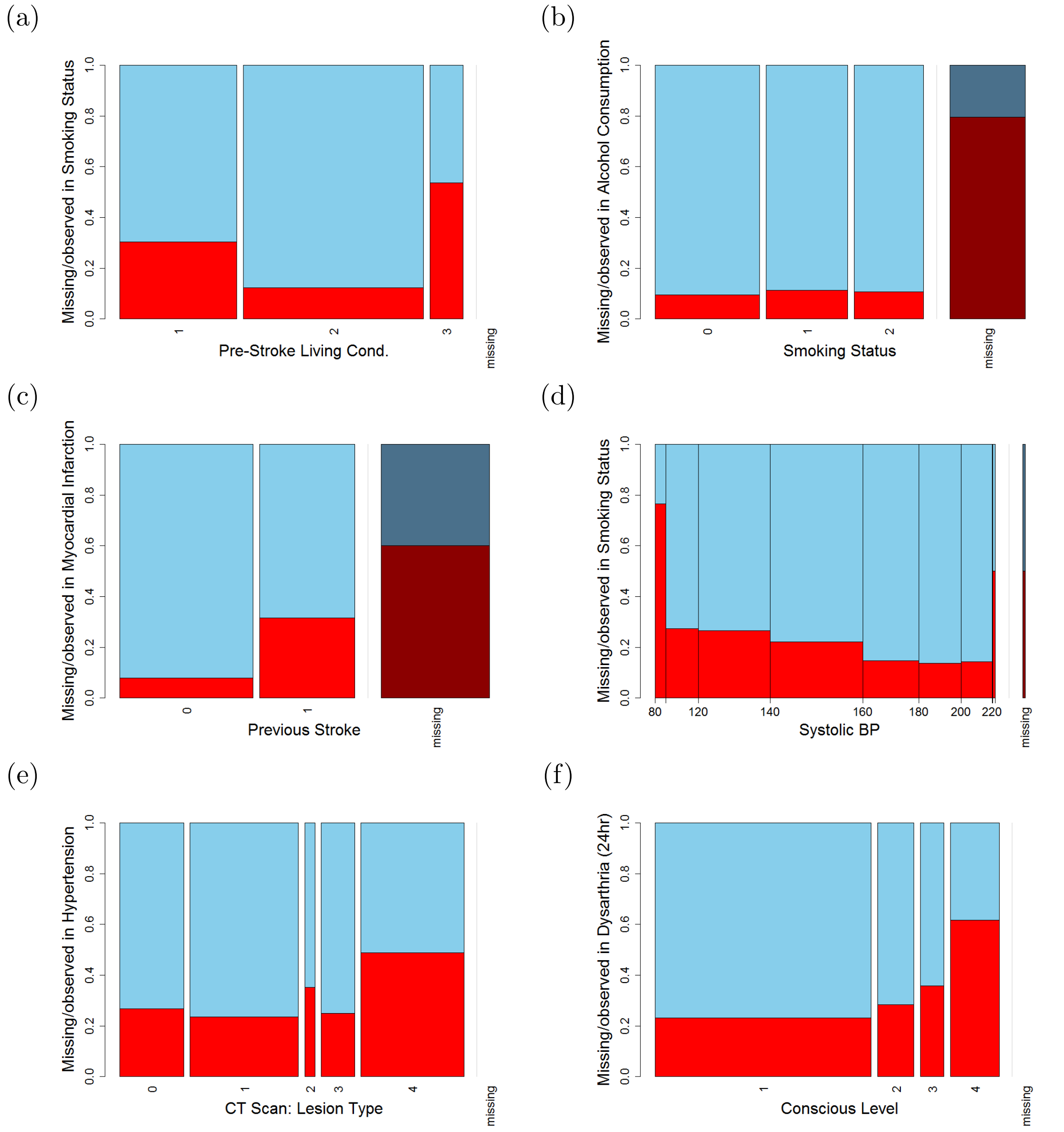


Figure 4.16: Spine plots showing colour coding of missing and available data for variables on y-axis against value of variable given on x-axis.

Further spine plots are given in Figure 4.16. The top left spine plot, Figure 4.16(a), shows the amount of missing observations for smoking status by prestroke living conditions, and shows that patients living either home alone or in an institution had more missing observations compared to those living at home with a companion. Looking at Figure 4.16(b), it can be seen that smoking status and alcohol consumption were often both missing together for a patient, and whether patients smoked did not seem to affect the amount of missing data for alcohol consumption.

The spine plot for history of myocardial infarction against previous stroke is given in Figure 4.16(c). This shows that patients who have had a stroke previously had more missing observations regarding myocardial infarction. The highest proportion of those missing data on myocardial infarction were those who were also missing data regarding previous stroke occurrence. This affect appeared to be common across the other variable, particularly for those relating to medical history.

Considering the effect systolic BP may have had on whether data is missing, Figure 4.16(d) gives the spine plot for smoking status against systolic BP. This shows that smoking status was more commonly missing for more extreme BP values, both high and low. This was a frequent occurrence for many variables, with missing data in other variables being more common for patients with systolic BP values below 100mmHg compared to other BP ranges.

Figures 4.16(e) and 4.16(f) present spine plots to look at the effect of lesion type shown in CT scan and worst consciousness level in the first 24 hours post-stroke, respectively. These spine plots further reiterate the findings of the matrix plots that ‘no scan’ and ‘coma’ seemed to increase the likelihood of patients having missing observations. Figure 4.16(e) gives the proportions of missing data for hypertension against lesion type in CT scan, showing ‘no scan’ and ‘PICH’ patients had the highest proportions of missing data regarding hypertension. In Figure 4.16(f) it can be seen that worsening consciousness levels related to increased missing data for dysarthria. This pattern was seen among many of the variables, where the proportions of missing data were found to increase with worsening consciousness levels.

Through exploration of the missing data, the missing data will be assumed to be missing at random (MAR) for the remaining analyses. The matrix and spine plots clearly show that the missingness was related across variables, and dependent upon the observed values of other variables. This suggests MAR is more plausible than missing completely at random (MCAR), and MCAR should not be assumed.

It is not possible to truly to detect between missing not at random (NMAR) and MAR, therefore care needs to be taken when deciding which of these assumptions to base analyses upon. Discussions with the clinicians responsible for data collection provided some assurance that the reasons for data being unavailable were not due to the underlying values of the missing observations. The arbitrary pattern of missingness, and associations found between missing data and the observed values of other variables, may further evidence this, though it cannot be truly ruled out that the data may be NMAR. Considering the findings of the missing data exploration, alongside assurances from clinicians and the added complications that the NMAR assumption creates within analyses, we will assume the missing data is MAR and use methods suitable for the MAR assumption when handling the missing data.

4.4 Multiple Imputation

4.4.1 Imputation Procedure

The missing data was handled using multiple imputation as outlined in Sections 3.3.2 and 3.3.3, where the imputation stage was implemented using multiple imputation using chained equations (MICE). The `mice` package (van Buuren and

Groothuis-Oudshoorn, 2011) in R (R Core Team, 2019) was used to perform the imputation procedure. There were many considerations to take into account during this stage to ensure the imputation models were appropriately specified.

In the previous section, the missing data was explored, concluding that the missing data could be assumed to be MAR and handled as such. This exploration also showed that eight variables had missing data for over 50% of patients, and therefore these variables are excluded from the imputation and analysis process. The 50% cut point was chosen through the recommendations by White et al. (2011), who outlined that carrying out imputation on variables with over 50% missingness can amplify any imperfections in the imputation procedure.

In order to specify an imputation model for each of the remaining incomplete variables, several considerations needed to be made. Firstly, variable type was considered to ensure appropriate form of the imputation models. Linear regression models were specified to impute the incomplete continuous variables, which included systolic and diastolic BP. Binary variables were imputed using logistic regression, where the incomplete binary variables included previous anti-hypertensives, anti-platelets and anti-coagulants, history of diabetes mellitis, hypertension, myocardial infarction and atrial fibrillation, previous stroke, hypertension at admission, facial, arm and leg weakness within the first 24 hours post stroke, and stroke symptoms: dysphasia, dysarthria and confusion. To impute ordered and unordered categorical variables, multinomial logistic regression was used, where the variables with imputation models of this form were pre-stroke mobility, pre-stroke Rankin, smoking status, alcohol consumption, and arm and leg weakness at admission to hospital.

The next consideration to be made was the predictors to be included in the imputation models. Each imputation model should include all variables in the analysis model, and in particular the outcome of the analysis model. As recommended by White and Royston (2009), the survival outcome was included in the imputation models using the censoring indicator, δ, and the Nelson-Aalen estimate of the cumulative hazard to the survival time. Additionally it is important for each imputation model to include all predictors of the incomplete variable, along with any variables which may predict whether or not the incomplete variable is missing. Given all these considerations, an all for all approach was taken when specifying the imputation models, where every variable was included as a predictor for each incomplete variable.

After specification of the imputation models, the number of imputations and iterations need to be considered. Given each of the variables included in the imputation procedure have under 50% missingness, Bodner's rule of thumb can be used here, where it is recommended to look at the overall percentage of missing data and impute a similar number of imputations. White et al. (2011) suggested the required number of imputations for $F = 0.05, 0.1, 0.2, 0.3, 0.5$ would be $m \geq 3, 6, 12, 24, 59$ respectively. The percentage of missing data for the stroke data is 22%, therefore this rule of thumb would suggest the number of imputation should be between 12 and 24. However, due to the amount of variables in the imputation models, and the limited computational resources available, m was chosen to be 10. This is sufficient to still give reasonable power to the imputations, and only a small loss of efficiency.

The imputations were carried out using the `mice` package in R, where the variables chosen to be included as predictors in the imputation model for each incomplete variable were specified using a prediction matrix, and the appropriate model forms for the variable types were specified within the imputation function. The imputation cycle was run for 1000 iterations to ensure convergence, and repeated 10 times to produce 10 imputed data sets.

4.4.2 Imputation Diagnostics

After the imputation procedure had been carried out, it was important to carry out checks on several aspects of the imputations. In particular, it was important to check the between and within imputation variability and the convergence of the imputations. Strip plots and histograms were used to examine the distributions of the imputations, enabling assessment of the between and within imputation variability. Convergence was assessed using trace plots.

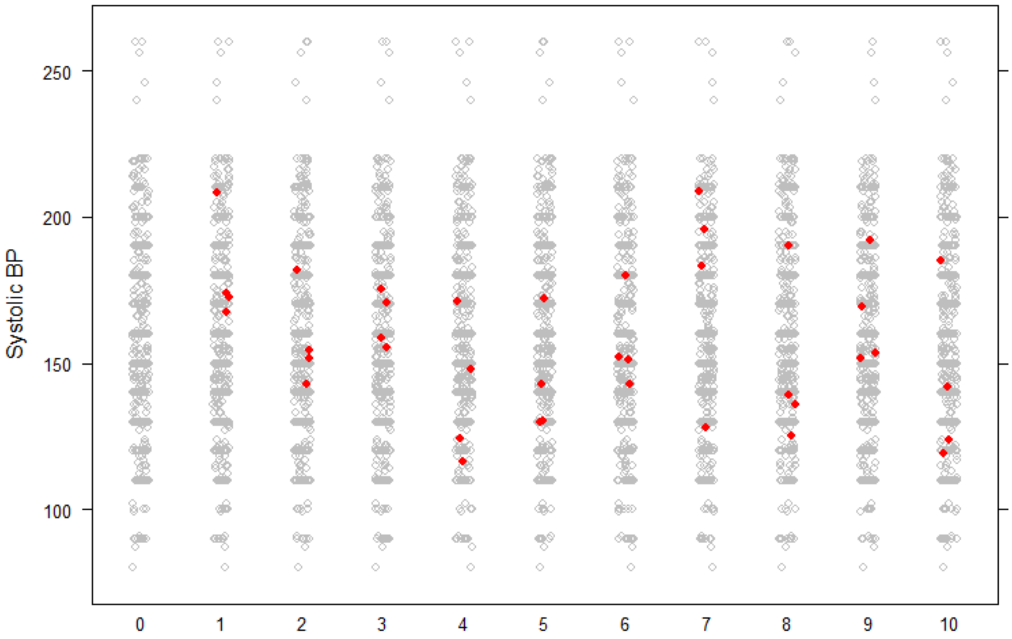


Figure 4.17: Strip plot showing the values of Systolic BP for the incomplete data, labelled 0, and the 10 imputed data sets, where imputed values are displayed as red points.

Firstly looking at the distribution of imputations for continuous covariates, strip plots indicate how the imputed values relate to the observed values of the variables, and show how the imputed values differ across the imputations $m = 1, ...10$, assessing both the within-imputation and between-imputation variability. Figures 4.17 and 4.18 gives the strip plots for the imputations of systolic and diastolic BP, respectively, where the imputed values are given in red and the observed values in

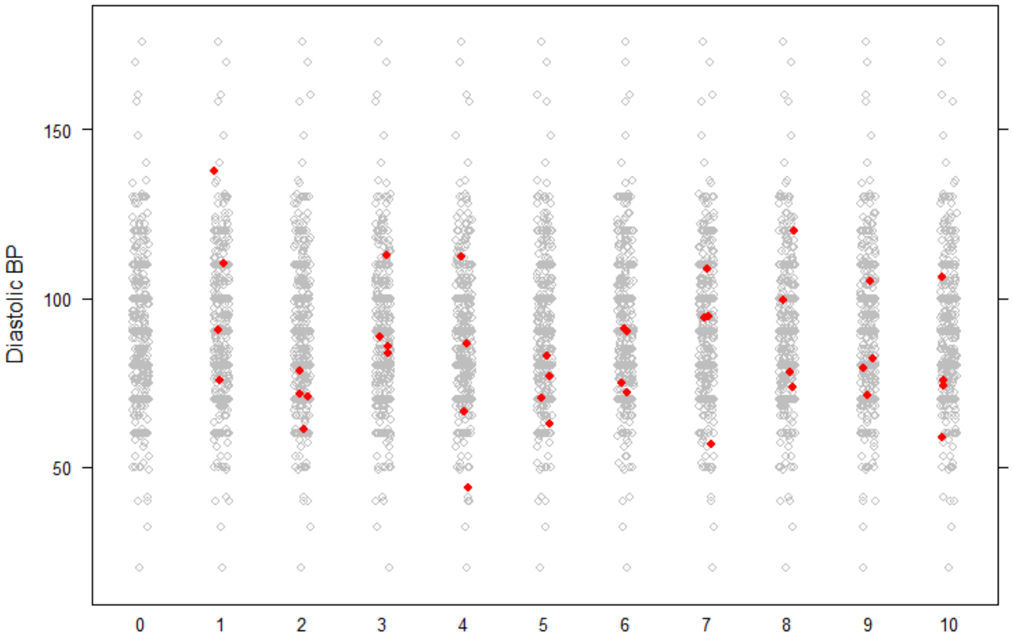


Figure 4.18: Strip plot showing the values of Diastolic BP for the incomplete data, labelled 0, and the 10 imputed data sets, where imputed values are displayed as red points.

grey. The observed data prior to imputation is represented as 0 on the x-axis, and the remaining values 1 to 10 represent each imputation $m = 1, ..., 10$. For both systolic and diastolic BP, it can be seen in Figures 4.17 and 4.18 that, for each imputation m, the imputed values fall within the range of the observed values, with no extreme values imputed. This suggests the imputations are sensible given the observed data. On the other hand, the strip plots indicates that there appears to some variability in the imputed values between the imputations. This is not a large concern though as only four values have been imputed for each of the BP variables.

For categorical and binary covariates, the distribution of the imputations was explored using histograms. Plotting histograms of the imputed values for each imputation m of each variable shows the assignment of patients to each level, checking between-imputation variability. Figure 4.19 gives the histograms showing the distribution of the imputed values for alcohol consumption for each of the imputed data sets. It can be seen in Figure 4.19 that each of the imputed data sets had a similar proportion of patients imputed into each of the levels of alcohol consumption, showing minimal between-imputation variability. Considering diabetes mellitus, the histograms in Figure 4.20 again show that similar proportions of patients have been imputed into each level across the imputed data sets. Figure 4.21 also indicates a similar result for facial weakness, however this also shows that a similar number of patients were imputed into each level within each imputed data set.

The histograms for pre-stroke Rankin are given in Figure 4.22. These indicate a larger amount of between-imputation variability, where there are clear differences in the distribution of the imputed values across each of the imputed data sets. An example of this variability can be seen by looking at level 0 in data sets 3 and 5 in Figure 4.22, where we can see that data set 5 has twice as many patients imputed as level 0 than data set 3. Further, neither data set 3 or 5 have any patients imputed as level 5, whereas data set 1 has 4 patients in this level. This variability does not cause huge concern though, some between-imputation variability is expected due to the uncertainty around the true underlying value of the missing data, and the key reason for using multiple imputation is to reflect this.

Further histograms were plotted to examine the proportions of patients imputed into each level of the categorical variables, comparing the distribution of imputed values to the imputed variable as a whole and the observed data prior to imputation. Three histograms were plotted for each variable: one with imputed values only, one combing the 10 imputed data sets and one for the original observed data. In Figure 4.23, we include these histograms for pre-stroke Rankin as pre-stroke Rankin was shown to have some between-imputation variability previously in Figure 4.22. Looking at the histogram of the imputed values in Figure 4.23, which gives the total number of patients imputed into each level over all 10 imputations, we can see that the distribution of the imputed values is not overall that

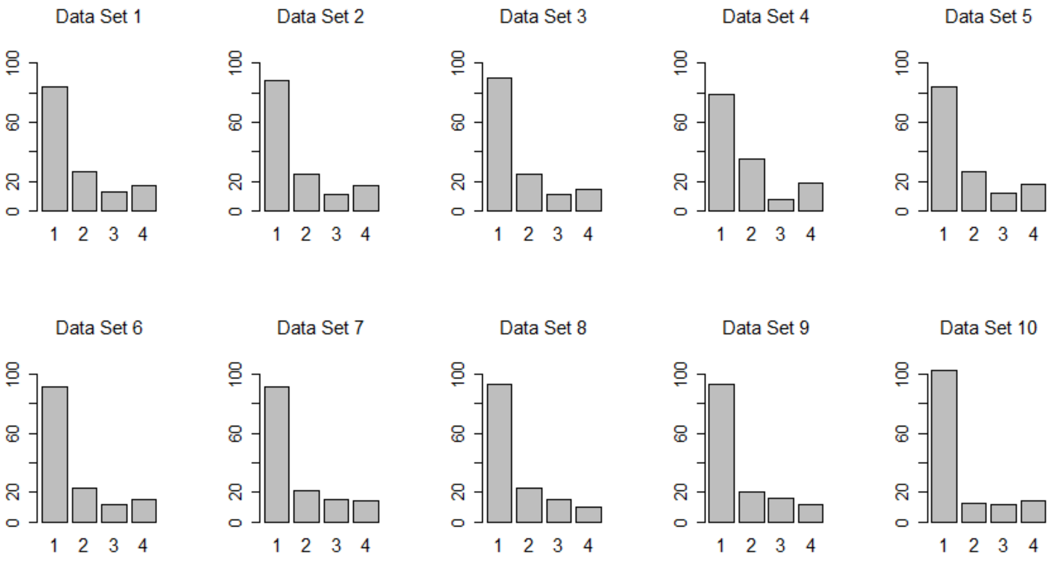


Figure 4.19: Histogram showing the number of patients imputed into each level of alcohol consumption for each imputed data set

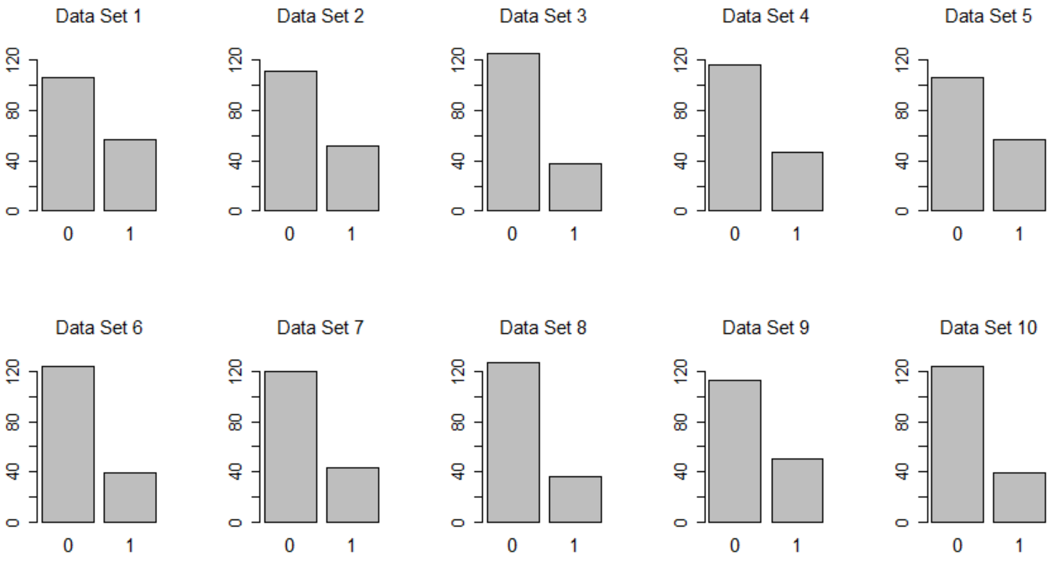


Figure 4.20: Histogram showing the number of patients imputed into each level of diabetes mellitus for each imputed data set

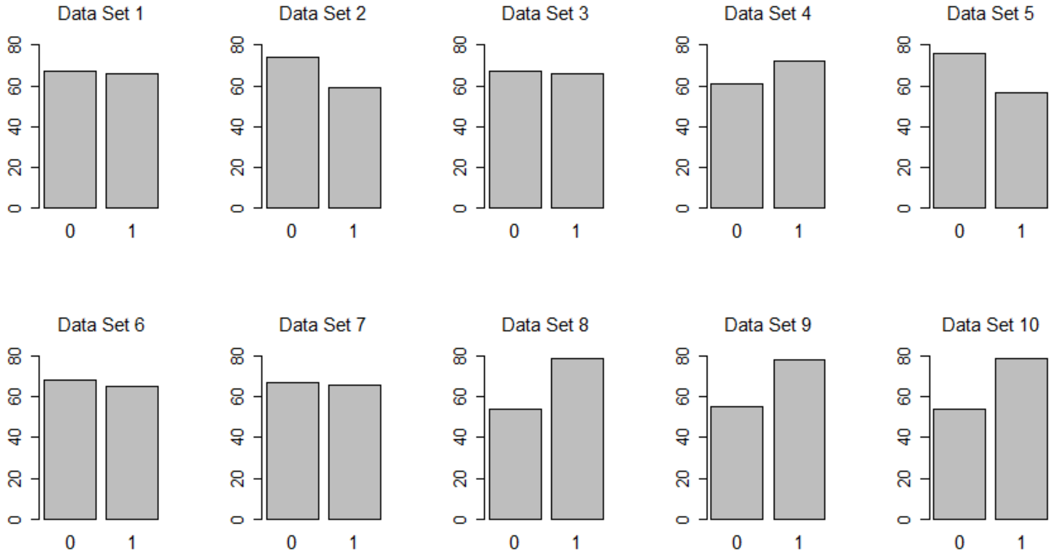


Figure 4.21: Histogram showing the number of patients imputed into each level of facial weakness for each imputed data set

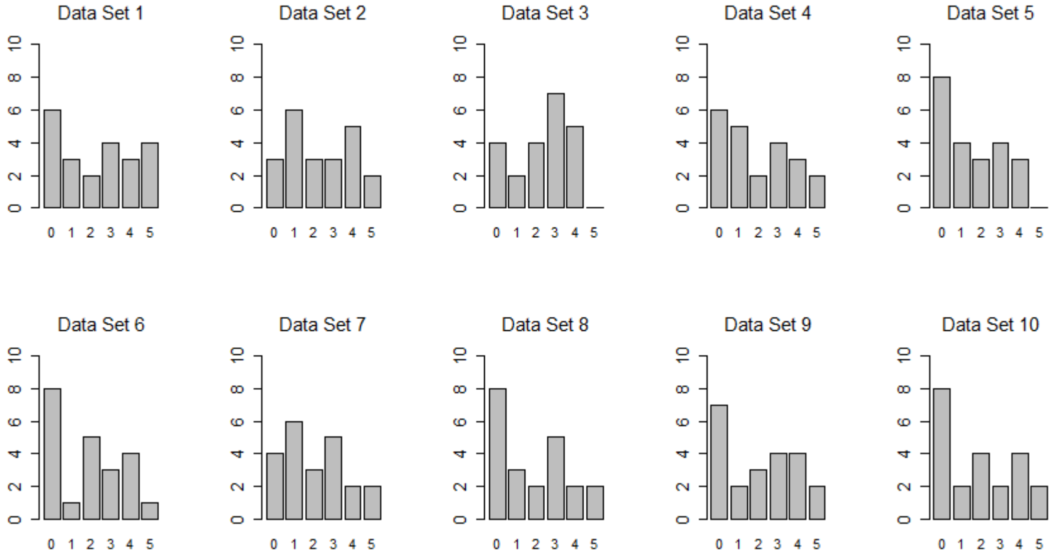


Figure 4.22: Histogram showing the number of patients imputed into each level of pre-stroke Rankin for each imputed data set

different to the distribution of the observed and imputed data as a whole, aside from levels 3 and 4 having higher proportions. Comparing the histograms of the observed data to the overall imputed data, Figure 4.23 shows that the imputation procedure has made minimal difference to the overall distribution of this variable. The finding of similar results on examination of the other variables indicates that the imputed data has been generated to be comparable with the observed data in terms of its distribution and variability.

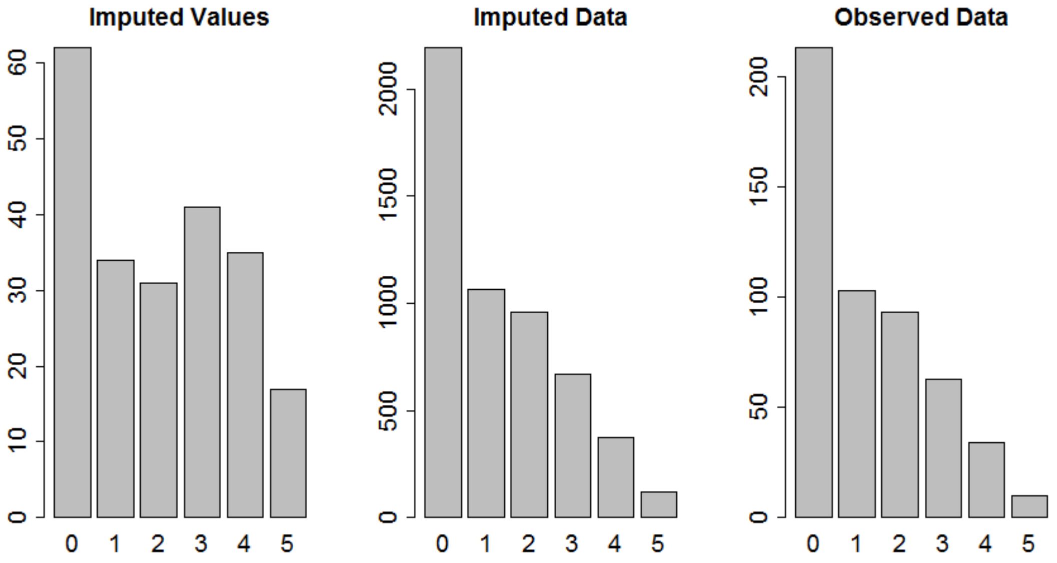


Figure 4.23: Histograms showing: the overall proportion of patients imputed into each level of pre-stroke Rankin over the 10 imputations, the total proportions of patients in each level over the 10 imputed data sets and the proportion of patients in each level of the observed data.

Finally, trace plots were used to assess the convergence of the imputations. A trace plot was drawn for each imputed variable, where each trace plot presents a trace of the mean and standard deviation for each of the m imputations, $m = 1, ..., 10$, over the iterations. Trace plots were plotted for each of the covariates imputed, where the trace of the mean is given in the left panel and the standard deviation given on the right. The majority of the trace plots showed good convergence, however there were a few giving possible cause for concern.

In particular, Figure 4.24 gives the trace plots for side of lesion, hypertension at admission and systolic BP at admission, where the mean and standard deviation for both side of lesion and hypertension at admission do not appear to converge very well. However, this is likely to be due to variable type and amount of missingness; side of lesion is categorical with only 1.5% missing data, and hypertension is binary with only 0.7% missing data. The small amount of values being imputed with these variable types results in more fluctuating trace plots.

The trace plots for arm and leg weakness in the first 24 hours post-stroke in Figure 4.25 highlight the need to ensure sufficient iterations, where there is clear change in the means within the early iterations. We can conclude that overall the trace plots show good convergence for the imputations over the 1000 iterations.

The imputation diagnostics as a whole seem to imply that in general there was not excessive between-imputation variability, the convergence was good, and the imputed values were in line with observed data. This supports the plausibility of the MAR assumption, and suggests the imputation procedure was satisfactory.

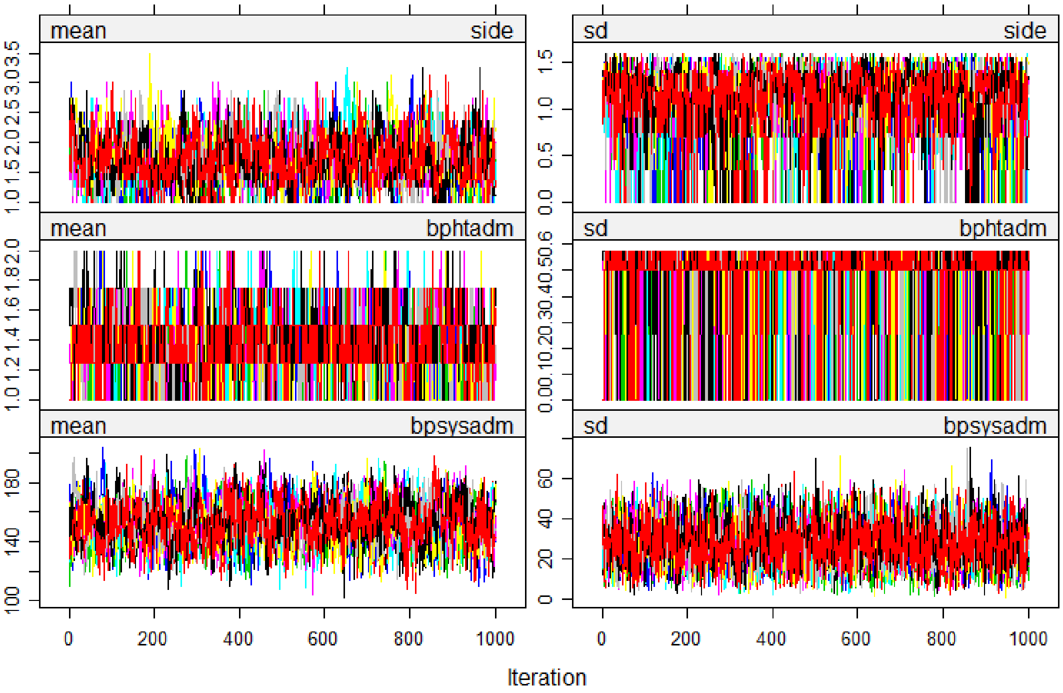


Figure 4.24: Trace plot for imputations of side of lesion, hypertension at admission and systolic BP at admission to assess their convergence.

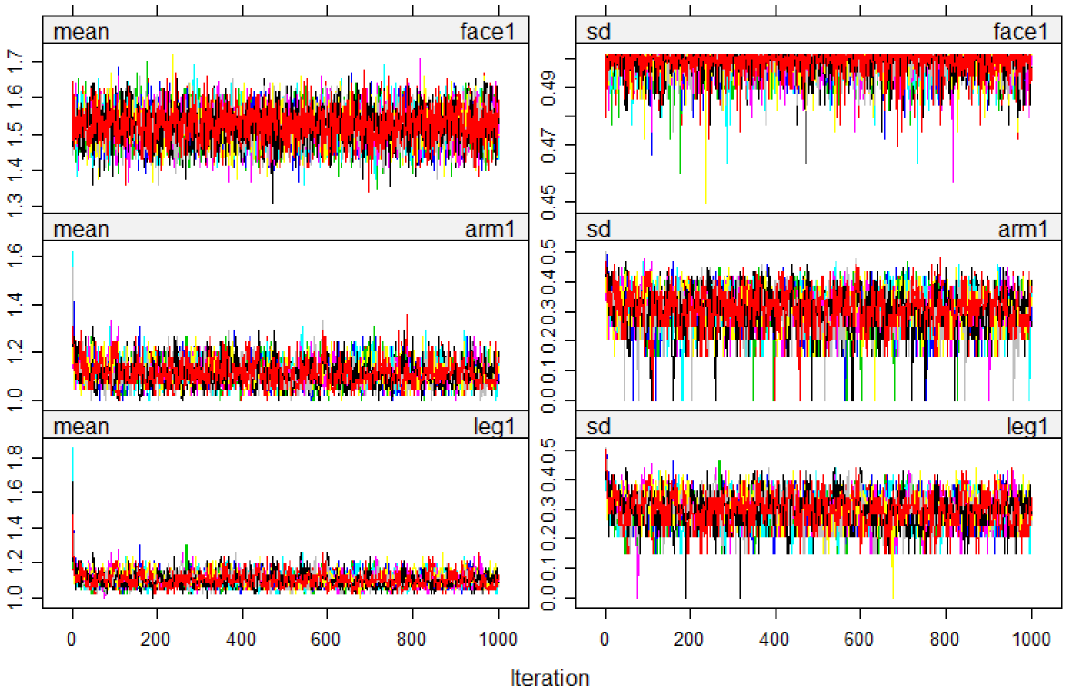


Figure 4.25: Trace plot for imputations of facial, arm and leg weakness in the first 24 hours post-stroke to assess their convergence.

4.5 Model Building

The analysis stage involved a model building procedure which was carried out on the imputed data sets using backwards selection and the Wald test, as outlined in Sections 3.2.5 and 3.3.4. This involved firstly fitting a fully adjusted Cox PH model to each of the 10 imputed data sets and using Rubin's rules to calculate the combined coefficients and standard errors. Using the Wald test and backwards selection, covariates were removed one by one in order to achieve the optimal set of covariates within the model.

Each stage of the backwards elimination involved fitting the Cox regression model to the new set of variables, after exclusion of one, to each of the imputed data sets separately and again combining the estimates using Rubin's rules. The variable of least importance for survival is then chosen for exclusion using the Wald test, and the model refitted to each of the imputed data sets excluding this variable, beginning the procedure again. This process was repeated until only variables significant for survival at the 5% significance level remained.

Using the combined estimates of the coefficients and standard errors, the hazard ratios, and their corresponding 95% confidence intervals and p-values were calculated. The results of this are given in Table 4.17, with a visualisation of the hazard ratios and confidence intervals provided in Figure 4.26.

The results of the model fitting show that the covariates important for survival are age, hospital, pre-stroke mobility, diabetes mellitus, side of lesion, lesion type shown in CT scan, and worst consciousness in the first 24 hours post-stroke, where each of these covariates were significant for survival at the 5% level in the adjusted Cox model, following backwards elimination.

Firstly looking at age, Table 4.17 shows that per each additional 10 years in age at time of stroke, there was a 29% increase in the hazard of death following stroke. The confidence interval gives 95% confidence that this increase in hazard is between 15% and 44%. This result coincides with the survival curves in Figure 4.2(a) split by age group, where the older age groups had worse survival rates overall.

Considering hospital admitted to, Figure 4.26 shows that stroke patients admitted to Hospital 2 had a reduced hazard of death compared to those admitted to Hospital 1 when adjusted for other important risk factors. Patients admitted to Hospital 2 had a reduction in hazard of almost 30%, as seen in Table 4.17. Referring back to the earlier data exploration, the survival curves in Figure 4.4(a), and the univariate Cox model for hospital in Table 4.13, showed that in the unadjusted setting, the opposite was the case, where Hospital 2 patients had an increased hazard, and where this effect was not significant. This change in effect will be considered further in the model validation.

The hazards for pre-stroke mobility, given in Table 4.17, suggest that patients with reduced mobility prior to stroke had an increased hazard of death post-stroke.

Table 4.17: Results of the Cox proportional modelling procedure on the multiply imputed data, showing the pooled estimates of the hazard ratios (HR), along with the corresponding 95% confidence intervals (CI) and *p*-values.

Variable	**HR**	**95% CI**	***p*-value**
Age (10 years)	1.288	(1.149,1.444)	<0.001
Hospital			
Hospital 1 (Baseline)			
Hospital 2	0.725	(0.558,0.940)	0.015
Pre-stroke Mobility			
200m Outdoors (Baseline)			
Indoors	1.458	(1.120,1.898)	0.005
Needs Help	1.688	(1.111,2.564)	0.014
Diabetes Mellitus			
No (Baseline)			
Yes	1.515	(1.069,2.147)	0.019
Side of Lesion			
No Lesion (Baseline)			
Right	0.884	(0.664,1.177)	0.398
Left	1.165	(0.871,1.558)	0.303
Both	1.884	(1.022,3.476)	0.043
Systolic BP (10mmHg)	0.954	(0.918,0.991)	0.015
CT Scan: Lesion Type			
No Lesion (Baseline)			
CI	1.288	(0.934,1.776)	0.123
HCI	2.192	(1.176,4.086)	0.014
PICH	1.844	(1.206,2.819)	0.005
No Scan	2.512	(1.771,3.563)	<0.001
Worst Conscious Level			
Alert (Baseline)			
Drowsy	1.900	(1.370,2.636)	<0.001
Stupor	2.308	(1.556,3.425)	<0.001
Coma	4.328	(3.144,5.957)	<0.001

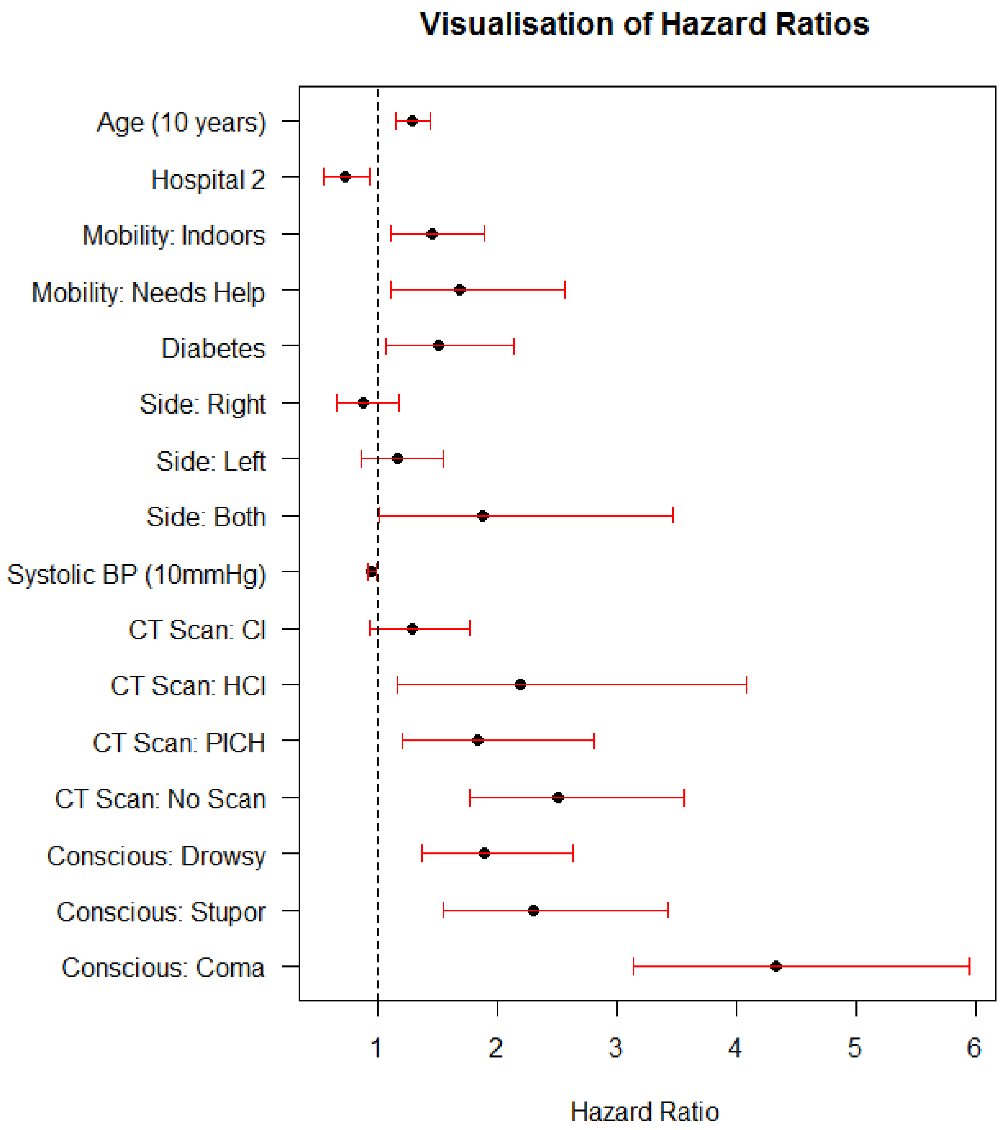


Figure 4.26: Visualisation of hazard ratios and 95% confidence intervals for the pooled Cox regression model results, showing the baseline reference hazard ratio at one as a dotted line.

Compared to the baseline of being capable of walking 200 metres outdoors, patients needing help to get around were most at risk of death with a 69% increase in hazard, whereas those able to move around indoors had an increase in hazard of 46%. This is concurrent with the data exploration findings, however, adjustment for other factors resulted in a reduction in the level of increase of the hazard ratios.

Table 4.17 and Figure 4.26 show that patients with diabetes mellitus had an increased hazard of death following stroke, where their risk of death was 1.5 times higher than patients who do not have diabetes. The confidence interval suggests this increase in hazard could be as little as 1.1 times higher, or up to over double the hazard. This finding is consistent with the data exploration, where the survival curves in Figure 4.3(c) showed that patients with diabetes had lower survival rates over the 5 year follow-up.

Side of lesion was also found to be significant for survival in the adjusted model, where compared to patients with no lesion, those most at risk of death were patients with a lesion on both sides of the brain; their hazard of death was almost double the hazard of those with no lesion. The 95% confidence interval shows this increase could be as little as 2%, or give a risk of up to 3 and half times higher, as shown in Figure 4.26. The effects of having a lesion either on the left side, or on the right side of the brain, were not found to be significant for survival, however the hazard ratios given in Table 4.17 imply that, compared to no lesion, a right-sided lesion could reduce hazard of death, whereas as a left-sided lesion would increase hazard.

The results of the pooled model also indicate that in the adjusted setting, for each 10mmHg increase in systolic BP at admission hospital, the hazard of death following stroke was reduced by almost 5%, with a hazard ratio of 0.95 given in Table 4.17. This finding coincides with the findings of the univariate, and age and sex adjusted, Cox models for systolic BP given in Table 4.13.

Lesion type shown in CT scan was also found to be important for survival in the adjusted model, where patients with no lesion were the baseline comparison. Figure 4.26 clearly shows that patients with the highest hazard of death were those who did not have a CT scan, where their hazard of death was 2 and half times higher than those found to have no lesion. Patients with a HCI had the next highest hazard of death, with a hazard ratio of 2.2, and patients diagnosed with PICH had a 84% higher risk of death than those with no lesion. CI was the only lesion type found to not be significant for survival at the 5% level, where the confidence interval spans one, however, the hazard ratio for CI given in Table 4.17 suggest patients with a CI had a 29% increase in hazard of death.

Finally, worst consciousness level in the first 24 hours post-stroke was found to be highly significant for survival in the adjusted Cox model, where worsening consciousness levels resulted in increased risk of death. Compared to alert patients, the hazard ratio for drowsy patients suggests they had a 90% higher risk of death. Further, the hazard ratios in Table 4.17 show that, compared to alert patients, stupor patients were over twice as likely to die post-stroke, and risk of death for patients in a coma was over quadruple that of alert patients. These findings regarding the effects of consciousness level on survival post-stroke are reiterated by the Kaplan-Meier survival curves in Figure 4.5(a), fitted during data exploration. However, like pre-stroke mobility, the adjustment for other factors resulted in a reduction in the level of increase of the hazard ratios for the consciousness levels, compared to the univariate Cox model, given in Table 4.15.

Overall, the results of the model building have shown that on adjustment for other risk factors, several variables are important for survival post-stroke. The findings show concurrence with the data exploration, whilst also highlighting the need for adjusted effects over univariate models. The next step is to assess the fit of this model by carrying out diagnostic checks.

4.6 Model Validation

In order to assess the fit of the pooled model, the key assumption to check is the proportional hazards assumption. This can be done through the test and visualisation outlined in Section 3.2.6. The pooled model gives the combined estimates of 10 Cox PH models fitted to each of the multiply imputed data sets, and current methodology allows for the proportional hazards assumption to be assessed by examining the Schoenfeld residuals and conducting the score test on each of the 10 Cox models separately.

Through use of the `cox.zph` function, here we present the results of the formal test, given in Table 4.18, for the model fitted to data set 1. In an attempt to gauge the fit of the pooled proportional hazards model, this test was carried out for each of the 10 imputed data sets, along with Schoenfeld residual plots, in order to assess the proportional hazards assumption in the models fitted to each of the imputed data sets.

The results in Table 4.18 show that several variables violated the proportional hazards assumption within the model fitted to imputed data set 1. These variables include age, the needs help level of pre-stroke mobility, no scan as a level of lesion type shown in CT scan and the stupor and coma levels for worst consciousness level, where each of these violated the proportional hazards assumption significantly at the 5% level. Table 4.18 additionally shows a marginal result for PICH, with a p-value less than 0.1, and hospital and systolic BP also have larger test statistics.

The results for imputed data set 1 indicate violation of the proportional hazards, however this only considers one of the imputed data sets. To look at each of the imputed data sets, and consider differences between these in terms of potential violation of the proportional hazards assumption, Table 4.21 presents the p-values produced by the `cox.zph` function for each of the imputed data sets.

Table 4.21 shows that there are clear differences between results for each data set, where for example, a lesion on both sides has a marginal result with a p-

value less than 0.1 for data set 6, but for other data sets the p-value is as large as 0.4. Overall these test shows that there is violation of the proportional hazards assumption in each of the models fitted to the imputed data set, with the global test all showing a violation. This violation and the differences between the p-values highlights the need to be able to assess the pooled model formally, and visually, to gain a better understanding of the violation of the proportional hazards assumption and how it can be handled in further analyses. This motivates the work conducted within Chapter 5, where a further assessment of the proportional hazards assumption is given.

Now, considering the unexpected interpretation, we refer back to the coefficient for hospital discussed in Section 4.5, where we noted that we had an opposing effect of hospital in the adjusted model compared to the univariate model. This indicates a possible interaction between hospital and another covariate included in the adjusted analysis model. On inspection of the variables included in the analysis model, we note the most probable interaction is between hospital and whether or not patients had a CT scan as this is directly related to hospital practice.

Through consideration of the number of patients not given a CT scan at each hospital by the worst consciousness level of patients, a difference in practice can be confirmed. Table 4.19 shows that Hospital 2 had a much higher number of patients who did not receive a CT scan, 130 compared to 44 at Hospital 1. It can also be seen in Table 4.19 that a higher proportion of alert patients were not given a CT scan at Hospital 2 with 64% of the unscanned patients being alert. Whereas only 34% of the unscanned patients were alert at Hospital 1. On the other extreme, 39% of the unscanned patients were in a coma in Hospital 1, compared to 24% in Hospital 2.

To explore how this may relate to survival, an interaction between 'no scan' and hospital can be incorporated into the levels of the CT scan variable. This gives 'no scan' as two levels: no scan at Hospital 1 and no scan at Hospital 2. Table

Table 4.18: Results of the formal test of the proportional hazards assumption, showing the estimates of ρ, χ^2, and the p-values of the score tests for each covariate effect in the adjusted Cox model fitted to imputed data set 1.

Variable	ρ	χ^2	***p*-value**
Age	0.268	33.105	<0.001
Hospital (Hospital 1)			
Hospital 2	0.061	1.547	0.214
Pre-stroke Mobility (200m)			
Indoors	0.006	0.016	0.901
Needs Help	0.153	10.402	0.001
Diabetes (No)			
Yes	-0.039	0.557	0.456
Side of Lesion (None)			
Left	0.068	1.828	0.176
Right	0.013	0.071	0.790
Both	0.040	0.692	0.405
Systolic BP	-0.057	1.692	0.193
CT Scan: Lesion Type (None)			
CI	0.037	0.553	0.457
HCI	-0.007	0.018	0.894
PICH	-0.085	2.812	0.094
No Scan	-0.182	12.248	<0.001
Worst Conscious Level (Alert)			
Drowsy	-0.075	2.216	0.137
Stupor	-0.119	5.567	0.018
Coma	-0.112	4.871	0.027
GLOBAL	NA	83.782	<0.001

4.20 presents the spread of patients across each of the levels, and also gives the incidence of death to compare how the incidence of death varies across the levels. Table 4.20 shows that more patients did not have a CT scan at Hospital 2, but the percentage of those that died who did not have a CT scan was higher for Hospital 1 compared to Hospital 2; 89% compared to 84%. This difference in incidence of death highlights the need to incorporate this interaction between hospital and no CT scan into further model building procedures for this stroke audit data.

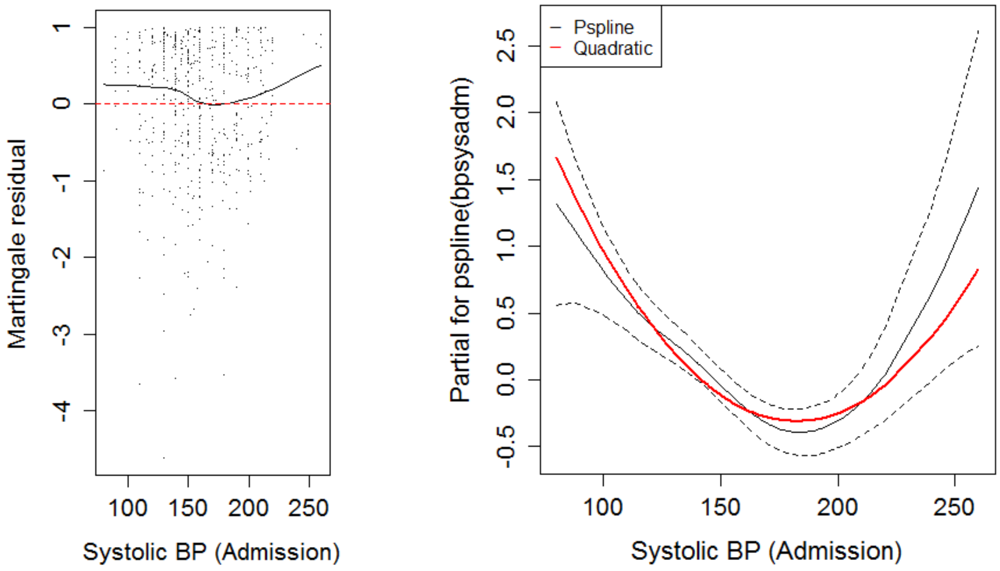
Table 4.19: Frequency table showing number of patients who did not receive a CT scan split by hospital admitted to and worst consciousness level

Hospital	**Consciousness Level**				**Total**
	Alert	Drowsy	Stupor	Coma	
Hospital 1	15	6	6	17	44
Hospital 2	83	7	8	32	130

Table 4.20: Number and percentage of patients within each level of lesion type shown in CT scan and the number and percentage of patients who died within each level, with hospital interaction for 'no scan'.

CT Scan: Lesion Type	**Spread**		**Died**	
	n	%	n	%
No lesion	108	20.1	58	53.7
CI	183	34.0	121	66.1
HCI	17	3.2	13	76.5
PICH	56	10.4	39	69.6
No scan @ Hosp. 1	44	8.2	39	88.6
No scan @ Hosp. 2	130	24.2	109	83.8

A further issue needing consideration was the functional form of systolic BP. A smooth fit of the Martingale residuals against the values of systolic BP, for imputed data set 1, shown in Figure 4.27(a), gives an indication of the functional form of Systolic BP, and shows a quadratic shape. To examine this more thoroughly, the functional form of systolic BP can be visualised by fitting systolic BP as a spline function in a Cox model. Given the Martingale residuals indicate a quadratic effect, the spline fit can be compared to a model fit with a quadratic effect of systolic BP. A plot of the spline fit for systolic BP is given in Figure 4.27(b), with the quadratic effect overlaid. The shape of these curves indicates that the effect of systolic BP on survival post-stroke takes a quadratic form, where extreme values, both high and low, cause an increase in hazard of death.



(a) Martingale Residuals

(b) Spline and Quadratic Fit

Figure 4.27: Visualisation of functional form of systolic BP: (a) Smooth fit of the Martingale residuals against systolic BP, (b) Plot of the spline fit (black) and quadratic fit (red) for the effect of systolic BP on survival.

4.7 Conclusion

This chapter has completed an initial analysis of the stroke audit data, and highlighted issues to be considered in further analyses. Data exploration indicated there were many baseline covariates important for survival post-stroke, with the presence of associations between the baseline covariates highlighting the need for an adjusted analysis.

The exploration of the missing data enabled the assumption of MAR to be concluded, allowing the use of MICE to impute the missing data values. Following imputation, the adjusted analysis demonstrated several baseline covariates to be important for survival post-stroke; these included age, hospital, side of lesion, lesion type shown in CT scan, diabetes mellitus, systolic BP at admission to hospital, pre-stroke mobility and worst consciousness level the first 24 hours post-stroke.

Interpretation of the analysis model in context resulted in the conclusion that hazard of death following stroke would be increased by higher age, a worse consciousness level, poorer mobility prior to stroke, and a lower systolic BP measurement on admission to hospital. Other findings indicated that a lesion on both sides of the brain would result in a higher hazard, along with not having a CT scan.

During validation of this model, however, several of these effects were found to be dependent on time post-stroke, violating the proportional hazards assumption. These results were for the models fitted to each imputed data set separately, and difference in the results between these motivates the work in Chapter 5 for assessing the proportional hazards assumption of the pooled model. Further, during model validation, an interaction between 'no scan' and hospital was found, and systolic BP was shown to have a quadratic functional form.

To resolve the issues found during model validation, and account for them appropriately in the analysis model, the imputation models need to be redefined. The imputation models need to be defined such that they are compatible with an analysis model which can handle time-dependent covariate effects and incorporate quadratic terms. To do this, further theoretical work is required, motivating the methodological developments in Chapters 6 and 7 prior to further analysis in Chapter 8.

Table 4.21: Results of the PH test for adjusted Cox model fitted to each of the 10 imputed data sets (DS), showing the p-values of the test only.

Variable	**DS 1**	**DS 2**	**DS 3**	**DS 4**	**DS 5**	**DS 6**	**DS 7**	**DS 8**	**DS 9**	**DS 10**
Age	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
Hospital 2	0.2140	0.1990	0.1120	0.2180	0.2150	0.1770	0.3030	0.2160	0.2200	0.2800
Indoors	0.9010	0.9270	0.9360	0.8660	0.3640	0.9390	0.7450	0.8220	0.8960	0.8740
Needs Help	0.0013	0.0009	0.0039	0.0013	0.0002	0.0009	0.0004	0.0038	0.0005	0.0006
Diabetes	0.4560	0.1030	0.2190	0.4570	0.2610	0.3410	0.6810	0.2320	0.5910	0.8220
Left	0.1760	0.1870	0.1960	0.1750	0.1970	0.1020	0.0773	0.0692	0.1430	0.1360
Right	0.7900	0.9710	0.9780	0.9160	0.8370	0.6570	0.7030	0.6570	0.7750	0.7500
Both	0.4050	0.3230	0.1000	0.3220	0.3300	0.0773	0.4760	0.1210	0.2860	0.2680
Systolic BP	0.1930	0.1660	0.1350	0.1680	0.1920	0.1270	0.2090	0.2050	0.2050	0.1930
CI	0.4570	0.3980	0.4330	0.4000	0.4420	0.4450	0.4270	0.4510	0.4320	0.3920
HCI	0.8940	0.8590	0.8760	0.9260	0.9790	0.9420	0.9500	0.8820	0.9590	0.9290
PICH	0.0936	0.0768	0.1180	0.0858	0.0842	0.1140	0.1020	0.0677	0.1290	0.1150
No Scan	0.0005	0.0005	0.0004	0.0004	0.0002	0.0004	0.0006	0.0005	0.0006	0.0005
Drowsy	0.1370	0.1290	0.1460	0.1160	0.0863	0.1160	0.1020	0.1320	0.1130	0.1080
Stupor	0.0183	0.0137	0.0302	0.0135	0.0056	0.0204	0.0177	0.0030	0.0172	0.0193
Coma	0.0273	0.0156	0.0089	0.0129	0.0086	0.0087	0.0166	0.0088	0.0200	0.0096
GLOBAL	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001

Chapter 5

Model Validation after Multiple Imputation

5.1 Introduction

This chapter discusses methods of model validation for the Cox proportional hazards model, and extends these model validation techniques to suitably assess the assumptions and fit of a Cox PH model when fitted to multiply imputed data.

After an analysis model has been fitted to data, it is important to assess if the assumptions of this model are satisfied and if the model fits the data well. There are many techniques available to assess the fit of a Cox proportional hazards model, however, fitting a model to multiply imputed data complicates model validation. There is minimal literature regarding validation of the Cox PH model in the setting of multiply imputed data, however given the complex procedures involved to produce multiply imputed data, it is important to ensure the analysis model is appropriate. Assessing the substantive model will not only ensure valid inferences can be made using the model, but will also ensure the imputation models have been fitted to be congenial to a suitable analysis model.

The key assumption of the Cox model is the proportional hazards assumption.

This chapter will primarily focus on assessing the proportional hazards assumption in the multiply imputed data setting, where both a formal test and visualisation tool will be considered. Initially this chapter will overview the current methods used to assess the proportional hazards assumption, leading on to provide alternative methods for application to multiply imputed data.

A simulation study will be presented to compare the test used in current practice, for assessing the proportional hazards assumption, against our proposed alternative for use with multiply imputed data. Application of our proposed alternative test and visualisation tool will be shown, where they will be used to assess the proportional hazards assumption of the pooled Cox model fitted to the stroke audit data.

5.2 Review of Methods to Assess the Proportional Hazards Assumption

As previously discussed in Sections 3.2.4 and 3.2.6, the proportional hazards assumption is the key assumption of the Cox regression model for survival data. Below we discuss several methods previously proposed to check for violation of the proportional hazards assumption in a fitted Cox regression model.

The proportional hazards assumption means that the relative hazard between two individuals is independent of time. More explicitly, for time-fixed covariates, the relative hazard for any two individuals, i and l say, will obey the relationship

$$\frac{h_0(t)e^{\beta' X_i}}{h_0(t)e^{\beta' X_l}} = \frac{e^{\beta' X_i}}{e^{\beta' X_l}}$$

which is independent of time, and this relationship should hold individually for each variable X_k in the model. For a time-varying covariate, proportional hazards means that the hazard for two individuals, expressed as

$$\frac{e^{\beta' X_i(t)}}{e^{\beta' X_l(t)}},$$

is not independent of time, however the relative impact of any two values of a covariate can still be given by a single coefficient β. In other words, the coefficient β for each covariate must be constant over time for the proportional hazards assumption to be satisfied.

An alternative way to look at proportional hazards assumption is to consider the form of the log-hazard function, expressed by Schoenfeld (1982) as

$$\log\left[h(t, \boldsymbol{X}, \boldsymbol{\beta})\right] = \log\left[h_0(t)\right] + \boldsymbol{X}'\boldsymbol{\beta}.$$

Considering a simple example, as given by Hosmer et al. (2008), it can be seen how the proportional hazards assumption can be assessed. Assume the Cox regression model contains a single binary covariate. Plotting the log-hazard over time would result in two curves; $\log\left[h_0(t)\right]$ when $X = 0$ and $\log\left[h_0(t)\right] + \beta$ when $X = 1$. Therefore, regardless of the form of the baseline hazard function, the difference between these curves at any point in time will be β. This concept works for further variable types, and hence generally, assessing the proportional hazards assumption is simply an examination of the extent to which plotted log-hazard functions are equidistant from each other over time.

Previously there have been many methods suggested for assessing the proportional hazards assumption, with Hess (1995) outlining eight graphical methods suggested for assessing this assumption. Firstly, one method suggested was to compare survival estimate through plotting the predicted survival curves based on the Cox model along with non-model-based estimates of the observed survival curves, such as Kaplan-Meier estimates. Early research by Breslow (1984) and Harrell and Lee (1986) suggested any departures when comparing these estimates would give evidence against the proportional hazards assumption, however, judging whether these discrepancies are a result of sampling fluctuations or real trends was not always possible.

Hess (1995) highlighted further approaches using the cumulative hazard functions, where plotting this against time, or against each other, allowed checking for a constant ratio, or constant slope respectively. Further, considering the log of the cumulative hazard functions, a plot of this against time could check for parallelism and a plot of difference in log cumulative hazard function against time could check for constancy. These methods can provide an indication regarding violation of the proportional hazards assumption but do not allow for inference about the shape of $\beta(t)$, a potentially time-varying coefficient.

A method suggested to be more direct by Harrell (2006) is to partition the follow-up time into intervals and fit models to each interval. There have been several approaches outlined for this method, including Gore et al. (1984), Kay (1986) and Anderson and Senthilselvan (1982), where these approaches allow the log hazard ratio, or Cox regression coefficient, to be a function of time by fitting specially stratified Cox models. This method can result in difficulties around choosing the number and location of the breakpoints between time intervals.

Incorporating a time-by-covariate interaction into the Cox regression model can also be used to assess the proportional hazards assumption, where choice of an appropriate functional form of time can be critical (Hess, 1995). Plotting the estimated relative hazard function against time can determine the magnitude of any violation of the proportional hazards assumption, where a formal test can assess if the coefficient of the interaction term is significant.

Finally, plotting the Schoenfeld residuals against time can be used to assess the proportional hazards assumption, where any trends shown in such a plot would indicate time dependence in covariate effects (Schoenfeld, 1982). Grambsch and Therneau (1994) developed this further, suggesting a scaled adjustment of the

Schoenfeld residuals. This adjustment enables a smoothing procedure to be carried out, improving interpretation of the plots. Grambsch and Therneau (1994) further derived a weighted residual score test analogous to generalised least squares which is now commonly used in practice. The methods outlined by Grambsch and Therneau (1994) are widely used in practice, and Ng'andu (1997) showed the weighted residual score test to have good power, with the added benefit of corresponding graphical plots which can be used to augment the results.

Hosmer et al. (2008) recommend use of both a visualisation and formal test together to support each other in assessing the proportional hazards assumption, therefore, we focus upon the methods outlined by Grambsch and Therneau (1994) and firstly consider how these can be derived. Expressing a model with a time-dependent coefficient,

$$h_i(t) = h_0(t) \exp(\boldsymbol{\beta}(t)'\boldsymbol{x}_i),$$

enables us to test for non-proportional hazards, by assessing if $\beta_k(t)$ is constant. Therefore, proportional hazards can be expressed as the restriction $\beta_k(t) = \beta_k$, where β_k is constant. If $\beta_k(t)$ is not constant, this suggests that the impact of the kth explanatory variable on the hazard may vary over time, violating the assumption of proportional hazards. If the proportional hazards assumption holds then a plot of $\beta_k(t)$ against time will be a horizontal line.

Hosmer et al. (2008) suggest the proportional hazards assumption should be assessed as a two-step procedure, where the results of the two steps should support each other. The steps to assess whether the proportional hazards assumption has been violated are visualising the relationship between $\beta_k(t)$ and time, and calculating a formal test statistic. Both these methods require the use of the Schoenfeld residuals. In Section 3.2.6, the Schoenfeld residual for the ith individual and kth covariate was defined as

$$s_{ki} = \delta_i \left\{ x_{ki} - \frac{\sum_{l \in R(t_i)} x_{kl} \exp(\hat{\boldsymbol{\beta}}' \boldsymbol{x}_l)}{\sum_{l \in R(t_i)} \exp(\hat{\boldsymbol{\beta}}' \boldsymbol{x}_l)} \right\}.$$

As Schoenfeld residuals are only non-zero at uncensored observations, rather than defining Schoenfeld residuals for each individual i, they can instead be defined for each event time t_j, $j = 1, ..., d$, where d is the total number of events. Letting $\boldsymbol{s}_j$ be the $q \times 1$ vector of Schoenfeld residuals for the jth event time, and $V(\beta, t)$ be the variance matrix of X at time t, Grambsch and Therneau (1994) show that, for the estimated coefficient $\hat{\beta}$ from a standard Cox regression model,

$$\mathbb{E}(s^*_{jk}) + \hat{\beta}_k \approx \beta_k(t_j),$$

where s^*_{jk} is the scaled Schoenfeld residual, defined as $V^{-1}(\hat{\beta}, t_j) s_j$, and $\beta_k(t_j)$ is the time-varying coefficient of X_k at the jth event time t_j.

This suggests that in order to examine the form of the time varying coefficient $\beta_k(t_j)$ of covariate X_k, a plot of $s^*_{jk} + \hat{\beta}_k$ versus time, or alternatively some function of time, say $G(t)$, can be used. This provides a method of visualising the nature and extent of non-proportional hazards. Further, fitting a line to the plot and testing for non-zero slope can be used to assess the proportional hazards assumption. A non-zero slope gives evidence that the proportional hazards assumption has been violated. This can lead to a more formal test.

Therneau and Grambsch (2013) suggest that an analogy to generalised least squares can be used to motivate a formal test statistic to carry out a score test, where as discussed in Section 3.2.6, the linear dependence of the coefficient of X_k on time can be expressed as a regression on some function of time $G(t)$, giving

$$\beta_k(t) = \beta_k + \theta_k G(t), \tag{5.1}$$

with the null hypothesis for proportional hazards corresponding to $\theta_k = 0$, for $k = 1, ..., q$.

In order to define the test statistic, first let G_j be a $q \times q$ diagonal matrix with diagonal entries $G(t_j)$, the function $G(t)$ evaluated at the jth event time, $j = 1, ..., d$. Letting $\hat{V}_k = V(\hat{\beta}, t_j)$ be the variance of the estimated coefficient at time t_j, Therneau and Grambsch (2013) suggest that θ can be estimated by

$$\hat{\theta} = Q^{-1} \sum_{j=1}^{d} G_j s_j, \tag{5.2}$$

where

$$Q = \sum_{j=1}^{d} G_j \hat{V}_j G_j - \left(\sum_{j=1}^{d} G_j \hat{V}_j \right) \left(\sum_{j=1}^{d} \hat{V}_j \right)^{-1} \left(\sum_{j=1}^{d} G_j \hat{V}_j \right)', \tag{5.3}$$

and Q^{-1} is the estimator of the variance of $\sum_{j=1}^{d} G_j s_j$.

Multivariate linear regression would estimate θ by

$$\tilde{\theta} = \left(\sum_{j=1}^{d} G_j \hat{V}_j G_j \right)^{-1} \sum_{j=1}^{d} G_j s_j,$$

however, since the Schoenfeld residuals are constrained to sum to zero, $\sum_{j=1}^{d} s_j = 0$, Q is defined in Equation (5.3) to add a correction term to $\sum_{j=1}^{d} G_j \hat{V}_j G_j$ accounting for the covariance among the scaled Schoenfeld residuals. The matrix Q^{-1} gives a consistent estimator of the variance of $\sum_{j=1}^{d} G_j s_j$ under the null hypothesis, and so a standard test of the null is

$$T = \hat{\theta}' Q \hat{\theta}, \tag{5.4}$$

where T has an asymptotic χ^2 distribution when the proportional hazards assumption holds. This test statistic can be used as a global test of the proportional hazards assumption for the Cox regression model, or can be used to assess proportionality for individual covariates within the model.

Rubin's rules cannot be directly applied to this test statistic, and thus in the multiple imputation setting, this needs further consideration to obtain a test of the proportional hazards assumption of the pooled substantive model, globally or univariate. It is possible to combine χ^2 statistics, using the method outlined by Li et al. (1991a), however this method is deficient compared to the use of Rubin's rules and the Wald test, as highlighted by Marshall et al. (2009). The amount of information wasted from using only a χ^2-statistic results in a loss of power of the significance test (Meng and Rubin, 1992), therefore, when available, point estimates and the covariance matrix should be combined to allow for use of the Wald test.

5.3 Assessing Proportional Hazards Assumption in Multiply Imputed Data Setting

When obtaining a pooled Cox regression model from M models fitted to M imputed data sets, the combined estimates consist of the pooled model coefficients and corresponding variance terms. This means that we do not have any combined information to inform regarding the model assumptions, such as the Schoenfeld residuals. Any additional informative features produced in the model fitting procedure remain as separate objects which can only be retrieved from the M fitted models, not the pooled model itself. Therefore in order to assess the proportional hazards assumption of a pooled Cox regression model, we need to assess the assumption separately in the M models fitted to each of the M imputed data sets, and combine the results of these to achieve a pooled test result.

As previously discussed, Rubin's rules cannot be directly applied to the test statistic outlined by Therneau and Grambsch (2013), given in Equation (5.4), to test the proportional hazards assumption within a pooled Cox regression model. An analogy to linear regression suggests a natural way to combine test statistics, combining individual regression coefficients and the corresponding variance estimates using Rubin's rules. We seek to show it is possible to test the proportional hazards assumption using an anology to linear regression, where the Wald test, as opposed to a score test, can be used to assess the proportional hazards assumption in the multiple imputation setting.

Further, with multiple sets of Schoenfeld residuals, one for each model fitted to each of the imputed data sets, visualisation of the nature and extent of non-proportional hazards for the pooled model requires further consideration to enable provision of a single combined visual assessment for each covariate effect in the pooled model.

5.3.1 Formal Test

In order to assess the proportional hazards assumption of a pooled model fitted to multiply imputed data, an analogy to linear regression suggests a natural way in which estimates can be combined using Rubin's rules, and under which the Wald test can apply. Using the approach outlined by Therneau and Grambsch (2013), we show that through use of a constant scaling factor for the Schoenfeld residuals, and by centering the function of time, $G(t)$, it is possible to test the proportional hazards assumption using standard linear regression of the scaled Schoenfeld residuals on some function of time. The results of this is that Rubin's rules can be applied to the coefficients and variance estimates, and the Wald test can be used to assess the null hypothesis of proportional hazards.

In order to show this, we take the formula for the variance estimator Q, given in Equation (5.3), and derive the score test statistic given in Therneau and Grambsch (2013) through use of a constant scaling factor and centering the function of time. This derivation also shows how the Wald test statistic can be used to assess the proportional hazards assumption through the use of linear regression models, with an extension given for use after the multiple imputation procedure.

Considering a constant scaling factor

The scaled Schoenfeld residuals are defined as

$$s^*_{jk} = \hat{V}_j^{-1} s_j,$$

where $\hat{V}_j = V(\hat{\beta}, t_j)$ is the variance matrix of a covariate, X, at time t_j. In reality, these $\hat{V}_j$'s can be unstable, particularly at the later few event times, therefore substituting this for a constant scaling factor can make it more stable. Denoting the Fisher's Information matrix for $\hat{\beta}$ as $\mathcal{I}(\hat{\beta})$, where $\sum_{j=1}^{d} \hat{V}_j = \mathcal{I}(\hat{\beta})$, an appropriate substitution would be to take the average of the $\hat{V}_j$'s, giving $\bar{V} = \mathcal{I}(\hat{\beta})/d$. This approximation results in the scaled Schoenfeld residuals being defined as

$$s^*_j = d\mathcal{I}^{-1}(\hat{\beta}) s_j, \tag{5.5}$$

and substituting $\bar{V}$ into the formula for Q as defined in Equation (5.3), we get

$$\begin{aligned} Q &= \sum_{j=1}^{d} G_j \bar{V} G_j - \left(\sum_{j=1}^{d} G_j \bar{V} \right) \left(\sum_{j=1}^{d} \bar{V} \right)^{-1} \left(\sum_{j=1}^{d} G_j \bar{V} \right)' \\ &= \sum_{j=1}^{d} G_j \bar{V} G_j - \left(\sum_{j=1}^{d} G_j \bar{V} \right) \left(d\bar{V} \right)^{-1} \left(\sum_{j=1}^{d} G_j \bar{V} \right)'. \end{aligned} \tag{5.6}$$

We can further simplify this by using $\bar{V} = \mathcal{I}(\hat{\beta})/d$, and that $(d\bar{V})^{-1}$ can be simplified to $\mathcal{I}^{-1}(\hat{\beta})$ since $(d\bar{V})^{-1} = (d\mathcal{I}(\hat{\beta})/d)^{-1}$. Working through algebraically, we can simplify the formula for Q in Equation (5.6) as follows:

$$\begin{aligned} Q &= \sum_{j=1}^{d} G_j \left(\mathcal{I}(\hat{\beta})/d \right) G_j - \left[\sum_{j=1}^{d} G_j \left(\mathcal{I}(\hat{\beta})/d \right) \right] \mathcal{I}^{-1}(\hat{\beta}) \left[\sum_{j=1}^{d} G_j \left(\mathcal{I}(\hat{\beta})/d \right) \right]' \\ &= d^{-1} \sum_{j=1}^{d} G_j \mathcal{I}(\hat{\beta}) G_j - d^{-2} \left(\sum_{j=1}^{d} G_j \mathcal{I}(\hat{\beta}) \right) \mathcal{I}^{-1}(\hat{\beta}) \left(\sum_{j=1}^{d} G_j \mathcal{I}(\hat{\beta}) \right)'. \end{aligned}$$

Since $\mathcal{I}(\hat{\beta})$ is independent of j, and G_j is a diagonal matrix, we then have:

$$\begin{aligned} Q &= d^{-1}\sum_{j=1}^{d} G_j\mathcal{I}(\hat{\beta})G_j - d^{-2}\left(\sum_{j=1}^{d} G_j\right)\left(\sum_{j=1}^{d} G_j\mathcal{I}(\hat{\beta})\right)' \\ &= d^{-1}\sum_{j=1}^{d} G_j\mathcal{I}(\hat{\beta})G_j - d^{-2}\sum_{j=1}^{d} G_j\mathcal{I}(\hat{\beta})G_j. \end{aligned}$$

Finally, pulling like terms together results in:

$$Q = \left(\frac{d-1}{d^2}\right)\sum_{j=1}^{d} G_j\mathcal{I}(\hat{\beta})G_j. \tag{5.7}$$

For a large number of events, or large d, the multiplicative term in Equation (5.7) can be approximated as

$$\frac{d-1}{d^2} \approx \frac{d}{d^2} = \frac{1}{d},$$

resulting in an approximation for Q simplified as

$$Q \approx d^{-1}\sum_{j=1}^{d} G_j\mathcal{I}(\hat{\beta})G_j.$$

Substituting this approximation for Q into Equation (5.2) would give a simpler approach to estimating θ; one which has potential to be combined using Rubin's rules. This result is not exact however and needs further consideration to avoid the need for the assumption of large d. Centering the function of time can be used to obtain an exact result, as shown in the following section.

Centering the function of time

Using the constant scaling factor $\bar{V} = \mathcal{I}(\hat{\beta})/d$ has allowed us to derive an approximate simplification of Q for large values of d, however, this simplification can be made exact for any value of d if we consider centering the function of time, $G(t_j)$.

Let g_j be a $q \times q$ diagonal matrix, with diagonal entries $g(t_j)$, a function of time $g(t)$ evaluated at the jth event time. Now suppose $G(t_j)$ is centered, so that $G_j = g_j - \bar{g}$, where $\bar{g} = d^{-1}\sum_{j=1}^{d} g_j$. Now considering the formula for Q given in Equation (5.6), with the initial substitution of the constant scaling factor $\bar{V}$, we can work through algebraically to achieve an exact simplification of Q by substituting in the centered function of time. This gives:

$$Q = \sum_{j=1}^{d}(g_j - \bar{g})\bar{V}(g_j - \bar{g}) - \left(\sum_{j=1}^{d}(g_j - \bar{g})\bar{V}\right)\left(\sum_{j=1}^{d}\bar{V}\right)^{-1}\left(\sum_{j=1}^{d}(g_j - \bar{g})\bar{V}\right)'. \quad (5.8)$$

Again, considering the Fisher's Information matrix, we have $\bar{V} = \mathcal{I}(\hat{\beta})/d$. Substituting this into Equation (5.8) gives

$$\begin{aligned} Q = &\sum_{j=1}^{d}(g_j - \bar{g})\left(\mathcal{I}(\hat{\beta})/d\right)(g_j - \bar{g}) \\ &- \left(\sum_{j=1}^{d}(g_j - \bar{g})\left(\mathcal{I}(\hat{\beta})/d\right)\right)\mathcal{I}^{-1}(\hat{\beta})\left(\sum_{j=1}^{d}(g_j - \bar{g})\left(\mathcal{I}(\hat{\beta})/d\right)\right)'. \end{aligned}$$

Now, as $\mathcal{I}(\hat{\beta})/d$ is independent of the event times, j, we can simplify Q further to get

$$\begin{aligned} Q = d^{-1}&\sum_{j=1}^{d}(g_j - \bar{g})\mathcal{I}(\hat{\beta})(g_j - \bar{g}) \\ &- d^{-2}\left[\mathcal{I}(\hat{\beta})\sum_{j=1}^{d}(g_j - \bar{g})\right]\mathcal{I}^{-1}(\hat{\beta})\left[\mathcal{I}(\hat{\beta})\sum_{j=1}^{d}(g_j - \bar{g})\right]'. \end{aligned}$$

At this stage, the centering of the function of time enables us to achieve an exact simplification of the correction factor within Q. By definition of $\bar{g}$, we have $\sum_{j=1}^{d}(g_j - \bar{g}) = 0$, and therefore

$$d^{-2}\left[\mathcal{I}(\hat{\beta})\sum_{j=1}^{d}(g_j - \bar{g})\right]\mathcal{I}^{-1}(\hat{\beta})\left[\mathcal{I}(\hat{\beta})\sum_{j=1}^{d}(g_j - \bar{g})\right]' = 0.$$

We have now shown that the correction term becomes zero after using a constant scaling factor and centering the function of time to achieve an exact simplification of Q as

$$Q = d^{-1} \sum_{j=1}^{d} (g_j - \bar{g}) \mathcal{I}(\hat{\beta}) (g_j - \bar{g}) .$$

By Equation (5.2), we can estimate θ as $\hat{\theta} = Q^{-1} \sum_{j=1}^{d} G_j s_j$, so when the scaling factor is constant and the time function is centered, θ can be estimated as

$$\hat{\theta} = \left[d^{-1} \sum_{j=1}^{d} (g_j - \bar{g}) \mathcal{I}(\hat{\beta}) (g_j - \bar{g}) \right]^{-1} \sum_{j=1}^{d} G_j s_j . \tag{5.9}$$

The constant scaling factor, $\bar{V} = \mathcal{I}(\hat{\beta})/d$, now also means that the scaled Schoenfeld residuals, s_j^*, can be defined as $s_j^* = \bar{V}^{-1} s_j$. Using the scaled Schoenfeld residuals within our estimation of θ gives

$$\hat{\theta} = \frac{\sum_{j=1}^{d} (g_j - \bar{g}) s_j^*}{\sum_{j=1}^{d} (g_j - \bar{g})^2} . \tag{5.10}$$

The estimate, $\hat{\theta}$, given in Equation (5.10) shows that under the conditions of using a constant scaling factor, $\bar{V}$, and centering the function of time, the proportional hazards assumption can be tested using a simple linear regression model of the scaled Schoenfeld residuals against a centered function of time, with no intercept term. This now provides a simple approach for testing the proportional hazards assumption, where the linear regression models give estimates of the coefficients and variance terms to which Rubin's rules can be easily applied. Combining these estimates enables assessment of the proportional hazards assumption for a pooled Cox regression model, fitted to multiply imputed data, using the Wald test.

Now, as Hosmer et al. (2008) recommend, covariate specific tests should be calculated. For a particular covariate, X_k, we have a linear regression model of the form

$$s_{jk}^* = \theta_k (g_j - \bar{g}) + \epsilon , \tag{5.11}$$

where θ_k is estimated by

$$\hat{\theta}_k = \frac{\sum_{j=1}^{d} (g_j - \bar{g})\, s^*_{jk}}{\sum_{j=1}^{d} (g_j - \bar{g})^2}.$$

The univariate test of the proportional hazards assumption for the kth covariate can therefore be based upon the linear regression model given in Equation (5.9), and the Wald test, with the test statistic T_k given as

$$T_k = \frac{\hat{\theta}_k^2}{\mathrm{Var}(\hat{\theta}_k)}.$$

Grambsch and Therneau (1994) suggested that since the (k, k)th element of Q^{-1} is a good estimator of the variance of $\hat{\theta}_j$, the test statistic can be defined as

$$T_k = \frac{\left\{\sum_{j=1}^{d} (g_j - \bar{g})\, s^*_{jk}\right\}^2}{d\mathcal{I}^{-1}(\hat{\beta})_{(k,k)} \sum_{j=1}^{d} (g_j - \bar{g})^2},$$

where $d\mathcal{I}^{-1}(\hat{\beta})_{(k,k)}$ gives the variance of $\sum_{j=1}^{d} (g_j - \bar{g})\, s^*_{jk}$ under the null. The difference between this test statistic and the Wald test statistic calculated directly from the linear regression is related to the variance term $d\mathcal{I}^{-1}(\hat{\beta})_{(k,k)}$. This variance relates to the the Fisher's Information matrix from the Cox regression model itself as opposed to the linear regression of the residuals on time, and thus the test statistic is not directly calculated from the linear regression model.

As previously discussed, Rubin's rules do not apply directly to the test statistics, and it is not sensible to combine the test statistic directly as this results in wasted information and a loss power of the test. We instead suggest that focus should be on using Rubin's rules to combine the coefficients and variance estimates from the linear regressions of the residuals against time.

In order to test the proportional hazards assumption for a particular covariate, k, in a pooled Cox regression model, we suggest that the linear regression model given in Equation (5.11) should be fitted to the scaled Schoenfeld residuals of each of the M Cox regression models separately. This obtains M estimates of

θ_k, denoted as $\hat{\theta}_k^{(m)}$, and M corresponding variance estimates, $\text{Var}(\hat{\theta}_k^{(m)})$, for $m = 1, ..., M$. Typically, if $\hat{\theta}$ is calculated using maximum likelihood estimation, this variance estimate, $\text{Var}(\hat{\theta}_k^{(m)})$, will the be the inverse Fisher's Information matrix of the regression.

Here Rubin's rules can be applied to obtain pooled coefficient and variance estimates, where the combined coefficient is given as

$$\hat{\theta}_k^* = \frac{1}{M} \sum_{m=1}^{M} \hat{\theta}_k^{(m)}. \tag{5.12}$$

The variance estimator of $\hat{\theta}_k^*$ is more complex as it needs to take into account both the between-imputation variance and the within-imputation variance. The between-imputation variance gives the variance between the estimates $\hat{\theta}_k^{(m)}$ and can be defined as

$$B_{(\theta_k)} = \frac{1}{M-1} \sum_{m=1}^{M} \left(\hat{\theta}_k^{(m)} - \hat{\theta}_k^* \right)^2, \tag{5.13}$$

and the within-imputation variance corresponds to the average variance of the estimates $\hat{\theta}_k^{(m)}$ and is given by

$$W_{(\theta_k)} = \frac{1}{M} \sum_{m=1}^{M} \text{Var}(\hat{\theta}_k^{(m)}). \tag{5.14}$$

The total variance of $\hat{\theta}_k^*$, denoted $\text{Var}(\hat{\theta}_k^*)$, is then defined as

$$\text{Var}(\hat{\theta}_k^*) = W_{(\theta_k)} + \left(1 + \frac{1}{M} \right) B_{(\theta_k)}. \tag{5.15}$$

Using the combined estimates of the coefficients and variance terms, given in Equations 5.12 and 5.15 respectively, the resulting combined test statistic can be calculated as

$$T_k^* = \frac{\hat{\theta}_k^{*2}}{\text{Var}(\hat{\theta}_k^*)}.$$

Under the null of proportional hazards, T_k^* has an asymptotic χ^2 distribution.

The significance of T_k^* can be assessed using the Wald test to check for violation of the proportional hazards assumption for covariate k in the pooled Cox regression model. This procedure can be carried out for each of the q covariates fitted within the pooled Cox proportional hazards model.

A simulation study is presented in Section 5.4 to compare the use of Wald tests against the score test in assessing the proportional hazards assumption, examining the type I error of each test.

5.3.2 Visual Examination

Now, we covered in Section 5.3.1 how to conduct a formal test of the proportional hazards assumption of a pooled Cox regression model, however as Hosmer et al. (2008) recommend, a visualisation should also be used to support the formal test.

To visualise the nature and extent of non-proportional hazards, it was recommended by Therneau and Grambsch (2013) to plot the scaled Schoenfeld residuals for each covariate against some function of time, where the addition of a smooth curve with confidence bands can aid in detecting departures from proportional hazards.

As in the formal test, in order to produce a visualisation for the pooled Cox regression model, we need to use the Schoenfeld residuals from the M Cox models fitted to the imputed data sets. Current practice would see this visualisation produced for each of the M imputed data sets separately for each of the q covariates. This would result in at least $M \times q$ plots in total, in which each one considers only one particular imputed data set and one covariate effect. We aim to show how an overall visualisation can be produced for each covariate effect within the pooled model by suggesting methods for combining the scaled Schoenfeld residuals and fitting a combined smooth curve to a plot of the combined residuals.

Below we discuss approaches for combining the scaled Schoenfeld residuals and fitting a smooth curve on a combined plot. We consider the current practice for fitting the smooth curve, as outlined by Therneau and Grambsch (2013) and provide an extension of this for use after multiple imputation.

Combining the Scaled Schoenfeld Residuals

Using the M sets of scaled Schoenfeld residuals for each covariate, two possible approaches were considered to produce a single visualisation of proportional hazards for each covariate.

Firstly, as we aim to produce a single plot incorporating all the residuals for a particular covariate, one possible approach would be to overlay all M sets of the scaled Schoenfeld residuals onto a single plot before adding a smoothed curve.

This approach could be highly informative, however it is more likely to be excessive, where each individual with an event would be represented by M points on the plot. The imputation variation would affect the scaling of the residuals, so even those individuals with observed data rather than imputed data, would have slight differences in their residual values for each of the M models. This method would therefore make interpretation difficult due to the amount of points presented within a single plot, particularly if M or d are large. It would also be difficult to distinguish if a group of residuals causing concern were from the same individual or not.

To produce a simpler plot, with more similarity in interpretation to the current practice, an alternative approach would be to take the average scaled Schoenfeld residual for each individual for each covariate and plot these against $g(t)$. Taking the average of the scaled Schoenfeld residuals across the imputations would not be a suitable approach to carry out formal tests given the reliance upon the variance for scaling, however, for visualisation purposes, the average would be sufficient to view the nature of the proportional hazards. This approach would therefore appear to be the more sensible option to allow for easier interpretation and clarity of the form of proportional hazards, and is the method we recommend.

Fitting a Smooth Curve

Further consideration is needed to appropriately add a smooth curve with confidence bands to a plot of combined residuals. It would not be appropriate to fit a smooth curve to the average scaled Schoenfeld residuals, particularly as this would prevent the confidence bands taking into account the between-imputation variance. Instead, a more rigorous approach would be needed.

Therneau and Grambsch (2013) outline that a common approach in statistical software packages to adding the smooth curve and confidence bands onto residual plots is to use a spline fit. In order to define the plotted values of the spline curve firstly suppose U is the matrix of basis vectors for the spline fit of the scaled Schoenfeld residuals on the $g(t_k)$s, and let C be the matrix for the same spline functions evaluated at the v plotting points. Let S be the $d \times q$ matrix of Schoenfeld residuals, so that the matrix of scaled Schoenfeld residuals is defined as $S^* = dS\mathcal{I}^{-1}$ under the simplification of constant variance. The plotted values of the spline curve for the kth covariate can then be defined as

$$\hat{y}_k = \mathbf{1}\hat{\beta}_k + C(U'U)^{-1}U'S_k^* \equiv \mathbf{1}\hat{\beta}_k + HS_k^*,$$

where S_k^* is the kth column of S^* and $\mathbf{1}$ is a v-vector of ones.

Through consideration of the variance matrix of S_k^*, Therneau and Grambsch (2013) provide an estimate for $\mathrm{Var}(\hat{y}_k)$. For notational simplicity, let $\mathcal{I}^{jk} = \mathcal{I}_{jk}^{-1}$, the (j,k)th element of $\mathcal{I}^{-1}$. Under the assumption of constant variance over time and controlling for the constraint of the Schoenfeld residual summing to zero, Therneau and Grambsch (2013) suggest $\mathcal{I}^{kk}\left[dI_d - J_d\right]$ gives a consistent estimator of the variance matrix of S_k^* under proportional hazards, where I_d is the $d \times d$ identity matrix and J_d is the $d \times d$ matrix of ones. Since the Schoenfeld residuals, and therefore also the scaled Schoenfeld residuals, are asymptotically uncorrelated with $\hat{\beta}_k$, we have

$$\mathrm{Var}(\hat{y}_k) = \mathcal{I}^{kk} J_v + \mathcal{I}^{kk} dHH' - \mathcal{I}^{kk} HJ_dH'.$$

Now, it is obvious that the smooth of a constant under the spline function equates said constant, so we have $HJ = J$. This results in the cancellation of terms to get

$$\mathrm{Var}(\hat{y}_k) = d\mathcal{I}^{kk} HH'. \tag{5.16}$$

Therneau and Grambsch (2013) therefore recommend that confidence intervals can be constructed using standard linear model calculations. This is due to the equivalence of Equation (5.16) to the standard linear model formula for the variance of predicted values, with the exception of $d\mathcal{I}^{kk}$ replacing the usual estimator of $\hat{\sigma}^2$.

Now, as the spline fits contain point and variance estimates, $\hat{y}_k$ and $\mathrm{Var}(\hat{y}_k)$ respectively, Rubin's rules can apply to yield a pooled smooth curve with appropriate confidence bands, allowing for extension to the multiple imputation setting. This can be done by fitting the smooth curve to each of the M sets of residuals separately using a spline fit of the scaled Schoenfeld residuals on time. The plotting points of the spline fits and their corresponding variance can then be combined using Rubin's rules to give an overall smooth curve. This curve will represent the form of the hazards of the covariates fitted within a pooled Cox regression model, and indicate if proportionality is satisfied.

The plotting points of the pooled smooth curve can be calculated as the average of the plotting points for the M imputed data sets, giving

$$\hat{y}_k^* = \frac{1}{M} \sum_{m=1}^{M} \hat{y}_k^{(m)}, \tag{5.17}$$

where $\hat{y}_k^{(m)}$ denotes the plotting points for the kth covariate of the mth imputed data set. Further, using Rubin's rules gives the between-imputation variance of

$\hat{y}_k$ to be $B_{(y_k)} = \frac{1}{M-1}\sum_{m=1}^{M}\left(\hat{y}_k^{(m)} - \hat{y}_k^*\right)^2$, and the within-imputation variance to be defined as $W_{(y_k)} = \frac{1}{M}\sum_{m=1}^{M}\text{Var}(\hat{y}_k^{(m)})$. This results in the total variance of $\hat{y}_k^*$ being defined as

$$\text{Var}(\hat{y}_k^*) = W_{(y_k)} + \left(1 + \frac{1}{M}\right) B_{(y_k)}, \tag{5.18}$$

which can then be used to form confidence intervals using standard linear model calculations.

This use of Rubin's rules ensures both the within-imputation and between-imputation variance are fully taken into account in the calculation of the pointwise confidence bands. The pooled smooth curve and confidence bands can then be superimposed onto the plot of the average residuals.

5.4 Simulation Study

In order to test the accuracy of the Wald test compared to the score test for assessing the proportional hazards assumption, here we present a simulation study of the Type I error of these tests. We compare the Type I error of the score test and Wald test used to assess the proportional hazards assumption of complete data, and also present the Type I error of the Wald test used to assess the assumption of a Cox model fitted to multiply imputed data.

We produced 5000 simulations of survival data for $n = 1000$ individuals, where in each simulation three covariates were generated. The first covariate, $X1$, was a continuous covariate simulated from a normal distribution with $\mu = 65$ and $\sigma = 10$, the second covariate, $X2$, was a binary covariate generated using the Bernoulli distribution with probabilities dependent upon the values of $X1$, and the third covariate, $X3$, was simulated as a continuous covariate dependent upon $X1$ and $X2$.

Survival times were generated from covariates $X1$, $X2$ and $X3$ using the `simsurv` package in R under the Weibull distribution. Survival times were generated under the assumption of proportional hazards. Four scenarios were simulated for the survival times, varying the proportion of individuals with events. The percentage of individuals simulated to have events in the four scenarios were 20%, 40%, 60% and 80%.

Cox regression models adjusted for all three covariates, $X1$, $X2$ and $X3$, were fitted in each of the scenarios. The proportional hazards assumption was tested for each covariate in these Cox models using two tests; the test commonly used in practice and our proposed alternative. Respectively, these are the score test in the `cox.zph` function in R, and the Wald test of the scaled Schoenfeld residuals regressed against the centered Kaplan-Meier function of time, as outlined in Section 5.3.1. To compare the performance of these tests, the Type 1 error was recorded for each covariate in each scenario.

Further, to examine the performance of our proposed Wald test method on a model fitted to multiply imputed data, missing data was introduced into covariates $X2$ and $X3$. The proportion of missing observations in each covariate were varied to produce three scenarios: 10% missing in $X2$ and 50% missing in $X3$, 20% missing in $X2$ and 40% missing in $X3$, and 30% missing in both $X2$ and $X3$. Covariate $X1$ remained complete in each scenario.

These three scenarios were simulated for each of the four survival rate scenarios, resulting in a total of twelve scenarios overall. The `mice` package in R was used to impute the missing data in each scenario, where the number of imputations was chosen to be $M = 10$, and the imputations were run over 1000 iterations.

A Cox model adjusted for all three covariates, $X1$, $X2$ and $X3$, was fitted to each of the $M = 10$ imputed data sets for each scenario in each simulation. The proportional hazards assumption was tested using the Wald test method, regressing the scaled Schoenfeld residuals against the centered Kaplan-Meier function of time as outlined in Section 5.3.1. The coefficient and variance estimates of the regressions were combined using Rubin's rules before applying the Wald test. The

Type 1 error was again recorded for each covariate in each scenario.

For each scenario within this simulation study, the Type 1 error was given as the proportion of the 5000 simulations where the p-value of the test of the proportional hazards assumption, for each covariate, was less than the significance level of $\alpha = 0.05$.

5.4.1 Type I Error Results

The results of the simulation study are presented in Table 5.1, where the Type 1 error is presented for the tests of the proportional hazards assumption for each of the covariates, under each of the scenarios outlined.

Initially comparing the Type 1 error between the results of the score and Wald tests for the complete data setting, it can be seen in Table 5.1 that the score and Wald test produce very similar Type 1 errors for each covariate, over the varying proportions of events. The Type 1 errors are all close to 0.05 for the complete data in both tests, thus we can conclude that the Wald test can be used as an alternative to the score test.

Now looking at the simulation study results in the situation where multiple imputation has been used to impute missing values, we can see that the Type 1 error is increased due to the uncertainty around the imputations. In particular, for the lower event proportion scenarios, $X2$ has increased Type 1 error with increased missing data. For larger event proportions, the missing information in covariate $X3$ appears to be causing an increase in Type 1 error for each of the covariates, with the Type 1 error for $X2$ increasing with increased missing data in $X3$, for 60% events and above. The Type 1 error for $X1$ and $X3$ can be seen in Table 5.1 to be increasing for increased missingness in $X3$, regardless of the amount of events. The increase in Type 1 error for $X1$ appears strange given $X1$ was complete in all scenarios, however, this may be explained by the model being adjusted for $X2$ and $X3$ which were incomplete. Any bias introduced through the imputation of $X2$

Table 5.1: Simulation results giving Type 1 error of test of proportional hazards assumption used in current practice and our proposed Wald test method on complete and multiply imputed data (n=1000, 5000 Simulations).

Events	**Missing (%)**		**Type of**	**Type 1 Error**		
(%)	$X2$	$X3$	**Test**	$X1$	$X2$	$X3$
20	0	0	Score Test	0.051	0.052	0.049
	0	0	Wald Test	0.050	0.052	0.049
	10	50	Wald-MI	0.055	0.053	0.052
	20	40	Wald-MI	0.049	0.058	0.049
	30	30	Wald-MI	0.048	0.060	0.045
40	0	0	Score Test	0.057	0.046	0.049
	0	0	Wald Test	0.053	0.046	0.049
	10	50	Wald-MI	0.072	0.054	0.071
	20	40	Wald-MI	0.068	0.054	0.072
	30	30	Wald-MI	0.068	0.060	0.070
60	0	0	Score Test	0.053	0.049	0.051
	0	0	Wald Test	0.050	0.048	0.052
	10	50	Wald-MI	0.102	0.064	0.115
	20	40	Wald-MI	0.096	0.058	0.111
	30	30	Wald-MI	0.090	0.047	0.100
80	0	0	Score Test	0.056	0.045	0.050
	0	0	Wald Test	0.056	0.045	0.052
	10	50	Wald-MI	0.114	0.068	0.151
	20	40	Wald-MI	0.117	0.055	0.156
	30	30	Wald-MI	0.107	0.042	0.140

and $X3$ may influence the association between $X1$ and survival within a model adjusted for all three covariates.

Given the consistency in Type 1 error between the score and Wald test in the complete data setting, the increase in Type 1 error of the test for proportional hazards following multiple imputation indicates that the imputation procedure may not have performed well. Given this, it it is possible that the imputations could be causing bias in the covariate effects for survival, and hence affecting the assumption of proportional hazards. The increase in Type 1 error of the test for violation of the proportional hazards assumption is therefore not necessarily an indication of poor performance of the test itself.

5.5 Application to Stroke Data

This chapter has outlined methods for assessing the proportional hazards assumption after multiple imputation, and here we provide an application of these methods to the model fitted to the stroke data in Section 4.5. A formal test of the proportional hazards assumption for each covariate effect was carried out by regressing the scaled Schoenfeld residuals for each covariate against the centered Kaplan-Meier function of time, as outlined in Section 5.3.1. The visualisation technique described in Section 5.3.2 was also carried out for each covariate effect.

This section presents the results of the formal test and visualisations, giving interpretaion of the results in the context of the model and providing a comparison of these methods against the results of the methods used in current practice.

Table 5.2 presents the results of the formal test, giving the pooled estimates of the regression coefficients of the residuals against time, and their corresponding p-values, calculated using the Wald test. Figures 5.1, 5.2, 5.3 and 5.3 present the visualisations of proportional hazards for each covariate effect, showing plots of the combined Schoenfeld residuals against time, with the pooled spline curves and confidence interval to aid with interpretation.

Firstly focusing upon the formal test, Table 5.2 shows that several covariate effects violate the proportional hazards assumption, where the regression coefficients of the residuals against time are significantly non-zero, with p-values less than 0.05. It can be seen in Table 5.2 that the effect of age violates the proportional hazards assumption, along with the ‘needs help’ level of pre-stroke mobility. Further, ‘no scan’ as a level of lesion type shown in CT scan violates the assumption, as do the effects of levels ‘stupor’ and ‘coma’ in the worst consciousness level covariate. This is the same set of covariate effects highlighted to be in violation of the proportional hazards assumption in Section 4.6. These results imply that the effects of these covariates on survival post-stroke are dependent upon time since stroke.

Table 5.2: Results of the formal test of the proportional hazards assumption, showing the pooled coefficient estimates of the scaled Schoenfeld residuals regressed against time, and their corresponding p-values, for each covariate effect in the adjusted Cox model.

Variable	**Coefficient**	***p*-value**
Age (10 years)	0.148	0.001
Hospital (Hospital 1)		
Hospital 2	0.756	0.265
Pre-stroke Mobility (200m Outdoors)		
Indoors	0.135	0.834
Needs Help	2.675	0.018
Diabetes Mellitus (No)		
Yes	-0.486	0.505
Side of Lesion (None)		
Right	0.981	0.191
Left	0.134	0.857
Both	1.557	0.329
Systolic BP (10mmHg)	-0.012	0.285
CT Scan: Lesion Type (No Lesion)		
CI	0.599	0.463
HCI	-0.148	0.923
PICH	-1.642	0.143
No Scan	-2.867	0.007
Worst Conscious Level (Alert)		
Drowsy	-1.198	0.168
Stupor	-2.229	0.047
Coma	-1.794	0.057

To further understand the violations of the proportional hazards assumption, the visualisations of the scaled Schoenfeld residuals against time can be used to examine the form of the hazards. Figure 5.1 gives these residual plots for age, diabetes mellitus and pre-stroke mobility. The violation of the proportional hazards assumption by the effect of age, determined in the formal test, is reiterated in Figure 5.1(a), which shows an initially increasing smooth curve, before levelling after 100 days post-stroke. The spline curve is below the coefficient reference line prior to 14 days post-stroke, before crossing the line, and the confidence interval of the smooth only contains the coefficient line around the time they cross. This suggests the effect of age on survival may change around this time, however given the shape of the spline fit, the effect of age may change any time within the first 100 days post-stroke.

The effect of diabetes mellitus on survival post-stroke is shown to satisfy the proportional hazards assumption in Figure 5.1(b), where the spline curve is reasonably constant, with the coefficient being within the confidence interval for the whole follow-up period. This coincides with the formal test, where the results in Table 5.2 suggest that the regression coefficient is not significantly different from zero, and thus diabetes mellitus has a constant effect on survival over time.

Figures 5.1(c) and 5.1(d) give the Schoenfeld residual plots for pre-stroke mobility, where Figure 5.1(c) shows the spline fit of the residuals against time to be constant for the 'indoors' level of pre-stroke mobility. This corresponds to the formal test result of this covariate effect satisfying the proportional hazards assumption. Figure 5.1(d) gives the Schoenfeld residual plot for the 'needs help' level of pre-stroke mobility. This displays a spline curve which is initially increasing, before levelling out around 10 to 30 days post-stroke. This again reiterates the findings of the formal test, where the covariate effect of 'needs help' was found to violate the proportional hazards assumption, since the coefficient is not contained within the confidence interval of the smooth at these early times. The shape of the smooth curve in Figure 5.1(d) indicates that there is likely to be a change in effect of 'needs help' on survival within the first month post-stroke.

The Schoenfeld residual plots for systolic BP and side of lesion are given in Figure 5.2. Firstly examining the residual plot for systolic BP in Figure 5.2(a), it can be seen that the coefficient reference line is within the confidence interval of

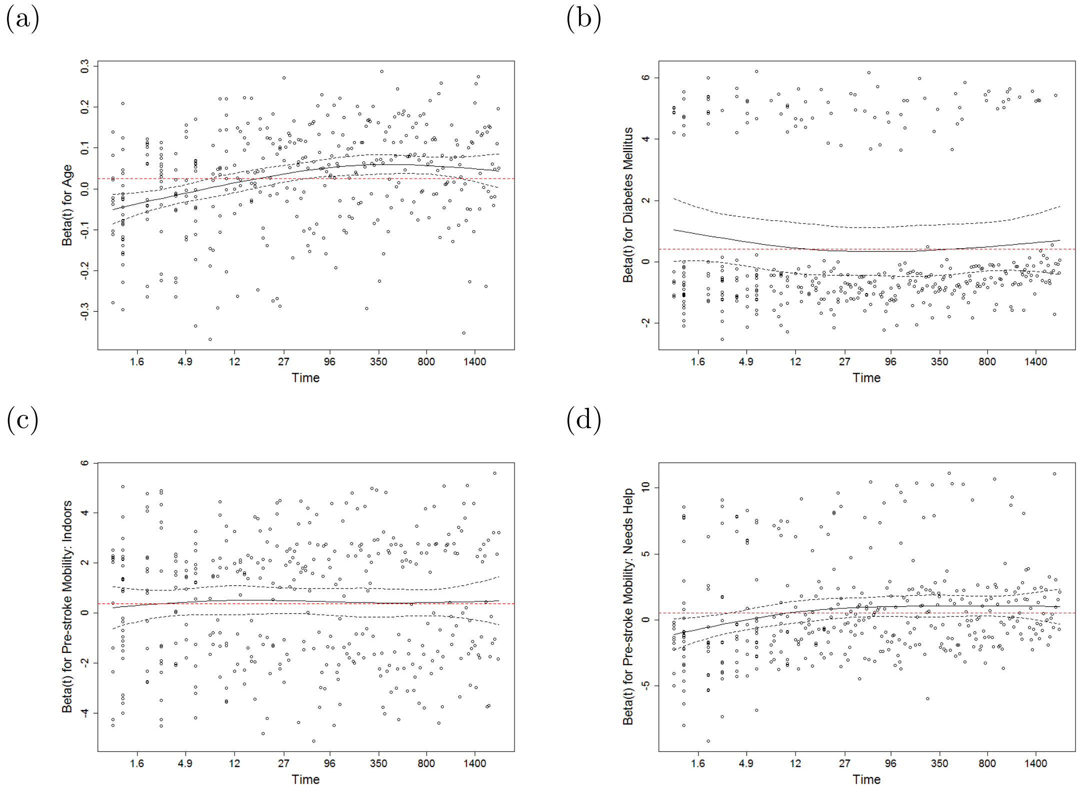


Figure 5.1: Combined plots of scaled Schoenfeld residuals against time (days) with smooth curve to visualise extent of violation of proportional hazards for: (a) Age; (b) Diabetes Mellitus; (c) Pre-stroke Mobility - Indoors; (d) Pre-stroke Mobility - Needs Help.

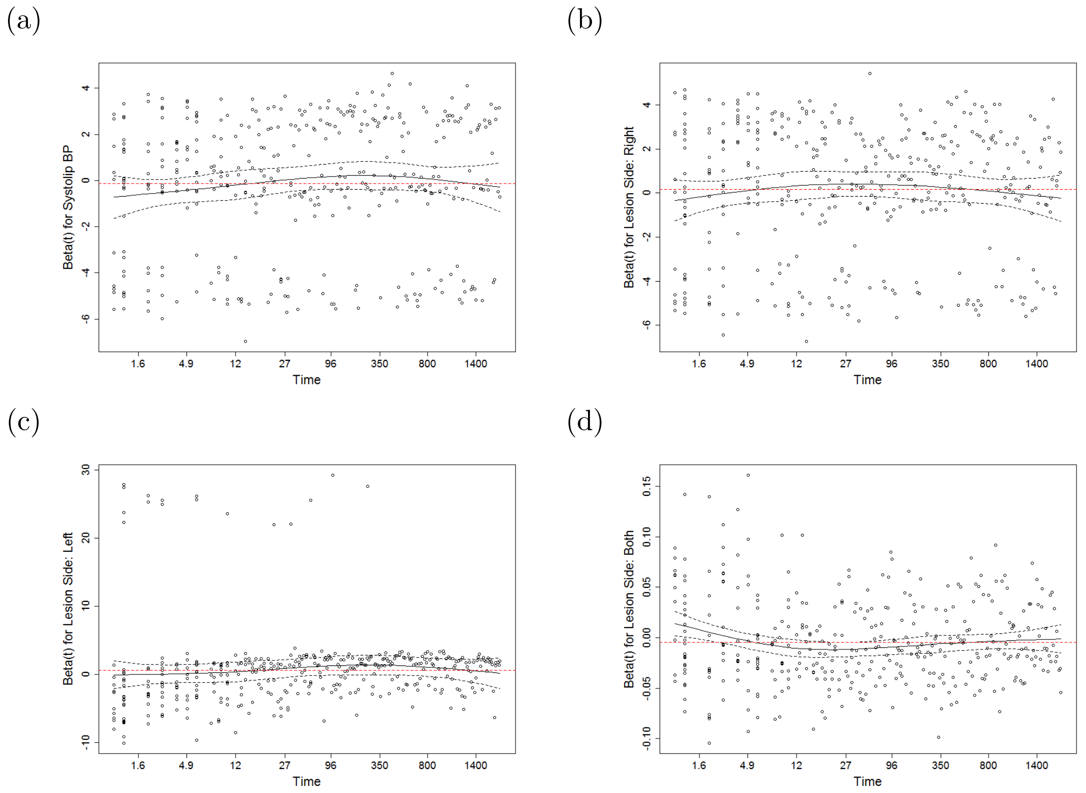


Figure 5.2: Combined plots of scaled Schoenfeld residuals against time (days) with smooth curve to visualise extent of violation of proportional hazards for: (a) Systolic BP; (b) Side of Lesion - Right; (c) Side of Lesion - Left; (d) Side of Lesion - Both.

the spline fit for the whole of follow-up, confirming the findings of the formal test that the covariate effect of systolic BP does not violate the proportional hazards assumption. Given systolic BP is a continuous covariate, the shape of the spline can also indicate any issues around the functional form of the covariate effect. Examining the spline fit closer, Figure 5.2(a) shows that the curve is increasing until around 350 days post-stroke, where the curve then becomes a decreasing function for the remainder of follow-up. This indicates that the effect of systolic BP may not be linear, and the functional form of this covariate will need further consideration.

The effects for the levels of side of lesion were all found to satisfy the proportional hazards assumption in the formal test; this seems reasonable given the Schoenfeld residual plots in Figures 5.2(b), 5.2(c) and 5.2(d). Within each of these plots, the spline curves remain close to the coefficient line for the full follow-up, with the coefficient reference lines mostly being contained with the confidence intervals of the spline fits. The curve in Figure 5.2(d), however, is initially decreasing, with the coefficient line outside the confidence interval, and could indicate non-proportional hazards for early survival times.

Figure 5.3 displays the Schoenfeld residual plots for hospital admitted to and worst consciousness level in the first 24 hours post-stroke. The results of the formal test in Table 5.2 indicated that the effect of hospital did not violate the proportional hazards assumption, which is further asserted by the residual plot, Figure 5.3(a). Although the smooth curve in Figure 5.3(a) appears to be on a slight incline over the 5 year follow-up, the curve remains close to the coefficient line, with the coefficient line remaining inside the confidence interval.

The Schoenfeld residual plots for worst consciousness level show some more interesting results, relating to the violation of proportional hazards shown in Table 5.2. Firstly considering the residual plot for 'drowsy' in Figure 5.3(b), it can be seen that the spline fit is constant initially, before reducing at around 10 days post-stroke to become level again from around 200 days post-stroke. This indicates that there may be a change of effect for survival of ‘drowsy’ patients between the early and late event times, however, the coefficient line remains within the confidence bands of the smooth for the whole follow-up time.

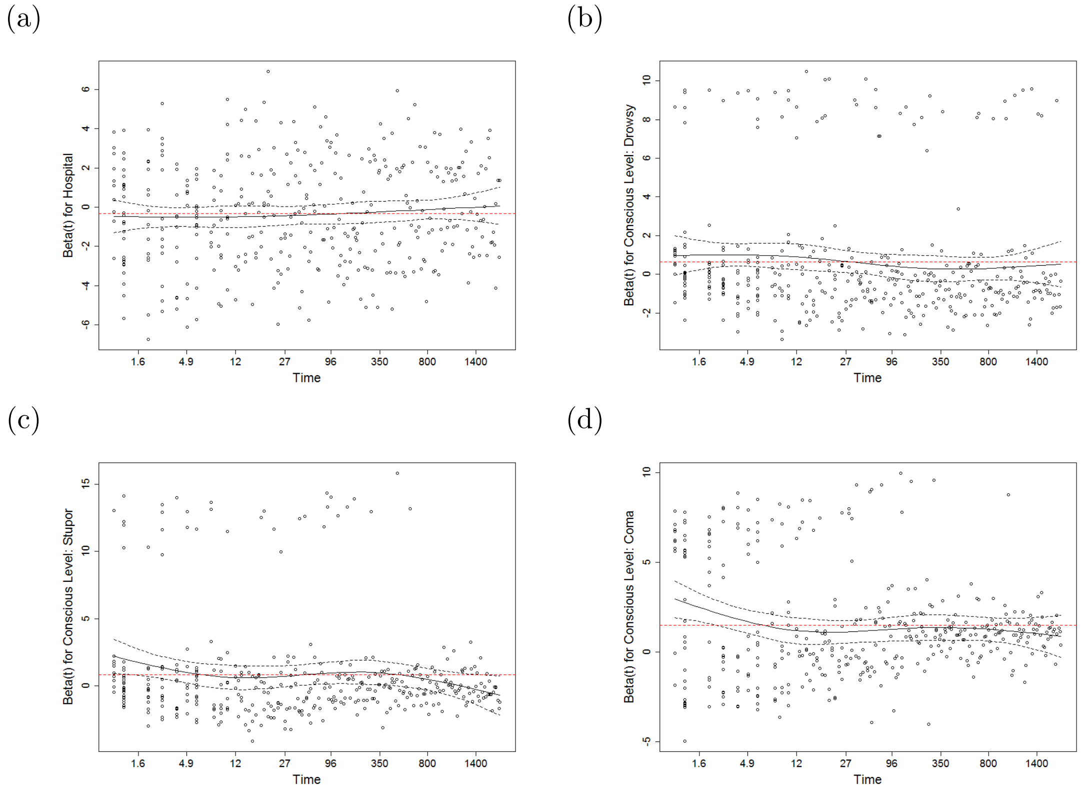


Figure 5.3: Combined plots of scaled Schoenfeld residuals against time (days) with smooth curve to visualise extent of violation of proportional hazards for: (a) Hospital; (b) Conscious Level - Drowsy; (c) Conscious Level - Stupor; (d) Conscious Level - Coma.

The spline fits in Figure 5.3(c) and 5.3(d), for ‘stupor’ and ‘coma’ respectively, appear to follow similar shapes, where the slopes are initially decreasing until around 7 to 14 days post stroke, where there is slight increase or levelling close to the coefficient line before decreasing again towards the end of follow. Within Figure 5.3(c), the coefficient line just remains inside the confidence interval of the spline fit for the follow-up period, but is close to the edge for early and late times. For ‘coma’ in Figure 5.3(d) coefficient line is not contained within the confidence interval for early follow-up times. The change in gradient of this curve is strongest at around one to two weeks post-stroke, where the spline fit then becomes more constant. The shape of these spline fits indicate that there is likely to be a change in effect of both ‘stupor’ and ‘coma’ within the first few weeks post-stroke.

Finally, Figure 5.4 gives the plots of the Schoenfeld residuals against time for the lesion type shown in CT scan variable. Figures 5.4(a) and 5.4(b) give the plots for ‘CI’ and ‘HCI’ respectively. In the formal test, the effects of ‘CI’ and ‘HCI’ were shown to not violate the proportional hazards assumption and this is reiterated in Figures 5.4(a) and 5.4(b). The spline fit in Figure 5.4(a) is on a slight incline, however the curve remains close to the coefficient line, with the confidence interval of the spline fit spanning the coefficient line over the full follow-up period. The spline curve in Figure 5.4(b) also remains close to the coefficient line.

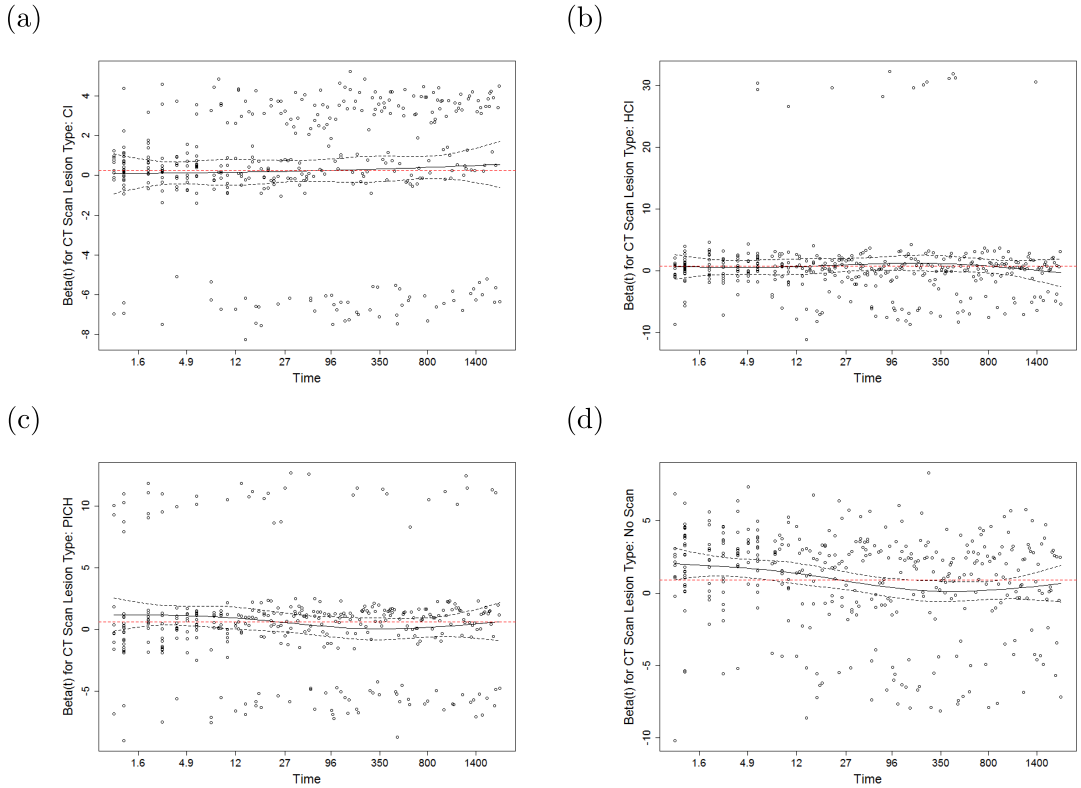


Figure 5.4: Combined plots of scaled Schoenfeld residuals against time (days) with smooth curve to visualise extent of violation of proportional hazards for lesion type shown in CT scan: (a) CI; (b) HCI; (c) PICH; (d) No Scan.

Considering the residual plot for ‘PICH’ in Figure 5.4(c), the shape of the spline curve indicates that there could be a violation of the proportional hazards assumption for the effect of this covariate level. The curve is initially constant, but is above the coefficient line. At around 2 weeks post-stroke, the curve then appears to decline, dropping below the coefficient line and gradually increasing again to reach the coefficient line at the end of follow-up. The results in Table 5.2 indicates that there is not a significant effect of time on the covariate effect for 'PICH' which is consistent with the residual plot as the confidence interval spans the coefficient line for the whole of follow-up. However given the shape of the smooth in Figure 5.4(c), it may be sensible to consider accounting for a possible time-dependent covariate effect for 'PICH', where the residual plot indicates the change in effect could be within the first month post-stroke.

The final level of lesion type shown in CT scan to consider is 'no scan'. The covariate effect of 'no scan' was shown to violate the proportional hazards assumption in the formal test, and this result is clearly supported by the residual plot in Figure 5.4(d). For around the first 7 days post-stroke, the coefficient line is not within the confidence bands of the spline fit, where the curve remains above the coefficient line until around a month post-stroke. The curve then crosses the coefficient line and remains below it for the remainder of follow-up, when the coefficient is close to the upper interval line. The curve is initially constant and begins decreasing around a week to 2 weeks post-stroke, indicating that a change in effect of 'no scan' could be around this time period.

5.6 Conclusion

This chapter has reviewed methods for assessing the proportional hazards assumption, and derived that the proportional hazards assumption can be assessed using a linear regression of the scaled residuals against time for each covariate, where this derivation relies on scaling the Schoenfeld residuals using a constant scaling factor, and using a centered function of time. Application of Rubin's rules to the regression estimates was shown to enable assessment of the proportional hazards assumption for a pooled Cox model fitted to multiply imputed data.

This chapter also presented a simulation study to show the Type 1 error of using linear regression and the Wald test, compared to the score test derived by Grambsch and Therneau (1994) which is used as standard in practice. The simulation study showed consistent Type 1 error between the two methods for the complete data, but highlighted possible issues around application to imputed data, with increased Type 1 errors for larger amounts of missing information, further increasing with number of events. We noted these issues are likely due to imputation procedure imperfections rather than the test itself however.

Application of our proposed test of proportional hazards to the pooled adjusted Cox model fitted in Chapter 4 has shown that the proportional hazards assumption was violated by several covariates within the stroke audit data, and thus any further analyses of this data needs to account for time-dependent covariate effects. This motivates the methodology developed in the next two chapters, prior to additional analyses in Chapter 8.

Chapter 6

Handling Non-Proportional Hazards in MICE

6.1 Introduction

Motivated by the violation of the proportional hazards assumption shown in the application to stroke data in Chapter 4 and Section 5.5, this chapter shows the theoretical extension for handling non-proportional hazards survival data within MICE, detailing the development of the appropriate form of the imputation models for multiple variable types.

Firstly, methods for handling non-proportional hazards will be reviewed, discussing how time-dependent covariate effects can be incorporated into the Cox proportional hazards model. We will highlight potential issues around handling time-dependent covariate effects in the presence of missing data and consider the current methods for handling missing data.

The Cox proportional hazards and piecewise-proportional hazards models will be defined. The derivation of the imputation model forms, approximately compatible with the standard Cox proportional hazards model, will be outlined in detail for each variable type, expanding upon the work by White and Royston (2009).

We will then present how this can be extended to derive the appropriate forms of the imputation models for the different variable types when the survival data has non-proportional hazards. The extension we present gives the form of imputation models approximately compatible with the piecewise-proportional hazards model.

6.2 Review of Current Methods

Previous chapters in this thesis have highlighted the importance of the proportional hazards assumption when fitting a Cox regression model, however, motivated by the model validation carried out in Section 5.5, here we consider alternatives to the standard Cox PH model in order to incorporate time-dependent covariate effects within survival analysis. Further, we explore approaches for handling missing data using multiple imputation, with a particular interest in their suitability for application to non-proportional hazards survival data.

Considering a time-constant explanatory variable, where the value of the variable itself does not change over time, such a variable can be defined to have a time-dependent covariate effect if its model coefficient is a function of time. By the definition of proportional hazards, a model coefficient which is a function of time violates this assumption. Collett (2015) states there are several alternative modelling approaches which do not require the assumption of proportional hazards, including accelerated failure time modelling and proportional odds models.

Interpretation of an accelerated failure time model is in terms of the speed of progression of a disease, where Tableman and Kim (2005) state that within an accelerated failure time model, the explanatory variables are assumed to act multiplicatively on the time-scale, affecting the rate of progression of an individual along the time-axis (Harrell, 2006). The Weibull, log-logistic and log-normal distributions are most commonly used for the survival times as the basis for a parametric accelerated failure time model, where an in depth discussion of these can be found in Collett (2015).

The proportional odds model, on the other hand, gives a non-parametric estimation of the baseline hazard function, and expresses the odds of an individual surviving beyond some time point t. Within this model the covariates act multiplicatively on these odds, resulting in a model which is a linear model for the log-odds ratio (Collett, 2015).

Hosmer et al. (2008) highlight, however, that the proportional hazards model is accepted in many applied settings as the standard method for analysis of survival, and thus focus will now be upon how time-dependent covariate effects can be incorporated into the Cox proportional hazards model.

Harrell (2006) state that the stratified proportional hazards model can be used to adjust for factors which are not modelled, such as those which do not satisfy the proportional hazards assumption, where Hosmer et al. (2008) reinforce its use for accommodating non-proportional hazards within a covariate. Within the stratified proportional hazards model, rather than incorporating covariates with non-proportional hazards as regressors, they can be incorporated as stratification factors, however Therneau and Grambsch (2013) highlight that stratified analyses are less efficient and this approach does not provide a test of the association between the stratification factor and survival.

An alternative approach, outlined by Therneau and Grambsch (2013), is to partition the time axis. This involves splitting the follow-up time into intervals, where the proportional hazards assumption may approximately hold within each interval. Models can then be fitted to each time interval separately, considering only those still at risk within the interval, and censoring those still at risk at the end of the interval.

A further method is to model time-varying covariate effects through creating time-dependent covariates. For a time-dependent coefficient $\beta(t)$, Therneau and

Grambsch (2013) suggest a time-dependent covariate $X^*(t)$ can be created so that

$$\beta(t)X = \beta X^*(t).$$

However, this approach can be computationally challenging and it can be difficult to specify $X^*(t)$ appropriately to produce valid inferences.

It is also possible to incorporate a time-dependent effect into the model directly, as a time-dependent coefficient, $\beta(t)$, which is some function of time. One of the simplest approaches to incorporate a time-dependent coefficient into a model is to define $\beta(t)$ as a step function. Zhou (2001) introduce this approach as a piecewise regression model, and Therneau et al. (2019) discuss how this can implemented within statistical software. This approach is another which involves partitioning the time axis into intervals, however rather than fitting separate models, a model is fitted with different coefficients for a covariate with non-proportional hazards, dependent upon the interval. We refer to this as the piecewise-proportional hazards model, where the coefficients for the covariate with non-proportional hazards are piecewise-constant. The coefficients for each time interval are therefore assumed to satisfy the proportional hazards assumption over the corresponding time interval. This is the approach we intend to use to handle the non-proportional hazards within the stroke audit data and will fully define the model in Section 6.3.2.

Now, we also need to consider possible approaches to handling non-proportional hazards within a multiple imputation framework, however there is currently minimal literature around how this can be incorporated into imputation models appropriately. Incorporating survival data into imputation models is a challenge in itself and requires special consideration for how the survival outcome is taken into account during imputation. Below we focus on a key set of papers which outline approaches for multiple imputation with survival data, discussing their findings and recommendations.

Multiple imputation using chained equations (MICE) has been identified as a flexible approach for multiple imputation, with the ability to handle multiple covariate types (White et al., 2011). Sterne et al. (2009) discuss the potentials and pitfalls of multiple imputation, particularly within clinical data. A key pitfall highlighted by Sterne et al. (2009) is the omission of the outcome variable from the imputation procedure as this can result in falsely weakening the association between covariates and the outcome.

White and Royston (2009) investigated approaches for incorporating the survival outcome within imputation models, and concluded a suitable approach is to include the censoring indicator and the Nelson-Aalen estimator of the cumulative hazard as terms within the imputation models. White et al. (2011) provide guidance around implementing MICE, discussing many considerations for the multiple imputations as we have previously outlined in Section 3.3.3. Further guidance on practical aspects of implementing MICE are provided by Azur et al. (2011), and van Buuren and Groothuis-Oudshoorn (2011) present an overview of the `mice` package in R.

Bartlett et al. (2015) expanded upon the work by White and Royston (2009) and White et al. (2011), discussing appropriate ways to incorporate non-linear or interaction terms into imputation models. The work by Bartlett et al. (2015) highlights that for an imputation model to be compatible with the substantive, or analysis, model, it must allow for any interaction or non-linear terms included within the analysis model. This is to avoid imposing any restrictions on the analyses of the imputed data sets, and suggests that time-dependent covariate effects should also be accounted for within imputation models to avoid restricting the form of the hazard functions for covariates with non-proportional hazards.

6.3 Imputation Model Theory

6.3.1 Background

The development of the MICE imputation models for application to non-PH survival data relies upon understanding the differences between a standard Cox PH model and a piecewise-proportional hazards model, and how this affects the derivation of the conditional distribution of the incomplete variable on the observed data.

Firstly we will define the Cox PH model and show the derivation of the conditional distribution of the incomplete variable on the observed data, as outlined by White and Royston (2009). Let t denote the survival time and δ be the censoring indicator. Suppose we have an incomplete variable X and complete variable Z. The Cox proportional hazards model is defined as

$$h(t|X, Z) = h_0(t) \exp(\beta_X X + \beta_Z Z).$$

In order to find the conditional distribution of the incomplete variable X on the observed data, we must first define the log-likelihood of the outcomes. Given the complete data, this is

$$\log p(t, \delta|X, Z) = \delta \log h(t|X, Z) - H(t|X, Z). \tag{6.1}$$

This can also be stated by substituting in the Cox PH model for the hazard function to give the log-likelihood

$$\log p(t, \delta|X, Z) = \delta(\log h_0(t) + \beta_X X + \beta_Z Z) - H_0(t) \exp(\beta_X X + \beta_Z Z).$$

By Bayes' theorem, we can determine the conditional distribution of the incomplete variable X on the observed data:

$$\log p(X|t, \delta, Z) = \log p(X|Z) + \log p(t, \delta|X, Z) - \log p(t, \delta|Z). \tag{6.2}$$

Substituting in the log-likelihood and considering a constant term C which depends only on δ, t, and Z, but not X, the conditional distribution of X on the observed data is equivalently

$$\log p(X|t, \delta, Z) = \log p(X|Z) + \delta(\beta_X X + \beta_Z Z) - H_0(t) \exp(\beta_X X + \beta_Z Z) + C. \tag{6.3}$$

Using appropriate exposure models and the conditional distribution of X on the observed data, the appropriate imputation models for different variable types can be formulated. Sections 6.3.3, 6.3.4 and 6.3.5 will cover this is depth for binary, categorical and continuous variables, respectively.

6.3.2 The Piecewise-Proportional Hazards Model

Now suppose the incomplete variable X has a time-dependent covariate effect, or non-proportional hazards. Suppose the hazard ratio associated with X is constant up until some survival time t_0, at which a change in the effect of X on survival occurs. Assume the new hazard ratio associated with X beyond survival time t_0 is again constant. We can define the hazard function using a piecewise-proportional hazards model as follows

$$h(t|X, Z) = h_0(t) \exp(\beta_{X1} X + \beta_{X2} \mathbb{I}(t > t_0) X + \beta_Z Z), \tag{6.4}$$

where

$$\mathbb{I}(t > t_0) = \begin{cases} 1, & \text{if } t > t_0 \\ 0, & \text{if } t \leq t_0 \end{cases}$$

is an indicator function for the effect of X dependent on survival time t, so that β_{X1} gives the effect of X before t_0, and the effect of X after t_0 is given by $(\beta_{X1} + \beta_{X2})$. This now results in the need for two censoring indicators, defined as

$$\delta_1 = \begin{cases} \delta, & \text{if } t \leq t_0 \\ 0, & \text{otherwise} \end{cases}, \qquad \delta_2 = \begin{cases} \delta, & \text{if } t > t_0 \\ 0, & \text{otherwise} \end{cases}.$$

Given the log-likelihood of the outcomes given the complete data, in Equation (6.1), we now derive the conditional distribution of X given the observed data for the piecewise-proportional hazards model. For notational ease, let $t_\mathrm{m} = \min(t_0, t)$, where $\min(t_0, t)$ is the earliest of the survival times t_0 and t. Considering the piecewise-proportional hazards model, the log-likelihood of the outcomes becomes

$$\begin{aligned} &\log p(t, \delta_1, \delta_2 | X, Z) \qquad (6.5) \\ &= \delta_1(\log h_0(t) + \beta_{X1}X + \beta_Z Z) + \delta_2(\log h_0(t) + (\beta_{X1} + \beta_{X2})X + \beta_Z Z) \\ &\quad - \left[H_0(t_\mathrm{m})e^{\beta_{X1}X + \beta_Z Z} + (H_0(t) - H_0(t_\mathrm{m}))e^{(\beta_{X1}+\beta_{X2})X + \beta_Z Z}\right], \end{aligned}$$

where $H_0(t_\mathrm{m}) = H_0(\min(t_0, t))$ is defined as the cumulative baseline hazard function at the minimum value of t and t_0.

Bayes' theorem was used by White and Royston (2009) to derive the conditional distribution of X given the observed data given in Equation (6.2). Now, by Bayes' theorem and using the log-likelihood derived from the piecewise-proportional hazards model, as defined in Equation (6.5), we derive that the conditional distribution of X given the observed data can be defined as

$$\begin{aligned} &\log p(X | t, \delta_1, \delta_2, Z) \qquad (6.6) \\ &= \log p(X|Z) + \delta_1(\beta_{X1}X + \beta_Z Z) + \delta_2(\beta_{X1}X + \beta_{X2}X + \beta_Z Z) \\ &\quad - \left[H_0(t_\mathrm{m})e^{\beta_{X1}X + \beta_Z Z} + (H_0(t) - H_0(t_\mathrm{m}))e^{\beta_{X1}X + \beta_{X2}X + \beta_Z Z}\right] + C, \end{aligned}$$

where the constant term C may depend on δ_1, δ_2, t and Z, but not X.

6.3.3 Imputation of Binary Variables

Here we will show how the imputation model can be derived to impute the missing observations of a binary variable using the conditional distribution of X given the observed data. Firstly, we will show how imputed values can be drawn from fitting an imputation model to the incomplete variable X on observed Z. We will then focus on the derivation of the imputation models to include the survival outcomes appropriately, giving both the derivations for imputation models approximately compatible to the standard Cox PH model and, as a novel contribution, to the piecewise-proportional hazards model. For these we will take three cases of the type of complete variable Z; categorical Z, no Z and the general case.

Drawing Imputations from the Imputation Model

Logistic regression is the model of choice for imputing binary X from observed $\mathbf{Z}$, where the logistic regression model is defined as

$$\operatorname{logit} p(X = 1|\mathbf{Z}; \beta) = \beta\mathbf{Z}.$$

Fitting this model to individuals with observed X gives estimated parameter $\hat{\beta}$ and estimated variance-covariance matrix $\mathbf{V}$. Approximate the posterior distribution of $\hat{\beta}$ by $\text{MVN}(\hat{\beta}, \mathbf{V})$, and let β^* be a draw from this posterior distribution. In order to impute each missing observation X_i, define u_i to be a random draw from a uniform distribution on $(0, 1)$, and let $p_i^* = [1 + \exp(-\beta^*\mathbf{Z}_i)]^{-1}$. An imputed value, X_i^* can be drawn as

$$X_i^* = \begin{cases} 1, & \text{if } u_i < p_i^* \\ 0, & \text{otherwise.} \end{cases}$$

This is a straightforward procedure to implement, however problems can arise due to perfect prediction (White et al., 2011). This occurs in logistic regression when a two-way table of the predictor and outcome variables contains a zero cell, or in other words, if there is a category within a predictor variable for which the outcome is always 0, or always 1. White et al. (2010) discuss the issue of perfect prediction in more depth.

Derivation for Standard Cox PH Model

Here we outline the derivation presented by White and Royston (2009) for the imputation model for binary incomplete variable X, expanding upon the theoretical rationale given. In order to derive the imputation model for binary incomplete variable X, take an exposure model for X on observed Z and extend this to account for the survival outcomes. Using exposure model, $\operatorname{logit} p(X = 1|Z) = \zeta_Z$, we have

$$\operatorname{logit} p(X = 1|t, \delta, Z) = \log p(X = 1|t, \delta, Z) - \log p(X = 0|t, \delta, Z).$$

Substituting in the conditional distribution of X given the observed data as defined in Equation (6.2), and simplifying gives

$$\begin{aligned}\operatorname{logit} p(X = 1|t, \delta, Z) &= \log p(X = 1|Z) + \delta(\beta_X + \beta_Z Z) - H_0(t)e^{\beta_X + \beta_Z Z} \\ &\quad - \left[\log p(X = 0|Z) + \delta(\beta_Z Z) - H_0(t)e^{\beta_Z Z}\right] \\ &= \log p(X = 1|Z) - \log p(X = 0|Z) + \delta\beta_X \\ &\quad - H_0(t)e^{\beta_X + \beta_Z Z} + H_0(t)e^{\beta_Z Z}.\end{aligned}$$

Referring back to the exposure model, and simplifying further, gives the imputation model for binary X as

$$\operatorname{logit} p(X = 1|t, \delta, Z) = \zeta_Z + \delta\beta_X - H_0(t)(e^{\beta_X} - 1)e^{\beta_Z Z}. \tag{6.7}$$

This is not a standard logistic regression model due to the $\exp(\beta_Z Z)$ term on the right hand side. For binary or categorical Z, Equation (6.7) gives a logistic regression on δ, Z, $H_0(t)$ and the interaction term between $H_0(t)$ and Z. This gives us the model previously defined in Section 3.3.3,

$$\operatorname{logit} p(X = 1|t, \delta, Z) = \alpha_0 + \alpha_1\delta + \alpha_2 H_0(t) + \alpha_{3Z} + \alpha_{4Z} H_0(t),$$

where the terms α_{3Z} and α_{4Z} each represent a set of dummy variables with their coefficients. Considering the case where there is no Z, the imputation model can be defined as

$$\operatorname{logit} p(X = 1|t, \delta) = \zeta + \delta\beta_X - H_0(t)(e^{\beta_X} - 1).$$

This gives a logistic regression of X on the censoring indicator, δ, and the baseline cumulative hazard, $H_0(t)$, giving the model specified in Section 3.3.3,

$$\operatorname{logit} p(X = 1|t, \delta) = \alpha_0 + \alpha_1\delta + \alpha_2 H_0(t).$$

For the most general case, where Z is continuous, there are no exact results. Here we assume the exposure model $\operatorname{logit} p(X = 1|Z) = \zeta_0 + \zeta_1 Z$, and make an approximation of the $\exp(\beta_Z Z)$ term. For small $\operatorname{Var}(\beta_Z Z)$, we can take the Taylor series approximation $\exp(\beta_Z Z) \approx \exp(\beta_Z \bar{Z})$. Substituting the new exposure model and Taylor series approximation into Equation (6.7) gives the imputation model of binary X on general Z as

$$\operatorname{logit} p(X = 1|t, \delta, Z) = \zeta_0 + \zeta_1 Z + \delta\beta_X - H_0(t)(e^{\beta_X} - 1)e^{\beta_Z \bar{Z}}.$$

This gives a logistic regression on δ, $H_0(t)$ and Z, and can be defined as

$$\operatorname{logit} p(X = 1|t, \delta, Z) = \alpha_0 + \alpha_1\delta + \alpha_2 H_0(t) + \alpha_3 Z.$$

A more accurate approximation can be made using the Taylor series, giving $\exp(\beta_Z Z) \approx \exp(\beta_Z \bar{Z})\{1 + \beta_Z(Z - \bar{Z})\}$. This results in the imputation model

$$\operatorname{logit} p(X = 1|t, \delta, Z) = \zeta_0 + \zeta_1 Z + \delta\beta_X - H_0(t)e^{\beta_Z \bar{Z}}(e^{\beta_X} - 1)\{1 + \beta_Z(Z - \bar{Z})\},$$

which is a logistic regression model on δ, Z, and $H_0(t)$, plus the interaction term $H_0(t) \times Z$.

Derivation for Piecewise-Proportional Hazards Model

The section above provides the derivation previously presented by White and Royston (2009) for the standard Cox PH setting, however, modification is needed to obtain an appropriateness imputation model form for the piecewise-proportional hazards setting. Taking a similar approach to the theoretical rationale for imputation models in the standard Cox PH model setting, here we present our derivation of the imputation model for binary X in the non-proportional hazards setting. This section aims to improve current practice and presents a novel contribution of this thesis.

We again take the most general exposure model $\operatorname{logit} p(X = 1|Z) = \zeta_Z$, but now consider the outcomes of survival time t and the two censoring indicators δ_1 and δ_2. This gives

$$\begin{aligned} &\operatorname{logit} p(X = 1|t, \delta_1, \delta_2, Z) = \\ &\qquad \log p(X = 1|t, \delta_1, \delta_2, Z) - \log p(X = 0|t, \delta_1, \delta_2, Z). \end{aligned} \tag{6.8}$$

Equation (6.6) states the conditional distribution of X given the observed data derived from the piecewise-proportional hazards model. Substituting this into

Equation (6.8) gives

$$
\begin{aligned}
&\operatorname{logit} p(X=1|t,\delta_1,\delta_2,Z) = \\
&\qquad \Big\{ \log p(X=1|Z) + \delta_1(\beta_{X1}+\beta_Z Z) + \delta_2(\beta_{X1}+\beta_{X2}+\beta_Z Z) \\
&\qquad\qquad - \left[H_0(t_{\mathrm{m}})e^{\beta_{X1}+\beta_Z Z} + (H_0(t)-H_0(t_{\mathrm{m}}))e^{\beta_{X1}+\beta_{X2}+\beta_Z Z}\right] \Big\} \\
&\qquad - \Big\{ \log p(X=0|Z) + \delta_1(\beta_Z Z) + \delta_2(\beta_Z Z) \\
&\qquad\qquad - \left[H_0(t_{\mathrm{m}})e^{\beta_Z Z} + (H_0(t)-H_0(t_{\mathrm{m}}))e^{\beta_Z Z}\right] \Big\},
\end{aligned}
$$

where again $H_0(t_{\mathrm{m}}) = H_0(\min(t_0,t))$ is the cumulative baseline hazard at the minimum of t_0 and t. Simplifying gives

$$
\begin{aligned}
&\operatorname{logit} p(X=1|t,\delta_1,\delta_2,Z) = \\
&\qquad \log p(X=1|Z) - \log p(X=0|Z) + \delta_1\beta_{X1} + \delta_2(\beta_{X1}+\beta_{X2}) \\
&\qquad - \left[H_0(t_{\mathrm{m}})e^{\beta_Z Z}(e^{\beta_{X1}}-1) + (H_0(t)-H_0(t_{\mathrm{m}}))e^{\beta_Z Z}(e^{\beta_{X1}+\beta_{X2}}-1)\right].
\end{aligned}
$$

Substituting in the exposure model and carrying out further simplification gives

$$
\begin{aligned}
\operatorname{logit} p(X=1|t,\delta_1,\delta_2,Z) &= \zeta_Z + \delta_1\beta_{X1} + \delta_2(\beta_{X1}+\beta_{X2}) \\
&- \left[H_0(t_{\mathrm{m}})(e^{\beta_{X1}}-1) + (H_0(t)-H_0(t_{\mathrm{m}}))(e^{\beta_{X1}+\beta_{X2}}-1)\right] e^{\beta_Z Z}
\end{aligned}
\tag{6.9}
$$

This is not a standard logistic regression due to the $\exp(\beta_Z Z)$ term again. For the simplest case where there is no Z we get

$$
\begin{aligned}
\operatorname{logit} p(X=1|t,\delta_1,\delta_2) &= \zeta + \delta_1\beta_{X1} + \delta_2(\beta_{X1}+\beta_{X2}) \\
&- \left[H_0(t_{\mathrm{m}})(e^{\beta_{X1}}-1) + (H_0(t)-H_0(t_{\mathrm{m}}))(e^{\beta_{X1}+\beta_{X2}}-1)\right],
\end{aligned}
$$

which gives a logistic regression on the censoring indicators, δ_1 and δ_2, and the two terms for the cumulative baseline hazard $H_0(t_{\mathrm{m}})$ and $H_0(t) - H_0(t_{\mathrm{m}})$;

$$\begin{aligned}&\operatorname{logit} p(X = 1|t, \delta_1, \delta_2) = \\ &\qquad \alpha_0 + \alpha_1\delta_1 + \alpha_2\delta_2 + \alpha_3 H_0(t_{\mathrm{m}}) + \alpha_4(H_0(t) - H_0(t_{\mathrm{m}})).\end{aligned}$$

For binary or categorical Z, Equation (6.9) gives a logistic regression on Z and the censoring indicators, δ_1 and δ_2, alongside the two terms for the cumulative baseline hazard $H_0(t_{\mathrm{m}})$ and $H_0(t) - H_0(t_{\mathrm{m}})$, and the interaction terms between these and Z. This can be represented as

$$\begin{aligned}\operatorname{logit} p(X = 1|t, \delta_1, \delta_2, Z) &= \alpha_0 + \alpha_1\delta_1 + \alpha_2\delta_2 + \alpha_3 H_0(t_{\mathrm{m}}) \\ &+ \alpha_4(H_0(t) - H_0(t_{\mathrm{m}})) + \alpha_{5Z} + \alpha_{6Z} H_0(t_{\mathrm{m}}) + \alpha_{7Z}(H_0(t) - H_0(t_{\mathrm{m}})),\end{aligned}$$

where terms such as α_{5Z} represent a set of dummy variables with their coefficients. Considering the general case for Z, we again have no exact results. Taking the same general case exposure model $\operatorname{logit} p(X = 1|Z) = \zeta_0 + \zeta_1 Z$, and using the Taylor series approximation $\exp(\beta_Z Z) \approx \exp(\beta_Z \bar{Z})$ for small $\operatorname{Var}(\beta_Z Z)$ again, we get the imputation model

$$\begin{aligned}\operatorname{logit} p(X = 1|t, \delta_1, \delta_2, Z) &= \zeta_0 + \zeta_1 Z + \delta_1\beta_{X1} + \delta_2(\beta_{X1} + \beta_{X2}) \\ &- \left[H_0(t_{\mathrm{m}})(e^{\beta_{X1}} - 1) + (H_0(t) - H_0(t_{\mathrm{m}}))(e^{\beta_{X1}+\beta_{X2}} - 1)\right] e^{\beta_Z \bar{Z}}.\end{aligned}$$

This gives a logistic regression model on Z, δ_1, δ_2, $H_0(t_{\mathrm{m}})$ and $H_0(t) - H_0(t_{\mathrm{m}})$, which can be written as

$$\begin{aligned}&\operatorname{logit} p(X = 1|t, \delta_1, \delta_2) = \\ &\qquad \alpha_0 + \alpha_1\delta_1 + \alpha_2\delta_2 + \alpha_3 H_0(t_{\mathrm{m}}) + \alpha_4(H_0(t) - H_0(t_{\mathrm{m}})) + \alpha_5 Z.\end{aligned}$$

Any transformation of Z needed for predicting X should also be included in the imputation model.

Previously we considered a more accurate approximation for $\exp(\beta_Z Z)$, which resulted in the imputation model containing an interaction term between the cumulative baseline hazard function and Z, however a simulation study conducted by White and Royston (2009) suggested that the addition of this interaction term does not increase the accuracy sufficiently to warrant the extra complication. Due to the added complication of accounting for non-proportional hazards here, we will not further complicate this by the addition of any such interaction terms.

6.3.4 Imputation of Categorical Variables

This section will show the derivation of the imputation model for a categorical variable X using the conditional distribution of X given the observed data. Initially we will outline multinomial logistic regression and describe how this can be used to draw an imputed value of X based on a fully observed variable Z. The approximate theoretical rationale for imputation models suitable for handling survival outcomes will then be given, both for approximate compatibility to the standard Cox PH model and to the piecewise-proportional hazards model. Here we will focus upon the general case for the variable Z.

Drawing Imputations from the Imputation Model

Categorical variables can be imputed using either multinomial logistic regression or a proportional odds model, however as proportional odds is only suitable for handling an ordered categorical variable, we will focus on multinomial logistic regression.

Let $l = 1, ..., L$ denote the levels of the categorical variable X and suppose the variable X has $L > 2$ levels modelled using multinomial logistic regression. Each of the levels, l, has a logistic regression equation comparing the level to some chosen baseline level, say level 1. This can be expressed as

$$p(X = l|\mathbf{Z}; \beta) = \left[\sum_{l'=1}^{L} \exp(\beta_{l'}\mathbf{Z})\right]^{-1} \exp(\beta_l \mathbf{Z}),$$

where $\beta_1 = 0$. Denote a random draw from a Normal approximation to the posterior distribution of $\beta = (\beta_2, ..., \beta_L)$ as β^*. For each missing observation X_i, let p^*_{il} denote the drawn level membership probabilities, where $p^*_{il} = p(X_i = l|\mathbf{Z}_i; \beta^*)$, for $l = 1, ..., L$, and let $P_{il} = \sum_{l'}^{l} p^*_{il'}$. Define u_i to be a random draw from a uniform distribution on $(0, 1)$. Each imputed value X^*_i is given as

$$X^*_i = 1 + \sum_{l=1}^{L-1} \mathbb{I}(u_i > P_{il}),$$

where the indicator function $\mathbb{I}(u_i > P_{il}) = 1$ if $u_i > P_{il}$, and 0 otherwise.

Derivation for Standard Cox PH Model

Here we expand upon the work by White and Royston (2009) to show the derivation of the imputation model for a categorical variable X. In order to derive the imputation model for a categorical variable X, with levels $l = 1, ..., L$, firstly the log-likelihood of the outcomes given the complete data needs to be reviewed. Consider the Cox proportional hazards model

$$h(t|X, Z) = h_0(t) \exp(\beta_{X_2} X_2 + \beta_{X_3} X_3 + ... + \beta_{X_L} X_L + \beta_Z Z),$$

where $\beta_{X_1} = 0$ for the baseline level of $X = 1$, and X_l is the indicator variable

$$X_l = \begin{cases} 1, & \text{if } X = l(l = 1, ...L) \\ 0, & \text{otherwise.} \end{cases} \tag{6.10}$$

Given the complete data, the log-likelihood for the outcomes for the case $X = l$ is given by

$$\log p(t, \delta | Z, X = l) = \delta \log h(t|Z, X = l) - H(t|Z, X = l). \tag{6.11}$$

This is equivalent to

$$\log p(t, \delta | Z, X = l) = \delta(\log h_0(t) + \beta_{X_l} + \beta_Z Z) - H_0(t) \exp(\beta_{X_l} + \beta_Z Z).$$

Given multinomial logistic regression consists of a logistic regression for each of the levels, assume we have a set of $L-1$ independent binary regressions for $l = 2, ..., L$, which can be defined as

$$\frac{p(X = l|t, \delta, Z)}{p(X = 1|t, \delta, Z)} = \frac{p(t, \delta|Z, X = l) \cdot p(X = l|Z)}{p(t, \delta|Z, X = 1) \cdot p(X = 1|Z)},$$

taking the form $\exp(\theta_l(t, \delta, Z))$, where $\theta_1 = 0$ since $X = 1$ is the baseline level. These can be expressed as the log-odds for each level against the baseline giving

$$\begin{aligned}&\log\left(\frac{p(t, \delta|Z, X = l) \cdot p(X = l|Z)}{p(t, \delta|Z, X = 1) \cdot p(X = 1|Z)}\right) = \\ &\quad \log p(t, \delta|Z, X = l) + \log p(X = l|Z) - \log p(t, \delta|Z, X = 1) - \log p(X = 1|Z).\end{aligned}$$

Considering the log-likelihood of the outcomes given the observed data, defined in Equation (6.11), the log-odds of each level can be expressed as

$$\begin{aligned}\log\left(\frac{p(t, \delta|Z, X = l) \cdot p(X = l|Z)}{p(t, \delta|Z, X = 1) \cdot p(X = 1|Z)}\right) &= \operatorname{logit} p(X = l|Z) \\ &+ \delta(\log h_0(t) + \beta_{X_l} + \beta_Z Z) - H_0(t) \exp(\beta_{X_l} + \beta_Z Z) \\ &- \left[\delta(\log h_0(t) + \beta_{X_1} + \beta_Z Z) - H_0(t) \exp(\beta_{X_1} + \beta_Z Z)\right],\end{aligned}$$

since $\operatorname{logit} p(X = l|Z) = \log p(X = l|Z) - \log p(X = 1|Z)$. Noting that $\beta_{X_1} = 0$, further simplification gives

$$\log\left(\frac{p(t,\delta|Z,X=l)\cdot p(X=l|Z)}{p(t,\delta|Z,X=1)\cdot p(X=1|Z)}\right) = \\ \operatorname{logit} p(X=l|Z) + \delta\beta_{X_l} - H_0(t)\exp(\beta_Z Z)(\exp(\beta_{X_l}) - 1).$$

We need to linearise this by making an approximation for $\exp(\beta_Z Z)$. For small $\operatorname{Var}(\beta_Z Z)$, the Taylor series can be used to approximate $\exp(\beta_Z Z)$. Firstly, consider the approximation $\exp(\beta_Z Z) \approx \exp(\beta_Z \bar{Z})$, and take the general exposure model $\operatorname{logit} p(X = l|Z) = \zeta_{l0} + \zeta_{l1} Z$. This gives the imputation model

$$\log\left(\frac{p(X=l|t,\delta,Z)}{p(X=1|t,\delta,Z)}\right) = \zeta_{l0} + \zeta_{l1}Z + \delta\beta_{X_l} - H_0(t)e^{\beta_Z Z}(e^{\beta_{X_l}} - 1),$$

which is a multinomial logistic regression on Z, the censoring indicator δ and the cumulative baseline hazard $H_0(t)$. This can be expressed as

$$\log\left(\frac{p(X=l|t,\delta,Z)}{p(X=1|t,\delta,Z)}\right) = \alpha_0 + \alpha_1 Z + \alpha_2\delta + \alpha_3 H_0(t).$$

Using the convention that $\log\left(\frac{p(X=1|t,\delta,Z)}{p(X=1|t,\delta,Z)}\right) = 0$, we can express this in terms of the probabilities as

$$p(X=l|t,\delta,Z) = \frac{\exp\{\zeta_{l0} + \zeta_{l1}Z + \delta\beta_{X_l} - H_0(t)e^{\beta_Z\bar{Z}}(e^{\beta_{X_l}} - 1)\}}{\sum_{l=1}^{L}\exp\{\zeta_{l0} + \zeta_{l1}Z + \delta\beta_{X_l} - H_0(t)e^{\beta_Z\bar{Z}}(e^{\beta_{X_l}} - 1)\}},$$

or equivalently,

$$p(X=l|t,\delta,Z) = \frac{\exp\{\alpha_0 + \alpha_1 Z + \alpha_2\delta + \alpha_3 H_0(t)\}}{\sum_{l=1}^{L}\exp\{\alpha_0 + \alpha_1 Z + \alpha_2\delta + \alpha_3 H_0(t)\}}.$$

Derivation for Piecewise-Proportional Hazards Model

Following our expansion on the work of White and Royston (2009) to demonstrate how the imputation model for a categorical variable can be derived for the standard Cox PH setting, this section provides a further novel contribution of this thesis by modifying this work for the piecewise-proportional hazards setting. Here we present our derivation of the imputation model for a categorical variable X with non-proportional hazards, where we approach the theoretical rationale for the imputation model for a categorical variable X with non-proportional hazards in a very similar format to the derivation for the standard Cox PH model. Firstly, let X be a categorical variable with a time-dependent covariate effect, where X has levels $l = 1, ..., L$. The piecewise-proportional hazards model fitted to X and some other variable Z is defined as

$$\begin{aligned} &h(t|X,Z) = \\ &\quad h_0(t)\exp(\beta_{X_2 1}X_2 + \mathbb{I}(t > t_0)\beta_{X_2 2}X_2 + ... + \beta_{X_L 1}X_L + \mathbb{I}(t > t_0)\beta_{X_L 2}X_L + \beta_Z Z), \end{aligned}$$

where $\beta_{X_1 1} = 0$ and $\beta_{X_1 2} = 0$ since $X = 1$ is the baseline level, and X_l is the indicator variable as defined in Equation (6.10). Then we have that $\beta_{X_l 1}$ gives the effect of $X = l$ before t_0 and $\beta_{X_l 1} + \beta_{X_l 2}$ gives the effect of $X = l$ after t_0, where t_0 is the time point at which the change in effect of X occurs.

The next step is to consider the log-likelihood of the outcomes given the observed data for the case of $X = l$. The outcomes to consider in the case of X having a time-dependent covariate effect are the two censoring indicators δ_1 and δ_2, and survival time t, giving the log-likelihood

$$\begin{aligned} &\log p(t, \delta_1, \delta_2 | Z, X = l) = \qquad (6.12) \\ &\quad \delta_1(\log h_0(t) + \beta_{X_l 1} + \beta_Z Z) + \delta_2(\log h_0(t) + \beta_{X_l 1} + \beta_{X_l 2} + \beta_Z Z) \\ &\quad - \left[H_0(t_{\mathrm{m}})\exp(\beta_{X_l 1} + \beta_Z Z) + (H_0(t) - H_0(t_{\mathrm{m}}))\exp(\beta_{X_l 1} + \beta_{X_l 2} + \beta_Z Z)\right], \end{aligned}$$

where $H_0(t_\mathrm{m}) = H_0(\min(t_0, t))$ is the cumulative baseline hazard at the minimum of t_0 and t. Assume again that we have $L - 1$ independent regressions, defined as

$$\frac{p(X = l|t, \delta_1, \delta_2, Z)}{p(X = 1|t, \delta_1, \delta_2, Z)} = \frac{p(t, \delta_1, \delta_2|Z, X = l) \cdot p(X = l|Z)}{p(t, \delta_1, \delta_2|Z, X = 1) \cdot p(X = 1|Z)},$$

for each level $l = 1, ..., L$, which takes the form $\exp(\theta_l(t, \delta_1, \delta_2, Z))$, where $\theta_1 = 0$ since $X = 1$ is the baseline level. Expressing these as log-odds for each level against the baseline gives

$$\log\left(\frac{p(t, \delta_1, \delta_2|Z, X = l) \cdot p(X = l|Z)}{p(t, \delta_1, \delta_2|Z, X = 1) \cdot p(X = 1|Z)}\right) = \log p(t, \delta_1, \delta_2|Z, X = l)$$
$$+ \log p(X = l|Z) - \log p(t, \delta_1, \delta_2|Z, X = 1) - \log p(X = 1|Z).$$

Since $\operatorname{logit} p(X = l|Z) = \log p(X = l|Z) - \log p(X = 1|Z)$ and $\beta_{X_1 1}$, $\beta_{X_1 2}$ are both zero, the log-odds can be expressed considering the log-likelihood given in Equation (6.12) as

$$\begin{aligned}
\log\left(\frac{p(t, \delta_1, \delta_2|Z, X = l) \cdot p(X = l|Z)}{p(t, \delta_1, \delta_2|Z, X = 1) \cdot p(X = 1|Z)}\right) &= \operatorname{logit} p(X = l|Z) \\
&+ \delta_1(\log h_0(t) + \beta_{X_l 1} + \beta_Z Z) + \delta_2(\log h_0(t) + \beta_{X_l 1} + \beta_{X_l 2} + \beta_Z Z) \\
&\quad - \big[H_0(t_\mathrm{m}) \exp(\beta_{X_l 1} + \beta_Z Z) + (H_0(t) - H_0(t_\mathrm{m})) \exp(\beta_{X_l 1} + \beta_{X_l 2} + \beta_Z Z)\big] \\
&+ \delta_1(\log h_0(t) + \beta_Z Z) + \delta_2(\log h_0(t) + \beta_Z Z) \\
&\quad - \big[H_0(t_\mathrm{m}) \exp(\beta_Z Z) + (H_0(t) - H_0(t_\mathrm{m})) \exp(\beta_Z Z)\big].
\end{aligned}$$

This expression can be simplified further to get

$$\begin{aligned}
&\log\left(\frac{p(t, \delta_1, \delta_2|Z, X = l) \cdot p(X = l|Z)}{p(t, \delta_1, \delta_2|Z, X = 1) \cdot p(X = 1|Z)}\right) = \\
&\quad \operatorname{logit} p(X = l|Z) + \delta_1 \beta_{X_l 1} + \delta_2(\beta_{X_l 1} + \beta_{X_l 2}) \\
&\quad - \big[H_0(t_\mathrm{m})(e^{\beta_{X_l 1}} - 1) + (H_0(t) - H_0(t_\mathrm{m}))(e^{\beta_{X_l 1} + \beta_{X_l 2}} - 1)\big]\, e^{\beta_Z Z}.
\end{aligned}$$

Again, we need to linearise this by making an approximation for $\exp(\beta_Z Z)$ using the Taylor series. For small $\mathrm{Var}(\beta_Z Z)$, a possible approximation is $\exp(\beta_Z Z) \approx \exp(\beta_Z \bar{Z})$. Using this approximation and taking the general exposure model, $\operatorname{logit} p(X = l|Z) = \zeta_{l0} + \zeta_{l1} Z$, gives the imputation model

$$\log\left(\frac{p(X = l|t, \delta_1, \delta_2, Z)}{p(X = 1|t, \delta_1, \delta_2, Z)}\right) = \zeta_{l0} + \zeta_{l1} Z + \delta_1 \beta_{X_l 1} + \delta_2(\beta_{X_l 1} + \beta_{X_l 2})$$
$$- \left[H_0(t_{\mathrm{m}})(e^{\beta_{X_l 1}} - 1) + (H_0(t) - H_0(t_{\mathrm{m}}))(e^{\beta_{X_l 1} + \beta_{X_l 2}} - 1)\right] e^{\beta_Z \bar{Z}}.$$

This gives a multinomial logistical regression on Z, δ_1, δ_2, $H_0(t_{\mathrm{m}})$ and $H_0(t) - H_0(t_{\mathrm{m}})$, which can be expressed as

$$\log\left(\frac{p(X = l|t, \delta_1, \delta_2, Z)}{p(X = 1|t, \delta_1, \delta_2, Z)}\right) =$$
$$\alpha_0 + \alpha_1 Z + \alpha_2 \delta_1 + \alpha_3 \delta_2 + \alpha_4 H_0(t_{\mathrm{m}}) + \alpha_5(H_0(t) - H_0(t_{\mathrm{m}})).$$

Using the convention that $\log\left(\frac{p(X=1|t,\delta_1,\delta_2,Z)}{p(X=1|t,\delta_1,\delta_2,Z)}\right) = 0$, we can express the imputation model in terms of the probabilities, giving the equation

$$\log\left(\frac{p(X = l|t, \delta_1, \delta_2, Z)}{p(X = 1|t, \delta_1, \delta_2, Z)}\right) =$$
$$\frac{\exp\left\{\alpha_0 + \alpha_1 Z + \alpha_2 \delta_1 + \alpha_3 \delta_2 + \alpha_4 H_0(t_{\mathrm{m}}) + \alpha_5(H_0(t) - H_0(t_{\mathrm{m}}))\right\}}{\sum_{l=1}^{L} \exp\left\{\alpha_0 + \alpha_1 Z + \alpha_2 \delta_1 + \alpha_3 \delta_2 + \alpha_4 H_0(t_{\mathrm{m}}) + \alpha_5(H_0(t) - H_0(t_{\mathrm{m}}))\right\}}.$$

6.3.5 Imputation of Continuous Variables

The final variable type to consider is continuous; this section will consider the imputation model for incomplete continuous variable X. Assuming X is a Normally distributed continuous variable, we will firstly show how imputed values for each missing observation can be obtained from a linear regression model. The derivation of the imputation model will then be shown to accommodate survival data outcomes, firstly focusing on the standard Cox PH model, and then extending to the non-proportional hazards setting where we derive the imputation model approximately compatible to the piecewise-proportional hazards model. The derivations will focus on the general case of observed variable Z.

Drawing Imputations from the Imputation Model

The most common choice of model for imputing a Normally distributed continuous variable is a linear regression model, defined as

$$X|\mathbf{Z};\beta \sim \mathrm{N}(\beta\mathbf{Z},\sigma^2).$$

Fitting this model to individuals with observed X gives estimated parameter, denoted $\hat{\beta}$ and the estimated variance-covariance matrix of $\hat{\beta}$, denoted $\mathbf{V}$. The estimated root mean-squared error is denoted as $\hat{\sigma}$. The imputation parameters β^* and σ^* can be drawn from the posterior distribution of β and σ. Let the number of individuals with observed X be denoted as n_{obs}, and let k be the number of parameters estimated. We firstly draw the imputation parameter σ^* as

$$\sigma^* = \hat{\sigma}\sqrt{(n_{obs}-k)/\phi},$$

where ϕ is a random draw from a χ^2-distribution on $n_{obs}-k$ degrees of freedom. Next, we draw imputation parameter β^* as

$$\beta^* = \hat{\beta} + \frac{\sigma^*}{\hat{\sigma}}\boldsymbol{u}_1\mathbf{V}^{1/2},$$

where $\mathbf{V}^{1/2}$ is the Cholesky decomposition of $\mathbf{V}$, and $\boldsymbol{u}_1$ is a row vector of k independent random draws from a standard Normal distribution. For each missing value of X, X_i, let X_i^* be the imputed value, which can be obtained as

$$X_i^* = \beta^*\mathbf{Z}_i + u_{2i}\sigma^*,$$

where u_{2i} is a random draw from a standard Normal distribution. If X is non-Normal, it may be possible to find a transformation of X, say $\psi(X)$, where $\psi(X)$ is Normally distributed. As the imputed values of $\psi(X)$ must be back-transformed to the original scale to get the imputed values of X, $\psi(X)$ must be monotonic and invertible.

Derivation for Standard Cox PH Model

Here we present the derivation for the imputation model of a continuous variable X, as outlined by White and Royston (2009). Theoretical rationale for the imputation model of a continuous variable X in the survival data setting requires the assumption of the exposure model $X|Z \sim \mathrm{N}(\zeta_0 + \zeta_1 Z, \sigma^2)$, for general Z. Using this exposure model, the conditional distribution of X given the observed data, in Equation (6.3), becomes

$$\log p(X|t, \delta, Z) = -\frac{(X - \zeta_0 - \zeta_1 Z)^2}{2\sigma^2} + \delta\beta_X X - H_0(t)\exp(\beta_X X + \beta_Z Z) + C, \quad (6.13)$$

where the constant term C may depend on t, δ or Z, but not on X. Due to the $\exp(\beta_X X)$ term, this is not a Normal distribution, or any other common distribution, therefore we need to make an approximation. For small $\mathrm{Var}(\beta_X X + \beta_Z Z)$, the Taylor series expansion gives the linear approximation $\exp(\beta_X X + \beta_Z Z) \approx \exp(\beta_X \bar{X} + \beta_Z \bar{Z})[1 + \beta_X(X - \bar{X}) + \beta_Z(Z - \bar{Z})]$. Substituting this approximation into Equation (6.13), and simplifying gives

$$\log p(X|t, \delta, Z) \approx -\frac{(X - \zeta_0 - \zeta_1 Z)^2}{2\sigma^2} + \beta_X X \left[\delta - H_0(t)\exp(\beta_X \bar{X} + \beta_Z \bar{Z})\right] + C,$$

where again the constant term C may depend on t, δ or Z, but not on X. Therefore, approximately we have

$$X|t, \delta, Z \sim \mathrm{N}\left(\zeta_0 + \zeta_1 Z + \beta_X \sigma^2 \left[\delta - H_0(t)\exp(\beta_X \bar{X} + \beta_Z \bar{Z})\right], \sigma^2\right),$$

This is a linear regression model on Z, δ and $H_0(t)$, and can be expressed as

$$X|t, \delta, Z \sim \mathrm{N}\left(\alpha_0 + \alpha_1\delta + \alpha_2 H_0(t) + \alpha_3 Z, \sigma^2\right),$$

for unknown parameters α. White and Royston (2009) further showed that using a quadratic approximation for $\exp(\beta_X X + \beta_Z Z)$ could be used, however, this would give a model which is not linear in the parameters. Using the quadratic approximation, $X|t, \delta, Z$ would be approximately Normally distributed with mean

$$\frac{\zeta_0 + \zeta_1 Z + \beta_X\sigma^2(\delta - H_0(t)e^{\beta_X\bar{X}+\beta_Z\bar{Z}}(1-\beta_X\bar{X})) - \beta_X\beta_Z\sigma^2 H_0(t)e^{\beta_X\bar{X}+\beta_Z\bar{Z}}(Z-\bar{Z})}{1+\beta_X^2\sigma^2 H_0(t)e^{\beta_X\bar{X}+\beta_Z\bar{Z}}}$$

and variance

$$\frac{\sigma^2}{1+\beta_X^2\sigma^2 H_0(t)e^{\beta_X\bar{X}+\beta_Z\bar{Z}}}.$$

In order to get a linear regression on δ, $H_0(t)$, Z and the interaction term $H_0(t)\times Z$, the quadratic terms β_X^2 would need to be ignored. This means that this approximation is only valid for small $\beta_X^2\sigma^2 H_0(t)$, or more specifically, when $\mathrm{Var}(\beta_X X)$ and/or $H_0(t)$ are small. If $H_0(t)$ is large, the approximation gives a slope which is too large.

Derivation for Piecewise-Proportional Hazards Model

The section above provides the derivation previously presented by White and Royston (2009) for the standard Cox PH setting, however, modification is needed to obtain an appropriateness imputation model form for the piecewise-proportional hazards setting. Following a similar process to White and Royston (2009), here we present our derivation of the imputation model for a continuous variable X which has a time-dependent covariate effect. This section aims to improve current practice and presents a novel contribution of this thesis.

Assuming the analysis model is a piecewise-proportional hazards model, as given in Equation (6.4), and considering a general Z, we can take the exposure model $X|Z \sim \mathrm{N}(\zeta_0 + \zeta_1 Z, \sigma^2)$. This results in the conditional distribution of X given the observed data being expressed as

$$\begin{aligned}
&\log p(X|t, \delta_1, \delta_2, Z) = \qquad (6.14)\\
&\quad -\frac{X - \zeta_0 - \zeta_1 Z)^2}{2\sigma^2} + \delta_1(\beta_{X1}X + \beta_Z Z) + \delta_2(\beta_{X1}X + \beta_{X2}X + \beta_Z Z)\\
&\quad - \left[H_0(t_\mathrm{m})e^{\beta_{X1}X+\beta_Z Z} + (H_0(t) - H_0(t_\mathrm{m}))e^{\beta_{X1}X+\beta_{X2}X+\beta_Z Z}\right] + C,
\end{aligned}$$

where the constant C may depend on δ_1, δ_2, t, and Z, but not X. As for the standard Cox PH model, this does not follow a Normal distribution, or any other common distribution, so an approximation is needed. In this case we need approximations for two terms; $\exp(\beta_{X1}X + \beta_Z Z)$ and $\exp(\beta_{X1}X + \beta_{X2}X + \beta_Z Z)$. Using the Taylor series, for small $\mathrm{Var}(\beta_{X1}X + \beta_Z Z)$ and small $\mathrm{Var}(\beta_{X1}X + \beta_{X2}X + \beta_Z Z)$, respectively, we can get the linear approximations

$$e^{\beta_{X1}X+\beta_Z Z} \approx e^{\beta_{X1}\bar{X}+\beta_Z\bar{Z}}\left[1 + \beta_{X1}(X - \bar{X}) + \beta_Z(Z - \bar{Z})\right],$$

and

$$\begin{aligned}
&e^{\beta_{X1}X+\beta_{X2}X+\beta_Z Z} \approx\\
&\qquad e^{\beta_{X1}\bar{X}+\beta_{X2}\bar{X}+\beta_Z\bar{Z}}\left[1 + \beta_{X1}(X - \bar{X}) + \beta_{X2}(X - \bar{X}) + \beta_Z(Z - \bar{Z})\right].
\end{aligned}$$

Substituting these approximations into the conditional distribution of X given the observed data, shown in Equation (6.14), gives

$$\begin{aligned}
&\log p(X|t,\delta_1,\delta_2,Z) \approx \\
&\quad -\frac{X-\zeta_0-\zeta_1 Z)^2}{2\sigma^2} + \delta_1(\beta_{X1}X+\beta_Z Z) + \delta_2(\beta_{X1}X+\beta_{X2}X+\beta_Z Z) \\
&\quad -\Big\{H_0(t_\mathrm{m})e^{\beta_{X1}\bar{X}+\beta_Z\bar{Z}}\left[1+\beta_{X1}(X-\bar{X})+\beta_Z(Z-\bar{Z})\right] \\
&\qquad -(H_0(t)-H_0(t_\mathrm{m}))e^{\beta_{X1}\bar{X}+\beta_{X2}\bar{X}+\beta_Z\bar{Z}}\big[1+\beta_{X1}(X-\bar{X}) \\
&\qquad\quad +\beta_{X2}(X-\bar{X})+\beta_Z(Z-\bar{Z})\big]\Big\}+C.
\end{aligned}$$

Many of these terms do not depend on X, therefore, we can rearrange and simplify to get the approximate conditional distribution of X given the observed data as

$$\begin{aligned}
&\log p(X|t,\delta_1,\delta_2,Z) \approx \\
&\quad -\frac{X-\zeta_0-\zeta_1 Z)^2}{2\sigma^2} + \beta_{X1}X\big(\delta_1 - H_0(t_\mathrm{m})e^{\beta_{X1}\bar{X}+\beta_Z\bar{Z}}\big) \\
&\quad +(\beta_{X1}+\beta_{X2})X\big(\delta_2-(H_0(t)-H_0(t_\mathrm{m}))e^{\beta_{X1}\bar{X}+\beta_{X2}\bar{X}+\beta_Z\bar{Z}}\big)+C,
\end{aligned}$$

where the constant term C may depend on Z, δ_1, δ_2 and t, but not X. Therefore approximately we have

$$X|t,\delta_1,\delta_2,Z \sim \mathrm{N}(\mu,\sigma^2),$$

where

$$\begin{aligned}
\mu = \zeta_0+\zeta_1 Z+\sigma^2\Big\{&\beta_{X1}\delta_1+(\beta_{X1}+\beta_{X2})\delta_2-\beta_{X1}H_0(t_\mathrm{m})e^{\beta_{X1}\bar{X}+\beta_Z\bar{Z}} \\
&-(\beta_{X1}+\beta_{X2})(H_0(t)-H_0(t_\mathrm{m}))e^{\beta_{X1}\bar{X}+\beta_{X2}\bar{X}+\beta_Z\bar{Z}}\Big\}.
\end{aligned}$$

This shows that the imputation model for a continuous variable X with a time-dependent covariate effect is a regression model on the two censoring indicators, δ_1 and δ_2, the cumulative baseline hazard terms, $H_0(t_\mathrm{m})$ and $H_0(t)-H_0(t_\mathrm{m})$, and observed variables Z. This can be expressed in terms of unknown parameters α as

$$X|t,\delta_1,\delta_2,Z \sim \mathrm{N}(\alpha_0+\alpha_1\delta_1+\alpha_2\delta_2+\alpha_3 H_0(t_\mathrm{m})+\alpha_4(H_0(t)-H_0(t_\mathrm{m}))+\alpha_5 Z,\sigma^2).$$

6.3.6 Further Considerations

Each of the imputation models defined above depend upon the baseline cumulative hazard function which is unknown. A simulation study conducted by White and Royston (2009) found that using the Nelson-Aalen estimator of the cumulative hazard function within the imputation models produced the lowest bias and highest power. Therefore it is recommended that using the Nelson-Aalen estimator is the best approach to incorporate the cumulative hazard function into the imputation models for each variable type.

A further issue to consider is how the observed variables Z should be incorporated into the imputation models, and whether a linear term is appropriate. Chapter 4 demonstrated a need to incorporate quadratic terms into the imputation models. Bartlett et al. (2015) considered methodology for this, as discussed in Section 6.2. Additionally, a covariate with a quadratic effect on survival may have missing values, and thus the linear and quadratic values of this covariate would need to be imputed appropriately.

A particular approach to imputing both linear and quadratic terms of the same covariate is passive imputation. This approach involves imputing the linear term and calculating the quadratic values as the square of the corresponding imputations of the linear term. However, Bartlett et al. (2015) showed that this approach can introduce biased estimates, which is reiterated by Von Hippel (2009), as using only the linear term for imputation of the quadratic dilutes the effect of the quadratic term on the outcome, thus causing underestimation of the coefficient.

An alternative approach is outlined by Von Hippel (2009) as ‘transform, then impute', which fits the quadratic term as an additional variable in the imputation models, and was further shown by Bartlett et al. (2015) to give unbiased results. The downfall of this approach is inconsistency in the imputations between the linear and quadratic terms, however, avoidance of bias in the coefficients is deemed to be of more importance by Von Hippel (2009), and therefore this would be the recommended approach.

6.4 Simulation Study

In order to assess if the specification of the imputation models proposed in Section 6.3, to account for time-dependent covariate effects, are an improvement over use of the naïve approach to ignore the non-proportional hazards, we conducted a simulation study to assess the bias of each imputation approach.

We produced 2000 simulations of survival data for $n = 500$ individuals, where in each simulation, three covariates were generated. The first covariate, $X1$ was continuous and generated from a normal distribution with $\mu = 65$ and $\sigma = 10$. The second covariate, $X2$, was a binary covariate generated using the Bernoulli distribution with probabilities dependent upon the values of $X1$. Finally, the third covariate, $X3$, was simulated to be a continuous covariate dependent upon both $X1$ and $X2$.

Survival times were generated from $X1$, $X2$ and $X3$ using the `simsurv` package in R, under the exponential distribution, where a time-dependent covariate effect was introduced for $X3$. This effect was piecewise-constant, where a changepoint for the effect was specified at $t = 200$, giving two piecewise-constant coefficients; $\beta_{X3(\leq)}$ for $t \leq 200$ and $\beta_{X3(>)}$ for $t > 200$, where $\beta_{X3(\leq)} = 0.3$ and $\beta_{X3(>)} = -0.3$. The effects of $X1$ and $X2$ on survival were assumed to satisfy the proportional hazards assumption, with coefficients $\beta_{X1} = 0.08$ and $\beta_{X2} = 1.1$. Four scenarios were simulated for survival, varying the proportion of events. The four scenarios specified the overall proportion of individuals with events to be: 20%, 40%, 60% and 80%, where for each scenario, the split of the amount of events within each time interval, $t \leq 200$ and $t > 200$, was simulated to be approximately equal.

Piecewise-proportional hazards models adjusted for all three covariates $X1$, $X2$ and $X3$, with a time-dependent effect for $X3$, were fitted to the complete data within each of the event proportion scenarios in each simulation. The bias of coefficients was computed for each covariate in each scenario, where the bias was calculated as the average over the simulations of the absolute difference between the model coefficients and the corresponding coefficient used to simulate the survival times.

Missing data was introduced into covariates $X2$ and $X3$ in order to examine the performance of the the imputation approaches. The proportion of missing values within each covariate was varied to produce three scenarios of missingness: 10% missing in $X2$ and 50% missing in $X3$, 20% missing in $X2$ and 40% missing in $X3$, and 30% missing in both $X2$ and $X3$. Covariate $X1$ remained complete in each scenario. This resulted in 12 missing data scenarios overall with the varying proportions of events, plus the 4 complete data scenarios.

For each of the twelve missing data scenarios, two imputation approaches were applied. The first approach ignored the time-dependent effect of $X3$ when incorporating the survival outcome into the imputation models, and fitted the imputation models with the outcome included through the Nelson-Aalen estimate of the cumulative hazard and the censoring indicator, as outlined by White and Royston (2009); this approach will be referred to as the 'naïve imputation' approach. The second imputation accounts for the time-dependent covariate effects using the imputation model forms proposed in Section 6.3, where the imputation models incorporated the survival outcome using the two censoring indicators δ_1 and δ_2, and Nelson-Aalen estimates for the two cumulative baseline hazard terms, $H_0(t_\mathrm{m})$ and $H_0(t) - H_0(t_\mathrm{m})$. This approach will be referred to as the 'TD imputation' approach.

Both imputation approaches were applied to each of the missing data scenarios within each event proportion scenario, where the imputations in each case were carried out using the `mice` package in R over 100 iterations for 10 cycles, to produce 10 imputed data sets within each approach for each scenario. A piecewise-proportional hazards model was fitted to each of the 10 imputed data sets from each imputation approach, where as in the complete case, the models were adjusted for all three covariates, $X1$, $X2$, and $X3$, with a time dependent effect on $X3$. For each approach, the 10 sets of model coefficients were combined using Rubin's rules to achieve a pooled piecewise-proportional hazards model for each approach in each scenario. The bias of the coefficients was again calculated as the average over the simulations of the absolute difference between the pooled model coefficients and the corresponding coefficients used to simulate the survival times.

This resulted is estimates of bias of model coefficients for each of the event proportion scenarios for both the complete case, and for the two imputation approaches for each missing data scenario. Smaller differences between the model coefficients and the simulation coefficients indicate lower bias.

6.4.1 Bias Results

The results of the simulation study are presented in Tables 6.1 and 6.2, which show the bias and percentage bias, respectively, in the coefficients for each of the covariates. The notation $X3_{(\leq)}$ and $X3_{(>)}$ represents the coefficients for $X3$ for time intervals $t \leq 200$ and $t > 200$, respectively.

The results in Tables 6.1 and 6.2 show that compared to the complete case, both imputation methods introduce some extra bias into the model coefficients of the piecewise-proportional hazards model.

In terms of the imputation approaches, Table 6.1 shows that the bias is lower for the coefficients of the model fitted following the TD imputation approach,

Table 6.1: Simulation results giving bias of coefficient of piecewise-proportional hazards model on complete and multiply imputed data, using naïve and TD imputation approaches (n=500, 2000 Simulations).

Events	**Missing (%)**		**Imp.**	**Bias**			
(%)	$X2$	$X3$	**Type**	$X1$	$X2$	$X3_{(\leq)}$	$X3_{(>)}$
20	0	0	Complete	0.018	0.288	0.029	0.024
	10	50	Naïve Imp	0.095	1.072	0.038	0.178
			TD Imp.	0.046	0.714	0.036	0.121
	20	40	Naïve Imp	0.088	1.159	0.040	0.171
			TD Imp.	0.042	0.784	0.034	0.116
	30	30	Naïve Imp	0.080	1.191	0.040	0.163
			TD Imp.	0.039	0.821	0.033	0.110
40	0	0	Complete	0.012	0.209	0.020	0.016
	10	50	Naïve Imp	0.099	1.090	0.029	0.194
			TD Imp.	0.040	0.647	0.027	0.094
	20	40	Naïve Imp	0.095	1.147	0.030	0.189
			TD Imp.	0.039	0.711	0.025	0.092
	30	30	Naïve Imp	0.087	1.162	0.031	0.180
			TD Imp.	0.035	0.720	0.024	0.088
60	0	0	Complete	0.010	0.165	0.016	0.015
	10	50	Naïve Imp	0.090	0.994	0.030	0.208
			TD Imp.	0.040	0.612	0.023	0.073
	20	40	Naïve Imp	0.086	1.037	0.026	0.201
			TD Imp.	0.038	0.677	0.021	0.070
	30	30	Naïve Imp	0.078	1.046	0.024	0.189
			TD Imp.	0.034	0.717	0.019	0.065
80	0	0	Complete	0.009	0.150	0.015	0.013
	10	50	Naïve Imp	0.080	0.894	0.058	0.230
			TD Imp.	0.040	0.596	0.025	0.081
	20	40	Naïve Imp	0.075	0.930	0.045	0.218
			TD Imp.	0.037	0.655	0.021	0.074
	30	30	Naïve Imp	0.067	0.938	0.034	0.203
			TD Imp.	0.033	0.693	0.018	0.066

compared to the naïve imputation approach. This is the case for all scenarios, with varying event and missingness proportions. Looking at the percentage bias in Table 6.2, it can be seen that the TD imputation approach reduced the bias by around 70% in some cases compared to the naïve imputation approach. The addition of separate terms for the cumulative hazard and censoring indicators for each time interval into the imputations models, to account for the time-dependent covariate effect, reduces the bias in the analysis stage compared to ignoring the non-proportional hazards.

These simulation results therefore suggest that the proposed method for incorporating time-dependent covariate effects, with approximate compatibility to a piecewise-proportional hazards analysis model, is an improvement on the naïve approach to ignore the non-proportional hazards in the imputation stage.

6.5 Conclusion

This chapter has reviewed methods for accounting for non-proportional hazards in a survival model, and discussed the issues this causes for multiple imputation. Following this, we showed the derivation of the imputation model form for each covariate type, as outlined by White and Royston (2009).

Using a similar process to White and Royston (2009), we defined the piecewise-proportional hazards model, and derived the form of the imputation models for binary, categorical and continuous covariates, to ensure approximate compatibility with the piecewise-proportional hazards model.

Finally, we presented a simulation study to compare the bias in model coefficients between the two imputation approaches; MICE ignoring non-proportional hazards and our derived imputation model forms incorporating a time-dependent hazard. This simulation study showed that the multiple imputation procedure causes bias in both cases, but the bias was reduced when time-dependent effects were accounted for within the imputation procedure. Therefore we can conclude our proposed imputation model forms produce more favourable results.

Table 6.2: Simulation results giving percentage bias of coefficient of piecewise-proportional hazards model on complete and multiply imputed data, using naïve and TD imputation approaches (n=500, 2000 Simulations).

Events	**Missing (%)**		**Imp.**	**Bias (%)**			
(%)	$X2$	$X3$	**Type**	$X1$	$X2$	$X3_{(\leq)}$	$X3_{(>)}$
20	0	0	Complete	22.5	26.2	9.7	8.0
	10	50	Naïve Imp	118.8	97.5	12.7	59.3
			TD Imp.	57.5	64.9	12.0	40.3
	20	40	Naïve Imp	110.0	105.4	13.3	57.0
			TD Imp.	52.5	71.3	11.3	38.7
	30	30	Naïve Imp	100.0	108.3	13.3	54.3
			TD Imp.	48.8	74.6	11.0	36.7
40	0	0	Complete	15.0	19.0	6.7	5.3
	10	50	Naïve Imp	123.8	99.1	9.7	64.7
			TD Imp.	50.0	58.8	9.0	31.3
	20	40	Naïve Imp	118.8	104.3	10.0	63.0
			TD Imp.	48.8	64.6	8.3	30.7
	30	30	Naïve Imp	108.8	105.6	10.3	60.0
			TD Imp.	43.8	65.5	8.0	29.3
60	0	0	Complete	12.5	15.0	5.3	5.0
	10	50	Naïve Imp	112.5	90.4	10.0	69.3
			TD Imp.	50.0	55.6	7.7	24.3
	20	40	Naïve Imp	107.5	94.3	8.7	67.0
			TD Imp.	47.5	61.5	7.0	23.3
	30	30	Naïve Imp	97.5	95.1	8.0	63.0
			TD Imp.	42.5	65.2	6.3	21.7
80	0	0	Complete	11.3	13.6	5.0	4.3
	10	50	Naïve Imp	100.0	81.3	19.3	76.7
			TD Imp.	50.0	54.2	8.3	27.0
	20	40	Naïve Imp	93.8	84.5	15.0	72.7
			TD Imp.	46.3	59.5	7.0	24.7
	30	30	Naïve Imp	83.8	85.3	11.3	67.7
			TD Imp.	41.3	63.0	6.0	22.0

Chapter 7

Piecewise-Proportional Hazards Model Validation

7.1 Introduction

This chapter provides methodology for assessing the proportional hazards assumption within the piecewise-proportional hazards model, detailing an alternative method for scaling the Schoenfeld residuals.

Initially, the form of the piecewise-proportional hazards model will be defined, outlining what the proportional hazards assumption means for this model and how it can be assessed. A review of the methods for assessing the proportional hazards assumption will be given, identifying their unsuitability for application to the piecewise-proportional hazards model.

A new method for scaling the Schoenfeld residuals is proposed, with discussion of how this can be used in assessing the proportional hazard assumption. An adapted visualisation technique is also provided to coincide with the formal test. Finally, further extensions of these methods are outlined for application within the multiple imputation setting.

7.2 Piecewise-Proportional Hazards Model

Prior to consideration of how we can assess the proportional hazards assumption within the piecewise-proportional hazards model, we must first define the piecewise-proportional hazards model and how it's form relates the proportional hazards assumption.

From the findings in Chapter 5, we would expect a single change point in covariate effects to be found for the stroke data, and thus, as the simplest case, we will focus on that scenario here. The methods we outline below will be able to be extended to additional change points and time intervals as needed however. Considering the scenario where there is a single change point in covariate effects, there will be two time intervals to consider.

To define the piecewise-proportional hazards model, suppose we have a covariate X with a time-dependent covariate effect and a further q explanatory variables Z_k, $k = 1, ..., q$, assumed to have time-constant effects. Fitting a piecewise-proportional hazards models assumes the effect of X is constant up until some survival time t_0, at which a change in the effect of X occurs, with the new effect of X assumed to again be constant beyond survival time t_0. Using a piecewise-proportional hazards model results in the hazard function being defined as

$$h(t|X) = h_0(t) \exp\left\{\beta_{X1}X + \beta_{X2}\mathbb{I}(t > t_0)X + \beta_1 Z_1 + ... + \beta_q Z_q\right\}, \tag{7.1}$$

where, as defined in Section 6.3.2,

$$\mathbb{I}(t > t_0) = \begin{cases} 1, & \text{if } t > t_0 \\ 0, & \text{if } t \leq t_0 \end{cases}$$

is an indicator function for the effect of X dependent on survival time t, so that β_{X1} gives the effect of X up to t_0, and the effect of X after t_0 is given by $(\beta_{X1} + \beta_{X2})$.

Given the piecewise-constant effect of X, we can consider the model given in Equation (7.1) as two separate models over two time intervals, defined as

$$h(t|X) = \begin{cases} h_0(t)\exp\left\{\beta_{X1}X + \beta_1 Z_1 + ... + \beta_q Z_q\right\}, & \text{if } t \leq t_0 \\ h_0(t)\exp\left\{(\beta_{X1} + \beta_{X2})X + \beta_1 Z_1 + ... + \beta_q Z_q\right\}, & \text{if } t > t_0. \end{cases}$$

The form of this model suggests the proportional hazards assumption should be examined separately for the two time periods for the variable X with a time-dependent covariate effect. This is required to assess whether the covariate effects of X satisfy the proportional hazards assumption within each of the time periods. For the remaining covariates, Z_k, the proportional hazards assumption should hold over the total follow-up time, and therefore should still be assessed as such.

Now, as outlined in Chapter 5, the proportional hazards assumption can be validated by assessing the linear dependence of the model coefficients on time. Equation (5.1) gave this linear dependence expressed as a regression on some function of time $G(t)$. Considering the form of the piecewise-proportional hazards model, we have three cases requiring assessment; the dependence of β_{X1} on time, the dependence of $\beta_{X1}+\beta_{X2}$ on time, and the dependence of the remaining β_k's on time. This results in the need to express three regressions for assessing the linear dependence of the coefficients on time:

1. $\beta_{X1}(t) = \beta_{X1} + \theta_{X1}G_1(t), \quad (t \leq t_0)$,

2. $(\beta_{X1} + \beta_{X2})(t) = \beta_{X1} + \beta_{X2} + \theta_{X2}G_2(t), \quad (t > t_0)$,

3. $\beta_k(t) = \beta_k + \theta_k G(t)$.

The null hypotheses for proportional hazards correspond to $\theta_{X1} = 0$, $\theta_{X2} = 0$ and $\theta_k = 0$ $(k = 1, ..., q)$ for regression models 1, 2 and 3 respectively.

In order to assess these null hypotheses, use of the scaled Schoenfeld residuals are required, as for a standard Cox proportional hazards model. This is where issues arise with the current methods.

7.3 Overview and Issues of Current Methods

As outlined in Section 5.2, in order to assess the proportional hazards assumptions, we first express a model with a time-dependent coefficient,

$$h_i(t) = h_0(t)\exp(\boldsymbol{\beta}(t)'\boldsymbol{x}_i),$$

and assess if $\beta_k(t)$ is constant. Grambsch and Therneau (1994) showed that for the estimated coefficient $\hat{\beta}$ from standard Cox regression, we have

$$\mathbb{E}(s^*_{jk}) + \hat{\beta}_k \approx \beta_k(t_j), \tag{7.2}$$

where s^*_{jk} is a scaled Schoenfeld residual.

An analogy to generalised least squares can be used, where Therneau and Grambsch (2013) suggest expressing the linear dependence of the coefficient on time as a regression on some function of time, as in Equation (5.1), giving

$$\beta_k(t) = \beta_k + \theta_k G(t), \tag{7.3}$$

where $G(t)$ is a specified function of time, and the null hypothesis for proportional hazards corresponds to $\theta_k = 0$, for $k = 1, ..., q$.

The relationship between the expected value of the scaled Schoenfeld residuals and the coefficients given in Equation (7.2) enables use of the scaled Schoenfeld to test the null hypothesis for proportional hazards of $\theta_k = 0$, as outlined in Section 5.2.

The rationale for the regression in Equation (7.3) relates to the rationale for the three regressions expressed in Section 7.2 for assessing the piecewise-proportional hazards model. The piecewise-proportional hazards model can also be expressed as a model with time-dependent coefficients, but this time for each interval separately.

Like the standard Cox regression model, Schoenfeld residuals are obtained in the output of the piecewise-proportional hazards model. However, scaled residuals have better diagnostic power for assessing the proportional hazards assumption and so are used more often (Lee and Wang, 2003). This weighting of the Schoenfeld residuals is where the issues arise for the piecewise-proportional hazards model.

Therneau et al. (2019) suggest the proportional hazards assumption for the piecewise-proportional hazards model can be assessed using the Therneau (2015b) `cox.zph` function in R. This scales the Schoenfeld residuals using the overall covariance matrix of the $\hat{\beta}$'s as described in Section 5.3, however, this scaling does not take into account the disjoint nature of the piecewise-proportional hazards model.

As outlined in Section 5.2, the Schoenfeld residuals can be scaled using $\hat{V}_j^{-1}$, where $\hat{V}_j^{-1}$ is a good approximation of the estimated variance of $\hat{V}_j^{-1} s_j$. Grambsch and Therneau (1994) suggested the use of the approximation $\bar{V}$ for $\hat{V}_j^{-1}$, where $\bar{V} = \mathcal{I}(\hat{\beta})/d$, and this is now used as standard practice for assessing the proportional hazards assumption. In particular, this approximation is used as the default in the R `survival` package.

The simulations presented by Grambsch and Therneau (1994) show that the approximation has minimal affect on analyses, however this simulation only considered the case of binary covariates. Winnett and Sasieni (2001) investigated the suitability and impact of using this approximation further, considering possible scenarios where it may not be appropriate. The work by Winnett and Sasieni (2001) highlighted that for any covariate with more than two values, or levels, the approximation may not be appropriate, where in this case, there can be a decrease in covariate values in the risk set, particularly for extreme covariate values resulting in early event times. Under this circumstance, Winnett and Sasieni (2001) state the approximation $\bar{V}$ will be not close to $\hat{V}_j^{-1}$.

As the purpose of the test of the proportional hazards assumption is to examine the deviation of the hazard ratio from constant, it is important to ensure the scaling is appropriate, where Winnett and Sasieni (2001) highlighted that inappropriate scaling can result in under or over estimation of the deviation being investigated. We consider how the use of the approximation $\bar{V}$ may impact the assessment of the proportional hazards assumption for piecewise-proportional hazards, where the assumption that all the Schoenfeld residuals have the same variance, which is assumed to be proportional to the inverse of the Fisher's information matrix, will not hold for the piecewise-proportional hazards model.

More specifically, considering the nature of the piecewise-proportional hazards model, and the difference in coefficients for a covariate with a time-dependent effect, over specified time intervals, it is expected that the estimates of $\hat{V}_j^{-1}$ will be different for time intervals $t_j \leq t_0$ and $t_j > t_0$, where t_j is the jth event time and t_0 is the changepoint of the time-dependent covariate effect. This suggests that $\bar{V}$ would not be good approximation of $\hat{V}_j^{-1}$.

Further, the output of the piecewise-proportional hazards model produces a vector of Schoenfeld residuals for each coefficient in the model. For time-dependent covariate X, this results in two vectors of Schoenfeld residuals, one for events occurring prior to time t_0, and one for events after t_0. However, the length of these vectors is the total number of events over the full follow-up time, resulting in these vectors having multiple consecutive zeros for events within the time interval for which that coefficient does not represent.

As the Schoenfeld residuals correspond to the score function of the model likelihood, the issue around the presence of zero residuals is likely to result in an underestimation of elements of the Fisher's Information matrix of the model fit, in turn interfering with the scaling of the Schoenfeld residuals and impacting the score test for proportional hazards.

The presence of zero residuals can lead to underestimation of the variance of the non-zero residuals corresponding to events, and considering the zero-valued residuals within the residual matrix, any scaling procedure will cause them to become marginally non-zero. The presence of these residuals can also influence the score test, where, given their closeness to zero, inclusion of them when testing for deviation of the hazard ratio from constant, can exaggerate the result towards accepting the null of zero-coefficient and non-violation of proportional hazards.

To illustrate this issue further, and highlight how this also affects the standard visualisation method for assessing the proportional hazards assumption, Figure 7.1 presents an example visualisation of the scaled Schoenfeld residuals for a variable with a time-dependent covariate effect in the piecewise-proportional hazards model, plotted using the `cox.zph` function in R. The left plot presents the scaled Schoenfeld residuals and smooth curve for the coefficient representing events in the first time interval of $t \leq t_0$, and the right plot presents the residuals related to the coefficient of the interval $t > t_0$ for the time-dependent covariate, where $t_0 = 30$. These plots clearly show the grouping of zero-valued residuals which have now been scaled to be marginally non-zero, whilst highlighting how these residuals aggregate the smooth to be more constant. Given the score test assesses significance of the slope of a regression of the residuals against time, any amplification of the slope towards zero will result in lower power of detecting any remaining violation of the proportional hazards assumption for the coefficients of the time-dependent covariate.

To illustrate how the disjoint nature of the piecewise-proportional hazards model can affect the scaling of the Schoenfeld residuals for covariates without time-dependent coefficients, example residual plots are presented in Figure 7.2, where as in Figure 7.1, the time point for the change in effect of the time-dependent covariate effects is at $t_0 = 30$. Examining the residuals on either side of this time point, Figure 7.2 indicates that there is some difference in the variation of

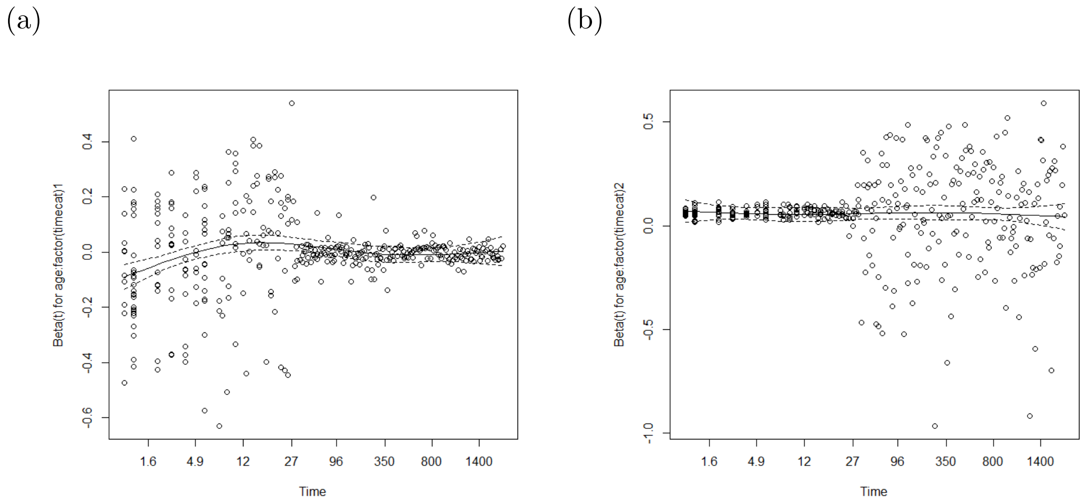


Figure 7.1: Plot of Schoenfeld residual plot for a time-dependent covariate in a piecewise-proportional hazards model, where the left plot represents the residuals for the coefficient of the first interval of $t \leq t_0$, and right for the interval $t > t_0$ ($t_0 = 30$).

the residuals between each interval for variables without a time-dependent effect. This suggests the variance of the scaled residuals may have been influenced by use of the estimation $\bar{V}$ and the presence of the zero-valued residuals.

Given these issues around the scaling of the Schoenfeld residuals, we propose the scaling of the Schoenfeld residuals should be handled separately for each time interval to avoid any influence of the poor variance estimate and zero-valued residuals on the assessment of the proportional hazards assumption.

7.4 Proposed Test of Proportional Hazards Assumption

In order to avoid the influence of the poor variance estimate and the zero-valued residuals when assessing the proportional hazards assumption, we propose it is necessary to consider the Schoenfeld residuals in separate pieces for each time interval. Below we assume a single change point and two time intervals, presenting a method for scaling the Schoenfeld residuals for each time interval and outlin-

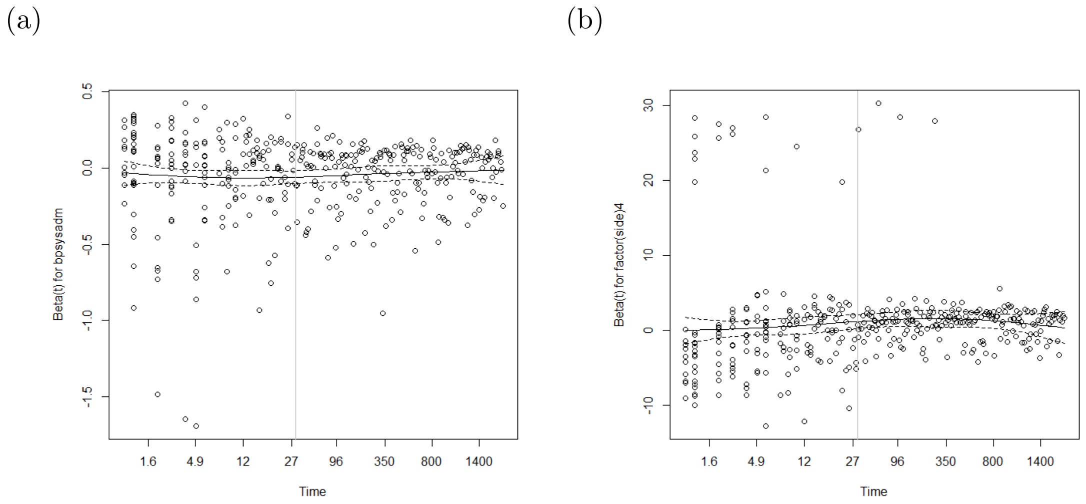


Figure 7.2: Examples of Schoenfeld residual plots for variables without a time-dependent covariate effect within a piecewise-proportional hazards model, with $t_0 = 30$.

ing how to proceed with testing the proportional hazards assumption afterwards. Further, we outline how this testing procedure can be supported by a suitable visualisation, and discuss how it may be extended when additional time intervals are needed.

7.4.1 Scaling Factor for Schoenfeld Residuals

In order to define a scaling factor for the Schoenfeld residuals, suppose there are $v = q+1$ explanatory variables included in a piecewise-proportional hazards model, one with a time-dependent covariate effect, X, and q with a constant effect, Z_k, $k = 1, ..., q$. Let t_0 be the time point at which the change in effect occurs for variable X, so that there are two time-intervals $t \leq t_0$, and $t > t_0$. The effects of X are constant but different either side of t_0. Suppose there are d_1 events prior to t_0 and d_2 events after t_0, where the total number of events is $d = d_1 + d_2$.

The matrix of Schoenfeld residuals would be a $d \times (v+1)$ matrix, where variable X has two columns of residuals, one for each coefficient. This would give a matrix of the form

$$S = \begin{bmatrix} s_{1X} & 0 & s_{1Z_1} & s_{1Z_2} & \cdots & s_{1Z_q} \\ s_{2X} & 0 & s_{2Z_1} & s_{2Z_2} & \cdots & s_{2Z_q} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ s_{d_1X} & 0 & s_{d_1Z_1} & s_{d_1Z_2} & \cdots & s_{d_1Z_q} \\ 0 & s_{d_1+1X} & s_{d_1+1Z_1} & s_{d_1+1Z_2} & \cdots & s_{d_1+1Z_q} \\ 0 & s_{d_1+2X} & s_{d_1+2Z_1} & s_{d_1+2Z_2} & \cdots & s_{d_1+2Z_q} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & s_{dX} & s_{dZ_1} & s_{dZ_2} & \cdots & s_{dZ_q} \end{bmatrix},$$

where s_{jk} is the Schoenfeld residual for covariate k at event time j. Now consider the Schoenfeld residuals as two separate matrices for the two time intervals and let S_1 be the matrix for events in the interval $t \leq t_0$, and S_2 be the matrix for events in the interval $t > t_0$. This would result in S_1 being a $d_1 \times v$ matrix and S_2 being a $d_2 \times v$ matrix, presented as

$$S_1 = \begin{bmatrix} s_{1_1X} & s_{1_1Z_1} & \cdots & s_{1_1Z_q} \\ s_{2_1X} & s_{2_1Z_1} & \cdots & s_{2_1Z_q} \\ \vdots & \vdots & \ddots & \vdots \\ s_{d_1X} & s_{d_1Z_1} & \cdots & s_{d_1Z_q} \end{bmatrix}, \qquad S_2 = \begin{bmatrix} s_{1_2X} & s_{1_2Z_1} & \cdots & s_{1_2Z_q} \\ s_{2_2X} & s_{2_2Z_1} & \cdots & s_{2_2Z_q} \\ \vdots & \vdots & \ddots & \vdots \\ s_{d_2X} & s_{d_2Z_1} & \cdots & s_{d_2Z_q} \end{bmatrix},$$

where s_{j_uk} is the Schoenfeld residual for covariate k at the jth event time of the corresponding time interval, u.

We propose each of these matrices, S_1 and S_2, can be scaled separately using an approximation of the Fisher's Information matrix for each time interval. Since the Cox regression model only outputs the overall Fisher's information matrix for the model, the variance would need to be computed for each event externally to the model output. Given the relationship between the Fisher's information, the score function, and how these relate to the Schoenfeld residuals, we have that $(S'S)$, the product of the Schoenfeld residual matrix and it's transpose, is asymptotically equivalent to the Fishers information matrix, and thus we can use this relationship to approximate the Fisher's information within each time interval.

Let u denote the time interval, so that $u = 1$ represents the time interval $t \leq t_0$ and $u = 2$ represents time interval $t > t_0$. We propose the Fisher's Information matrix for time interval u can be approximated as

$$\mathcal{I}_u(\hat{\beta}) \approx (S'_u S_u), \tag{7.4}$$

where $\mathcal{I}_u(\hat{\beta})$ is a $v \times v$ matrix.

Noting that, as in Equation (5.5), the Schoenfeld residuals are generally scaled as

$$S^* = d\mathcal{I}^{-1}(\hat{\beta})S,$$

we suggest the approximation of $\mathcal{I}_u(\hat{\beta})$ in Equation (7.4) can be used to weight the Schoenfeld residuals for time interval u. The matrices of Schoenfeld residuals, S_1 and S_2 respectively, can therefore be scaled as

$$S_1^* = S_1 d_1 (S'_1 S_1)^{-1}, \quad S_2^* = S_2 d_2 (S'_2 S_2)^{-1},$$

where S_1^* now denotes the $v \times d_1$ matrix of scaled Schoenfeld residuals for time interval $t \leq t_0$, and S_2^* denotes the $v \times d_2$ matrix of scaled Schoenfeld residuals for time interval $t > t_0$. The matrices of the scaled Schoenfeld residuals can now be presented as

$$S_1^* = \begin{bmatrix} s^*_{1_1X} & s^*_{1_1Z_1} & \cdots & s^*_{1_1Z_q} \\ s^*_{2_1X} & s^*_{2_1Z_1} & \cdots & s^*_{2_1Z_q} \\ \vdots & \vdots & \ddots & \vdots \\ s^*_{d_1X} & s^*_{d_1Z_1} & \cdots & s^*_{d_1Z_q} \end{bmatrix}, \qquad S_2^* = \begin{bmatrix} s^*_{1_2X} & s^*_{1_2Z_1} & \cdots & s^*_{1_2Z_q} \\ s^*_{2_2X} & s^*_{2_2Z_1} & \cdots & s^*_{2_2Z_q} \\ \vdots & \vdots & \ddots & \vdots \\ s^*_{d_2X} & s^*_{d_2Z_1} & \cdots & s^*_{d_2Z_q} \end{bmatrix}.$$

For each covariate, we now have two vectors of scaled Schoenfeld residuals, one for each time-interval. For variable X, the covariate with a time-dependent effect, the two vectors can be used individually to assess the proportional hazards assumption over each time period separately. The vectors of scaled Schoenfeld residuals for variable X can be defined as $\boldsymbol{s}^*_{Xu}$ for time interval u, where $\boldsymbol{s}^*_{Xu} = (s^*_{1_uX}, s^*_{2_uX}, ..., s^*_{d_uX})'$, the column of residuals for X in S^*_u. The vector of residuals $\boldsymbol{s}^*_{Xu}$ can now be used in the test for proportional hazards in time interval u for covariate X.

In order to check the proportional hazards assumption for the remaining covariates, Z_k, which do not have a time-dependent covariate effect in the piecewise-proportional hazards model, the two vectors of scaled residuals need stacking together so the assumption can be assessed over the full follow-up time. Before these vectors can be stacked, the residuals need re-weighting to account for the difference in the scaling factors.

Let Λ_u be a $v \times v$ matrix denoting the scaling factor for the Schoenfeld residual matrix for time interval u, so that $\Lambda_u = d_u \mathcal{I}_u^{-1}(\hat{\beta})$. Now, exclude the rows and columns of Λ_u corresponding to variable X to form a $q \times q$ matrix whose diagonal entries, (k, k), correspond to the variance elements of $\mathcal{I}_u^{-1}(\hat{\beta})$ relating to Z_k. Denote this matrix as Λ_{uZ}.

To stack the scaled Schoenfeld residual vectors for Z_k from each time interval, the residuals from time interval $u = 2$ can be re-weighted using the square root of the ratio of the variance elements of $\mathcal{I}_u^{-1}(\hat{\beta})$ relating to Z_k, or rather the ratio of the (k,k) entries of Λ_{1Z} and Λ_{2Z}. Let $\tilde{\lambda}_k$ denote the weighting factor for Z_k's residuals, defined as

$$\tilde{\lambda}_k = \sqrt{\frac{\Lambda_{1Z(k,k)}}{\Lambda_{2Z(k,k)}}}, \tag{7.5}$$

and note the vector of scaled Schoenfeld residuals for time interval $u = 2$ and covariate Z_k from S_2^* is defined as $\boldsymbol{s}^*_{Z_k2} = (s^*_{1_2Z_k}, s^*_{2_2Z_k}, ..., s^*_{d_2Z_k})'$. The re-weighted vectors of residuals for time interval $u = 2$ for Z_k, denoted $\tilde{\boldsymbol{s}}^*_{Z_k2}$, are then given as

$$\tilde{\boldsymbol{s}}^*_{Z_k2} = \tilde{\lambda}_k \boldsymbol{s}^*_{Z_k2}.$$

The residual vectors for Z_k can then be stacked to become $\tilde{\boldsymbol{s}}^*_{Z_k} = (\boldsymbol{s}^*_{Z_k1}, \tilde{\boldsymbol{s}}^*_{Z_k2})'$, or $\tilde{\boldsymbol{s}}^*_{Z_k} = (s^*_{1_1Z_k}, s^*_{2_1Z_k}, ..., s^*_{d_1Z_k}, \tilde{s}^*_{1_2Z_k}, \tilde{s}^*_{2_2Z_k}, ..., \tilde{s}^*_{d_2Z_k})'$, a $d \times 1$ vector which can now be used to assess the proportional hazards assumption for variable Z_k.

Recall that $\boldsymbol{s}^*_{X1}$ and $\boldsymbol{s}^*_{X2}$ are respectively the $d_1 \times 1$ and $d_2 \times 1$ vectors of scaled Schoenfeld residuals for variable X for the time intervals $t \leq t_0$, and $t > t_0$, and $\tilde{\boldsymbol{s}}^*_{Z_k}$ denotes the vector for variable Z_k, $k = 1, ..., q$, including the residuals for the whole of follow-up. Let g_j be a function of time, where g_{1j} and g_{2j} denote the function of time at the event times in time interval $u = 1$ and $u = 2$ respectively. Note that $\bar{g} = \sum_{j=1}^{d} g_j$ and $\bar{g_u} = \sum_{j=1}^{d_u} g_{uj}$. The tests for proportional hazards given in Section 7.2 can now be expressed as three corresponding linear regression models:

1. $s^*_{j_1X} = \theta_{X1}(g_{1j} - \bar{g}_1)$
2. $s^*_{j_2X} = \theta_{X2}(g_{2j} - \bar{g}_2)$
3. $\tilde{s}^*_{jZ_k} = \theta_{Z_k}(g_j - \bar{g})$,

where $\theta_{X1} = 0$, $\theta_{X2} = 0$ and $\theta_{Z_k} = 0$ correspond to the null hypothesis of proportional hazards for the regressions 1, 2 and 3 respectively.

Referring back to the test outlined in Section 5.3.1, we can estimate these θ coefficients and calculate a test statistic to conduct a score test for assessing the proportional hazards assumption. For the variables Z_k, we can estimate θ_{Z_k} from model 3 as

$$\hat{\theta}_{Z_k} = \frac{\sum_{j=1}^{d} (g_j - \bar{g})\, s^*_{jZ_k}}{\sum_{j=1}^{d} (g_j - \bar{g})^2},$$

with test statistic

$$T_k = \frac{\left\{\sum_{j=1}^{d} (g_j - \bar{g})\, s^*_{jZ_k}\right\}^2}{\left(\Lambda_{1Z(k,k)} + \Lambda_{2Z(k,k)}\right) \sum_{j=1}^{d} (g_j - \bar{g})^2}.$$

For covariate X with a time-dependent covariate effect, the estimates of the coefficients θ_{X1} and θ_{X2}, and their corresponding test statistics, are

$$\hat{\theta}_{X1} = \frac{\sum_{j=1}^{d_1} (g_{1j} - \bar{g}_1)\, s^*_{jX_1}}{\sum_{j=1}^{d_1} (g_{1j} - \bar{g}_1)^2}, \qquad T_{X1} = \frac{\left\{\sum_{j=1}^{d_1} (g_{1j} - \bar{g}_1)\, s^*_{jX_1}\right\}^2}{d_1 (S_1' S_1)^{-1}_{(X,X)} \sum_{j=1}^{d_1} (g_{1j} - \bar{g}_1)^2},$$

and

$$\hat{\theta}_{X2} = \frac{\sum_{j=1}^{d_2} (g_{2j} - \bar{g}_2)\, s^*_{jX_2}}{\sum_{j=1}^{d_2} (g_{2j} - \bar{g}_2)^2}, \qquad T_{X2} = \frac{\left\{\sum_{j=1}^{d_2} (g_{2j} - \bar{g}_2)\, s^*_{jX_2}\right\}^2}{d_2 (S_2' S_2)^{-1}_{(X,X)} \sum_{j=1}^{d_2} (g_{2j} - \bar{g}_2)^2}.$$

respectively, where $d_u (S_u' S_u)^{-1}_{(X,X)}$ is the diagonal entry of $d_u (S_u' S_u)^{-1}$ corresponding to covariate X.

7.4.2 Visualisation

As recommended by Hosmer et al. (2008), it is also useful to use a visualisation of the residuals alongside the formal test. Figures 7.1 and 7.2 highlighted that the visualisations produced by the `cox.zph` function in R are not appropriate for the piecewise-proportional hazards model since they do not handle the residuals for the time-dependent covariate effects separately.

We propose that for a variable with a time-dependent covariate effect, two plots should be produced, one for the residuals from each time interval. Using the scaling technique outlined in Section 7.4.1, we have vectors of scaled residuals which can be plotted against the chosen function of time for each time interval. More specifically $\boldsymbol{s}^*_{X1}$ can be plotted against g_1 to visualise any violation of proportional hazards in the effect of X for $t \leq t_0$, and likewise, $\boldsymbol{s}^*_{X2}$ can be plotted against g_2 for time interval $t > t_0$.

Further, as outlined in Section 5.3.2, a smooth curve and corresponding confidence interval can be added to these plots. To define the plotted values of the spline curves, firstly suppose U_{Xu} is the matrix of basis vectors for the spline fit of the scaled Schoenfeld residuals on g_u, and let C_{Xu} be the matrix for the same spline functions evaluated at the p plotting points. The plotted values of the spline curve for variable X, with a time-dependent covariate effect, on interval u can be defined as

$$\hat{y}_{Xu} = \mathbf{1}\hat{\beta}_{Xu} + C_{Xu}(U'_{Xu}U_{Xu})^{-1}U'_{Xu}\boldsymbol{s}^*_{Xu} \equiv \mathbf{1}\hat{\beta}_{Xu} + H_{Xu}\boldsymbol{s}^*_{Xu},$$

and the variance of $\hat{y}_{Xu}$ is given by

$$\mathrm{Var}(\hat{y}_{Xu}) = d_u(S'_uS_u)^{-1}_{(X,X)}H_{Xu}H'_{Xu}.$$

As stated in Section 5.3.2, the confidence intervals can be constructed using standard linear model calculations.

Now, for variables Z_k, a single plot for each Z_k is required to visualise proportional hazards over the full follow-up. This can be done by plotting the stacked vector of scaled and re-weighted Schoenfeld residuals against some function of time, or more specifically plotting $\tilde{\boldsymbol{s}}^*_{Z_k}$ against g. Again, a smooth curve needs to be added to help in the interpretation. Let U_Z be the matrix of basis vectors for the spline fit of $\tilde{\boldsymbol{s}}^*_{Z_k}$ on g, with C_Z being the matrix of the spline functions evaluated at the plotting points. The plotted values of the spline curve for variable Z_k

can be defined as

$$\hat{y}_{Z_k} = \mathbf{1}\hat{\beta}_{Z_k} + C_Z(U_Z'U_Z)^{-1}U_Z'\tilde{\boldsymbol{s}}^*_{Z_k} \equiv \mathbf{1}\hat{\beta}_{Z_k} + H_Z\tilde{\boldsymbol{s}}^*_{Z_k},$$

where the variance of $\hat{y}_{Z_k}$ is defined as

$$\mathrm{Var}(\hat{y}_{Z_k}) = (\Lambda_{1Z} + \Lambda_{2Z})_{(k,k)}H_ZH_Z'.$$

Again, the confidence intervals can be constructed using standard linear model calculations.

7.4.3 Consideration of Number of Changepoints

For completeness, it is important to note how the number and location of changepoints in effect can be determined, and how the methods outlined above in Sections 7.4.1 and 7.4.2 can be applied when more than one changepoint is needed.

Firstly, considering the choice of time points for any changes in effects, the standard visualisation of the Schoenfeld residuals can be used in the first instance to identify possible changepoints. Any turning points or changes in gradient to the smooth curve can indicate possible time points for a change in effect. If there are multiple occurrences of turning points or large changes to gradient of the curve, this may indicate multiple changepoints. To assess the suitability of possible change points, a range of possible choices can be compared through fitting piecewise-proportional hazards models with varying changepoints and comparing their fit using the Akaike Information Criterion.

If more than one change in effect is found, resulting in more than two time intervals for fitting the model over, the methods outlined in Sections 7.4.1 and 7.4.2 can be easily extended. Considering the form of the residual matrices outlined in section 7.4.1, the number of coefficients for time-dependent covariate X would equate the number of time intervals, and this would equate the number of columns for X in the residual matrix S. Again, residuals matrices can be specified for each time interval, u, separately. Here it is important to ensure to use the column of residuals which corresponds to the appropriate coefficient of X and the appropriate rows of residuals for Z_k which correspond to time interval u. These residual matrices can be scaled in the same way, using the approximation of the Fisher's Information matrix for each interval as in Equation (7.4).

In order to re-weight the residuals for variables Z_k with time-constant effects, define $\tilde{\lambda}_k$ to be as in Equation (7.5), but replace $\Lambda_{2Z(k,k)}$ with $\Lambda_{uZ(k,k)}$, for each time interval u. The scaled, re-weighted residuals for each Z_k from each time interval can then be stacked as previously outlined. The test statistics and smooth curves can be fitted as outlined in Sections 7.4.1 and 7.4.2, ensuring to account for the variance terms for each of the additional time intervals.

7.5 Extension to Multiply Imputed Data

The methods outlined in Section 7.4 for testing the proportional hazards assumption for a piecewise-proportional hazards model can be extended to the multiple imputation setting.

Firstly, we consider the formal test. Fitting a piecewise-proportional hazards model to multiply imputed data sets, results in M sets of the residual matrix S, denoted $S^{(m)}$, $m = 1, ..., M$. For each imputed data set m, the Schoenfeld residual matrix can be be split and scaled separately using the methods outlined in Section 7.4.1. This results in M vectors of scaled Schoenfeld residuals $\boldsymbol{s}_{Xu}^{(m)*}$ for the variable X with a time-dependent covariate effect, and M vector $\tilde{\boldsymbol{s}}_{Z_k}^{(m)*}$ for the q variables, Z_k, $k = 1, ..., q$, assumed to have time-constant effects.

As outlined in Section 7.4.1, the tests for proportional hazards can be expressed as three linear regression models:

1. $s_{j_1 X}^{(m)*} = \theta_{X1}^{(m)}(g_{1j} - \bar{g_1})$

2. $s_{j_2X}^{(m)*} = \theta_{X2}^{(m)}(g_{2j} - \bar{g}_2)$

3. $\tilde{s}_{jZ_k}^{(m)*} = \theta_{Z_k}^{(m)}(g_j - \bar{g})$.

For covariates with time-dependent coefficients, the methods outlined in Section 5.3 still hold for assessing the proportional hazards assumption over multiply imputed data sets, for the both the formal test and visualisation. This holds since, for each time interval separately, the Schoenfeld residuals are scaled by a constant scaling factor, and can be regressed against the centered function of time for the corresponding interval, as in regressions 1 and 2 above.

For the covariates considered to have a constant coefficient over the whole of follow-up, assessing the proportional hazards assumption requires extra consideration as we have stacked two vectors of residuals, which have been scaled using different scaling factors, and thus do not necessarily have constant variance. This affects the derivations shown in Section 5.3.1, where the correction factor within Q will now no longer be exactly zero.

Considering the formula for Q, given in Equation (5.3), it is possible to derive Q for the piecewise-proportional hazards model, where instead of $\bar{V}$ as an estimate of the overall variance, we can take approximations $\bar{V}_1$ and $\bar{V}_2$ as the variance estimates for the two time intervals. An exact result for the correction factor being zero cannot be shown, but we aim to show that the correction factor within Q is small enough so that the proportional hazards can be tested using the regression given in model 3 above.

We can estimate $\bar{V}_u$, for $u = 1, 2$ using the approximation of the Fisher's information matrix for each interval, as given in Equation (7.4). For $u = 1$, this gives

$$\bar{V}_1 = d_1^{-1}\mathcal{I}_1(\beta) \approx d_1^{-1}(S_1'S_1),$$

where d_1 is the number of events in time interval $u = 1$. For $u = 2$, we also need to consider the additional scalar $\tilde{\lambda}_k$ used to re-weight the residuals. This gives the approximation of the variance for interval $u = 2$ as

$$\bar{V}_2 = d_2^{-1}\tilde{\boldsymbol{\lambda}}^{-1}\mathcal{I}_2(\beta) \approx d_2^{-1}\tilde{\boldsymbol{\lambda}}^{-1}(S_2'S_2),$$

where $\tilde{\boldsymbol{\lambda}}$ is a $k \times k$ matrix with the (k, k)th diagonal elements being λ_k, as in Equation (7.5), and d_2 is the number of events in time interval $u = 2$, so that $d_1 + d_2 = d$, the total number of events. Substituting the $\bar{V}_u$ terms into Equation (5.3) gives Q as

$$\begin{aligned} Q = &\sum_{j=1}^{d_1} G_j\bar{V}_1G_j + \sum_{j=d_1+1}^{d} G_j\bar{V}_2G_j \\ &- \left(\sum_{j=1}^{d_1} G_j\bar{V}_1 + \sum_{j=d_1+1}^{d} G_j\bar{V}_2\right)\left(d_1\bar{V}_1 + d_2\bar{V}_2\right)^{-1}\left(\sum_{j=1}^{d_1} G_j\bar{V}_1 + \sum_{j=d_1+1}^{d} G_j\right)' \end{aligned} \tag{7.6}$$

As in Section 5.3, we consider centering the function of time, so that $G_j = g_j - \bar{g}$, where $\bar{g} = d^{-1}\sum_{j=1}^{d} g_j$. Note that the function of time is centered over the whole follow-up and not within the individual time intervals. We aim to show that the correction factor term in Equation (7.6) is small and thus inconsequential.

Substituting the centered function of time into Equation (7.6), and showing the $\bar{V}_u$ in terms of $\mathcal{I}_u(\beta)$, gives

$$\begin{aligned} Q = &\sum_{j=1}^{d_1}(g_j - \bar{g})[d_1^{-1}\mathcal{I}_1(\beta)](g_j - \bar{g}) + \sum_{j=d_1+1}^{d}(g_j - \bar{g})[d_2^{-1}\tilde{\boldsymbol{\lambda}}^{-1}\mathcal{I}_2(\beta)](g_j - \bar{g}) \\ &- \left(\sum_{j=1}^{d_1}(g_j - \bar{g})d_1^{-1}\mathcal{I}_1(\beta) + \sum_{j=d_1+1}^{d}(g_j - \bar{g})d_2^{-1}\tilde{\boldsymbol{\lambda}}^{-1}\mathcal{I}_2(\beta)\right)[\mathcal{I}_1(\beta) + \tilde{\boldsymbol{\lambda}}^{-1}\mathcal{I}_2(\beta)]^{-1} \\ &\times \left(\sum_{j=1}^{d_1}(g_j - \bar{g})d_1^{-1}\mathcal{I}_1(\beta) + \sum_{j=d_1+1}^{d}(g_j - \bar{g})d_2^{-1}\tilde{\boldsymbol{\lambda}}^{-1}\mathcal{I}_2(\beta)\right)'. \end{aligned} \tag{7.7}$$

Taking the first set of brackets within the correction factor, we can simplify to get

$$d_1^{-1} d_2^{-1} \left(d_2 \mathcal{I}_1(\beta) \sum_{j=1}^{d_1} (g_j - \bar{g}) + d_1 \tilde{\boldsymbol{\lambda}}^{-1} \mathcal{I}_2(\beta) \sum_{j=d_1+1}^{d} (g_j - \bar{g}) \right). \tag{7.8}$$

Centering the function of time over the full follow-up means that the two summation terms in Equation (7.8) are of opposing signs, where $\sum_{j=1}^{d_1}(g_j - \bar{g})$ will be a negative term, and $\sum_{j=d_1+1}^{d}(g_j - \bar{g})$ will be a positive term, since the early event times will be less than $\bar{g}$, and the later event times will be greater than $\bar{g}$. Subject to the additional scalars of $\mathcal{I}_u(\beta)$ and d_u, the addition of these two summation terms should therefore be reasonably close to zero, particularly when the variance terms $\bar{V}_1$ and $\bar{V}_2$ are similar. The re-weighting term $\tilde{\lambda}_k$ is intended to ensure closeness between the variance terms.

The term in Equation (7.8) is the transpose of the third bracket term in the correction factor in Equation (7.7), and thus overall from this simplification in Equation (7.8), we can see that the correction factor is divided through by the term $(d_1 d_2)^2$. With possible multiplicative terms of d_1 and d_2 in each, this gives the smallest divisor of the correction term as d_1^2, assuming $d_1 < d_2$. For a survival model with a time-split, there should be a sufficient proportion of events within each interval, and thus the inequality $d_1^2 > d$ should hold. Therefore for large d, the correction factor can be assumed to be negligibly small, and hence the linear regression model, numbered 3 above, can be used to assess the proportional hazards assumption of covariates Z_k with constant coefficient, where the Wald test can be used to assess if $\hat{\theta}_{Z_k}$ is significantly non-zero.

Now, through fitting the linear regression models defined above, the M estimates of the θ coefficients, denoted $\hat{\theta}_{Xu}^{(m)}$ and $\hat{\theta}_{Z_k}^{(m)}$, and their corresponding variance, can be combined using Rubin's rules. The coefficients can be combined over the M imputed data sets using Equation (5.12), and the variance estimates can be combined by taking the between-imputation and within-imputation variance estimates, as stated in Equations (5.13) and (5.14) respectively, and combining these using Equation (5.15) to give the pooled variance estimate. For each of the three regressions, the Wald test statistic can then be calculated from the combined estimates to assess the proportional hazards assumption.

For the visualisation, the M sets of the scaled Schoenfeld residuals, $\boldsymbol{s}_{Xu}^{(m)*}$ and $\tilde{\boldsymbol{s}}_{Z_k}^{(m)*}$, can be plotted by taking their average and plotting this against the corresponding functions of time, g_u or g respectively. The smooth curve to help with interpretation can then be added to the plots using the method outlined in Section 5.3.2. This involves fitting spline fits to each of the M sets of scaled residuals separately, combining their plotting points $\hat{y}_{Xu}^{(m)}$ and $\hat{y}_{Z_k}^{(m)}$, and corresponding variance estimates. This can be done using Rubin's rules, as given in Equations (5.17) and (5.18) for the plotting points and their variance respectively, to achieve the pooled smooth curves for each of the covariates.

7.6 Simulation Study

In order to assess the appropriateness of the using the approximation in Equation (7.4), a simulation study was carried out to examine the differences in Type 1 error of testing the proportional hazards assumption, dependent upon which scaling technique is used for the Schoenfeld residuals. The two scaling methods differ in the use of the Fisher's information matrix of the model fit, or our proposed approximation for this in Equation (7.4).

This simulation was an add-on to the simulation study presented in Chapter 5, so has the same set-up, where again 5000 simulations were produced for $n = 1000$ individuals. Three covariates were simulated, $X1$, $X2$, and $X3$, where $X1$ and $X3$ were continuous and $X2$ was binary, with $X2$ dependent on $X1$, and $X3$ dependent on both $X1$ and $X2$, as outlined in Section 5.4.

The survival times were generated from covariates $X1$, $X2$, and $X3$ using the simsurv package in R under a Weibull distribution, where the covariate effects of $X1$, $X2$, and $X3$ were all assumed to satisfy the proportional hazards assumption. The number of events was varied to produce four scenarios, where the proportions of individuals simulated to have events in the four scenarios were 20%, 40%, 60% and 80%.

Within each scenario, a Cox regression model was fitted to the survival data, adjusted for all three covariates, $X1$, $X2$, and $X3$. The proportional hazards assumption was assessed in four ways for each covariate within each scenario, where we considered both the score test, as outlined by Grambsch and Therneau (1994), and the Wald test, as discussed in Section 5.3.1. For each of these tests, we considered two approaches to scaling the Schoenfeld residuals. These were the use of the standard scaling factor $d\mathcal{I}^{-1}(\hat{\beta})$, or our proposed scaling factor, $d(S'S)^{-1}$. These are referred to as 'Fisher' and 'Approx', respectively, within the results section.

In order to assess if the approximation in Equation 7.4 is a reasonable alternative to using the Fisher's information from the model fit, the Type 1 error was recorded for each covariate for each test. The Type 1 error was given as the proportion of the 5000 simulations where the p-value of the test of the proportional hazards assumption was less than the significance level of $\alpha = 0.05$.

7.6.1 Type 1 Error Results

Table 7.1 presents the results of the simulation study, and shows overall that there is very minimal difference between the two scaling methods in terms of the Type 1 error of the tests. The maximum difference in Type 1 error shown in Table 7.1 is for the score test for covariate $X3$, with 20% events, but this difference was still small at 0.005. Most of the results in Table 7.1 show a difference in Type 1 error 0.001 or less between the two scaling types.

Given how minimal the difference in Type 1 error is between the two scaling techniques, for both the Wald and score tests, we can conclude that use of the approximation in Equation (7.4) for the Fisher's information is a reasonable alternative. Therefore this approximation can be used to provide a more computationally convenient approach within the setting of piecewise-proportional hazards to give an approximation of the Fisher's information for each time interval separately.

Table 7.1: Simulation results giving Type 1 error of test of proportional hazards assumption using Score or Wald test method on complete data for two scaling methods of the Schoenfeld residuals: $d\mathcal{I}^{-1}(\hat{\beta})$ (Fisher) or $d\,(S'S)^{-1}$ (Approx) (n=1000, 5000 Simulations).

Events	**Test**	**Scale**	**Type 1 Error**		
(%)	**Type**	**Type**	$X1$	$X2$	$X3$
20	Score	Fisher	0.051	0.052	0.053
		Approx	0.051	0.054	0.048
	Wald	Fisher	0.050	0.052	0.049
		Approx	0.050	0.053	0.048
40	Score	Fisher	0.057	0.046	0.049
		Approx	0.054	0.047	0.050
	Wald	Fisher	0.053	0.046	0.049
		Approx	0.054	0.046	0.049
60	Score	Fisher	0.053	0.049	0.051
		Approx	0.051	0.049	0.052
	Wald	Fisher	0.050	0.048	0.052
		Approx	0.051	0.049	0.052
80	Score	Fisher	0.056	0.045	0.050
		Approx	0.054	0.045	0.051
	Wald	Fisher	0.056	0.045	0.052
		Approx	0.055	0.046	0.050

7.7 Conclusion

This chapter provided an adaptation of the standard method for assessing the proportional hazards assumption so that the piecewise-proportional hazards model can be validated appropriately.

Initially, the unsuitably of the use of the standard method for a piecewise-proportional hazards model was highlighted, showing the potential bias of the test towards the null of proportional hazards if the Schoenfeld residuals are not scaled appropriately. An alternative scaling technique for the Schoenfeld residuals was then proposed, taking into the account the disjoint nature of the piecewise-proportional hazards model. This alternative scaling was shown to be applicable for both a formal test and visualisation of proportional hazards, where the simulation study evidenced the suitability of the approximation used within the scaling.

Methods for conducting a formal test and producing suitable visualisations were outlined, where emphasis was placed on the disjoint nature of the piecewise-proportional hazards model. The methods proposed suggested separate assessment of the proportional hazards assumption over the two or more time periods for the time-dependent covariate effects, whilst still allowing for assessment over the full follow-up time for covariates with time-constant effects.

Finally, an extension of these methods was provided for application to the multiply imputed data setting. This brought together the methods in Chapters 5 and 7, ready for application in Chapter 8.

Chapter 8

Application to Stroke: Part 2

8.1 Introduction

Following on from the initial analyses of the stroke audit data in Chapter 4, this chapter addresses the issues found during the model validation of the initial model building procedure by applying the methods outlined in Chapters 6 and 7.

Firstly this chapter recaps the results of initial analyses in Chapter 4. Issues around interactions and functional form of the baseline covariates are addressed, and the non-proportional hazards are further examined to identify the time post-stroke of the changepoint of the time-dependent covariate effects.

Subsequently, the multiple imputation procedure is detailed, demonstrating the application of the methodology outlined in Chapter 6. Here this chapter discusses how the imputation procedure accounts for time-dependent covariate effects, interactions, and non-linear covariate effects.

Following the imputation process, this chapter outlines the model building procedure, giving interpretation of the resulting fitted model in the context of stroke survival.

Finally, assessment of the proportional hazards assumption is provided, demonstrating the application of the model validation techniques given in Chapter 7.

8.2 Summary of Application to Stroke: Part 1

The initial analyses outlined in Chapter 4 identified several covariates to be important for survival post-stroke, where the model validation in Sections 4.6 and 5.5 highlighted that many of these effects may depend upon time post-stroke.

The baseline covariates identified to be significant for survival in the adjusted model were age at time of stroke, hospital admitted to, pre-stroke mobility, side of lesion, diabetes mellitus, systolic BP at hospital admission, lesion type shown in CT scan and worst consciousness level in the first 24 hours post-stroke.

The results of the pooled adjusted model, as shown in Table 4.17 and Figure 4.26, indicated that increased age, poorer mobility prior to stroke, diabetes, and worsened consciousness all increased hazard of death following stroke. Further, considering lesion types, a lesion on both sides of the brain, and PICH shown in CT scan, were found to result in the highest hazard. Patients who did not receive a CT scan had the highest hazard overall compared to any of the identified lesion types. Admission to Hospital 2 was found to give a reduced hazard of death, as was higher systolic BP on admission to hospital.

Model validation highlighted several issues with the fit of the model. Section 8.3 discusses these in more depth, and shows how these can be handled.

8.3 Handling issues found during model validation

The validation of the initial modelling procedure, as given in Sections 4.6 and 5.5, highlighted three key issues which need addressing: a possible interaction between hospital and no CT scan, non-linear covariate effects, and time-dependent covariate effects. Throughout this section, we explore each of these issues further and show how these can be handled to improve the model fit.

8.3.1 Interaction between hospital and no CT scan

During the initial model building procedure, we found the effect of hospital to have an unexpected interpretation, where the hazard ratio for hospital in the adjusted model opposed the univariate hazard ratio of hospital. Table 4.19 in Section 4.6 showed that there appeared to be differing practices between the hospitals regarding whether or not patients received a CT scan, dependent upon the level of consciousness of patients. Further, Table 4.20 indicated that there was a 5% higher incidence of death for patients who did not have a CT scan at Hospital 1 compared to Hospital 2.

To examine if an interaction between 'no scan' and hospital resolves the unexpected hazard ratio for hospital in the adjusted model in Chapter 4, we incorporated this interaction into a Cox model. The results of the Cox regression model for the effect of lesion type shown in CT scan on survival, including an interaction between 'no scan' and hospital, are given in Table 8.1. The results in Table 8.1 highlight the need incorporate this interaction into further model building procedures for the stroke audit data, where the results show that the hazard ratio for hospital is now concurrent with the results of the univariate model for hospital, in Table 4.13. Table 8.1 also indicates there is a difference in the effects of 'no scan' between the two hospitals on survival post-stroke, where the hazard ratios are 4.7 and 3.2 for 'no scan' at hospitals 1 and 2, respectively.

8.3.2 Functional form of non-linear covariate effects

The model validation in Section 4.6 also found issues around the functional form of systolic BP, suggesting a quadratic term for systolic BP may need to be incorporated into the model building. Given the effect of systolic BP was found to be non-linear in Section 4.6, here we also examine the functional form of diastolic BP to ensure both these covariates are included in the imputation procedure in the correct functional form.

Table 8.1: Results of Cox regression model fitted to complete data for survival of stroke patients on lesion type shown in CT scan, with an interaction between 'no scan' and hospital.

Variable	HR	95% CI	p-value
Hospital (Hospital 1)			
Hospital 2	0.925	(0.699,1.224)	0.586
CT Scan: Lesion Type (None)			
CI	1.376	(1.005,1.882)	0.046
HCI	1.944	(1.048,3.605)	0.035
PICH	1.889	(1.253,2.846)	0.002
No Scan @ Hosp. 1	4.692	(3.080,7.147)	<0.001
No Scan @ Hosp. 2	3.228	(2.231,4.672)	<0.001

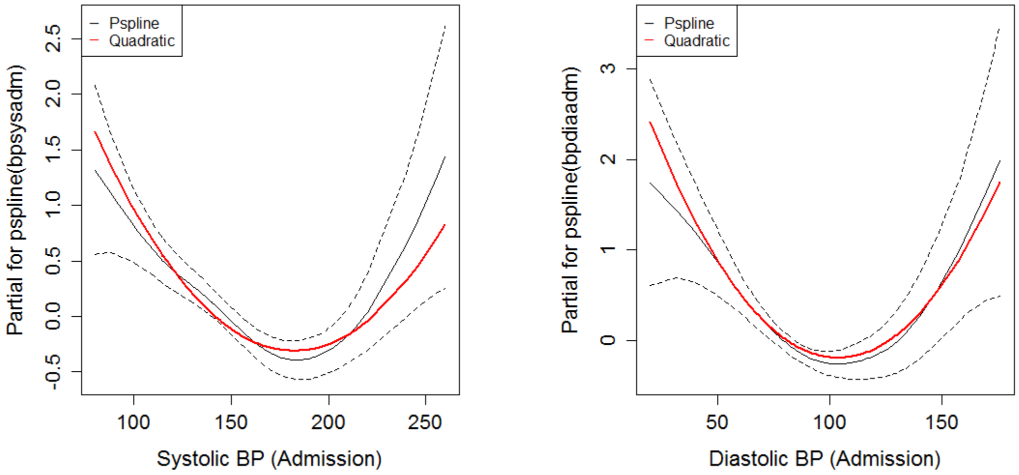
Figure 8.1 gives plots of the spline fits for the blood pressure measures. The corresponding quadratic effects of the BP measures are overlaid onto the plots in Figure 8.1, where the quadratic effect is the resulting effect of the measure when it is fitted in a Cox model as a linear term plus the quadratic term for the measure.

As previously stated in Section 4.6, adding a quadratic term for systolic BP appears to be an appropriate fit of its functional form. Figure 8.1(a) reiterates this, with the spline and quadratic curves being close. Figure 8.1(b) also shows the shape of the quadratic fit to be close to the spline fit for diastolic BP, indicating a quadratic term should also be included for diastolic BP during further imputation and model building procedures for the stroke audit data.

8.3.3 Change point of time-dependent covariate effects

The main issue highlighted by model validation, in both Sections 4.6 and 5.5, was that several covariates in the adjusted Cox model violated the proportional hazards assumption, indicating the presence of time-dependent covariate effects.

To incorporate the time-dependent covariate effects, a piecewise-proportional hazards model can be used. In order to appropriately fit the time-dependent covariate effects within a piecewise-proportional hazards model, it is necessary to



(a) Systolic BP

(b) Diastolic BP

Figure 8.1: Visualisation of functional form of BP, showing plots of the spline fit (black) and quadratic fit (red) for the effects of: (a) Systolic BP; (b) Diastolic BP, on survival post-stroke.

identify the time point, or time points, at which the change in the effect of the covariates occur.

The plots of the Schoenfeld residuals in Section 5.5 indicated that the change-point appears to be within the first 100 days post-stroke, possibly around the first month post-stroke. Considering previous research, a week and a month post-stroke have previously been identified to be key time points for survival (Easton et al., 2014; Andersen and Olsen, 2011; Petty et al., 2000).

To identify an appropriate changepoint, adjusted piecewise-proportional hazards models with varying changepoints were fitted to each of the imputed data sets from the initial analysis in Chapter 4. The Akaike Information Criterion (AIC) was used to compare the fit of the models for each changepoint. We assumed a single change in effect and explored changepoints at each of the first 10 days post-stroke, then in increments of 5 days up to 30 days. We also examined changepoints at 40, 50, 70 and 100 days post-stroke.

Figure 8.2 gives a plot of the AIC values for each of the piecewise-proportional hazards models, with different time points of change in effects, fitted to multiply imputed data set 1. It is clear in Figure 8.2 that a change in effect at 7 days post-stroke minimises the AIC and thus could be the most appropriate choice. Similar plots for the remaining imputed data sets also showed 7 days to be the optimal changepoint which minimised the AIC.

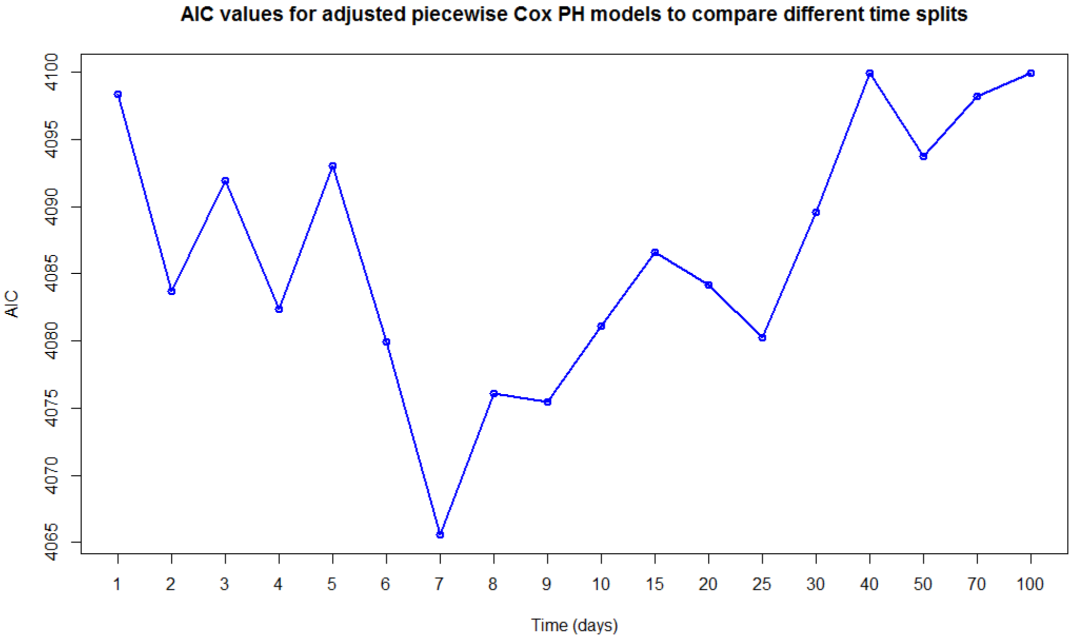


Figure 8.2: Plot of AIC results for piecewise proportional hazards models with varying time points of change in effects to identify optimal change point

8.4 Multiple Imputation

8.4.1 Imputation Procedure

After the naïve imputation approach in Chapter 4, the incomplete data were re-imputed using the methods outlined in Chapter 6 to incorporate the time-dependent covariate effects found during the initial model validation. The imputation procedure was again implemented through multiple imputation using chained equations (MICE) using the mice package in R, and the missing data mechanism was assumed to be MAR, as discussed in Section 4.3. Several additional considerations were taken into account during this imputation procedure.

The key additional consideration within this imputation procedure was accounting for the non-proportional hazards. The imputation models were specified using the methods outlined in Section 6.3, where the imputation models were specified to be approximately compatible with a piecewise-proportional hazards model as the analysis model.

This involved incorporating additional terms within the imputation models which account for the change in hazard at the 7 day time-point post-stroke. Section 6.3 showed that in order to incorporate this change in hazard, the imputation models needed to contain two terms for the cumulative baseline hazard; the cumulative baseline hazard up to and including 7 days post-stroke, and the other for after 7 days post-stroke. These were calculated using the Nelson-Aalen estimator. Additionally, the imputation models also needed to incorporate two censoring indicator terms, one for up to 7 days post-stroke, and the other identifying events after 7 days.

A further issue found in the model validation was that the BP measures did not have linear functional form, but instead would need to be included as quadratic terms in further analyses. This also means that these measures needed to be incorporated into the imputation procedure as quadratic terms. This resulted in two issues, how to include them in imputation models as predictors for the missing values of other incomplete covariates, and also how to impute the missing values of the BP measures themselves, given we need both the measure and the quadratic measure.

In terms of including the BP measures as predictors for the missing values of other incomplete covariates, the quadratic terms could simply be added as an additional term in the imputation models. For imputing the missing BP values, more consideration was needed, with several possible approaches, as outlined in

Section 6.3.6. As there was minimal missingness within these covariates with only 4 observations missing, a passive approach was taken where the quadratic BP measures were calculated using the square of the imputed values for the linear variables.

Model validation also highlighted an interaction between hospital and no CT scan. This interaction also needed to be incorporated into the imputation procedure. As the interaction was added to a single level in the lesion type shown in CT scan variable, this interaction was incorporated into the variable itself and thus no additional terms were needed. The key thing to note for this was that hospital needed to be included along side the CT scan variable, as a main effect in the imputation models, due to the interaction. There were no missing values in either of these covariates so this did not complicate the imputation procedure.

During this new imputation procedure accounting for additional issues, we also had to choose a cut off for the amount of missing observations to be imputed, and ensure the imputations were of the appropriate form for the variable type. As with the previous imputation procedure, 50% was chosen as the cut off for the amount of missing data to be imputed, to again avoid amplifying imperfections in the imputation procedure. Linear regression models were used for the imputation models of continuous covariates, binary variables were imputed using logistic regression models and multinomial logistic regression was used for both unordered and ordered categorical covariates.

The imputation models also needed to include all covariates to be included in the analysis model, and all covariates which are predictors of either the incomplete covariate or whether the observation is missing. As with the previous imputation procedure, an all for all approach was taken where every baseline covariate was included as a predictor in every incomplete covariate's imputation model.

The imputation procedure was again carried out using the `mice` package in R, where a prediction matrix was used to specify the variables to be included as predictors in the imputation model. Each imputation cycle was run for 1000 iterations to ensure convergence, and the number of imputations was also again chosen to be 10, to produce 10 imputed data sets.

In terms of the computational aspect of implementing the imputation procedure, it should be noted that the new methods applied here, as outlined in Chapter 6, have not been any more computationally intensive than the standard methods upon application. The methods can be easily implemented in the MICE package in R, and the only part that would result in higher computational intensity is the addition of extra predictive terms to account for the non-proportional hazards. In our case, this only caused the imputation procedure to run marginally slower.

For both types of imputation procedure, we found that the procedure took around 24-36 hours to run on a standard computer, and this was reduced to 6-12 hours when a high-speed computer cluster was used. Although this is reasonably computationally intensive, the key things which affect the running time are the number of imputations, the number of iterations and the number of variables with missing data. It is possible to run the imputation procedure on a standard computer, regardless of which procedure is used, though it would be preferable to use a high-speed cluster if that resource were available. Compared to the standard approach, no additional computing facilities are required to the run imputation procedure developed to handle non-proportional hazards.

8.4.2 Imputation Diagnostics

Imputation diagnostics were again carried out to assess several aspects of the imputation procedure. Strip plots and histograms were used to examine the distributions of the imputed values compared to the observed values, assessing the within and between imputation variability. Trace plots were again used to assess the convergence of the imputations.

Strips plots for the imputation of the systolic and diastolic BP missing values are shown in Figures 8.3 and 8.4, respectively. These show the distribution of the observed values and how the imputed values relate, and also show how the imputed values vary between the imputed data sets. For systolic BP, Figure 8.3 shows that imputed data set 4 has an extreme imputed value slightly below the range of the observed data, however all the remaining imputed data sets are shown to have imputed values within the range of the observed data. Figure 8.4 shows that the imputed values for diastolic BP are all within the range of the observed data, with reasonable consistency in the imputed values between the imputed data sets.

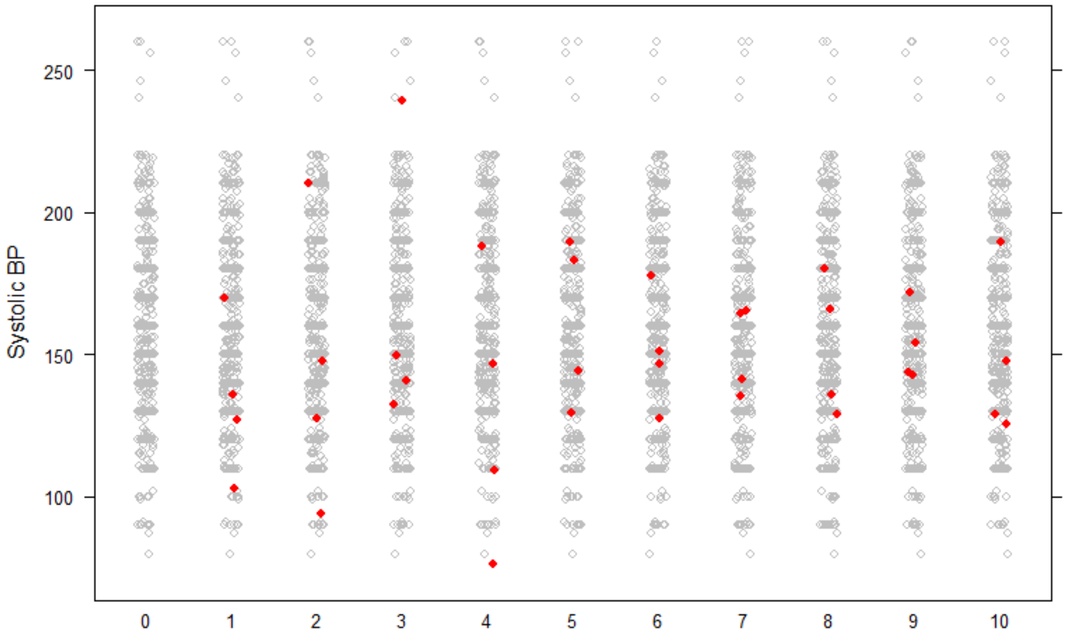


Figure 8.3: Strip plot showing the observed and imputed values of systolic BP, for both incomplete data, labelled 0, and the 10 imputed data sets, where imputed values are displayed as red points.

For categorical and binary covariates, the distribution of the imputations were explored using histograms. Histograms showing the imputed values for each imputed data set m for each variable were plotted to check between imputation variability. Figure 8.5 gives these histograms for pre-stroke mobility, chosen for inclusion since pre-stroke mobility was an incomplete covariate with a time-dependent covariate effect. Figure 8.5 shows that in general, similar proportions of patients

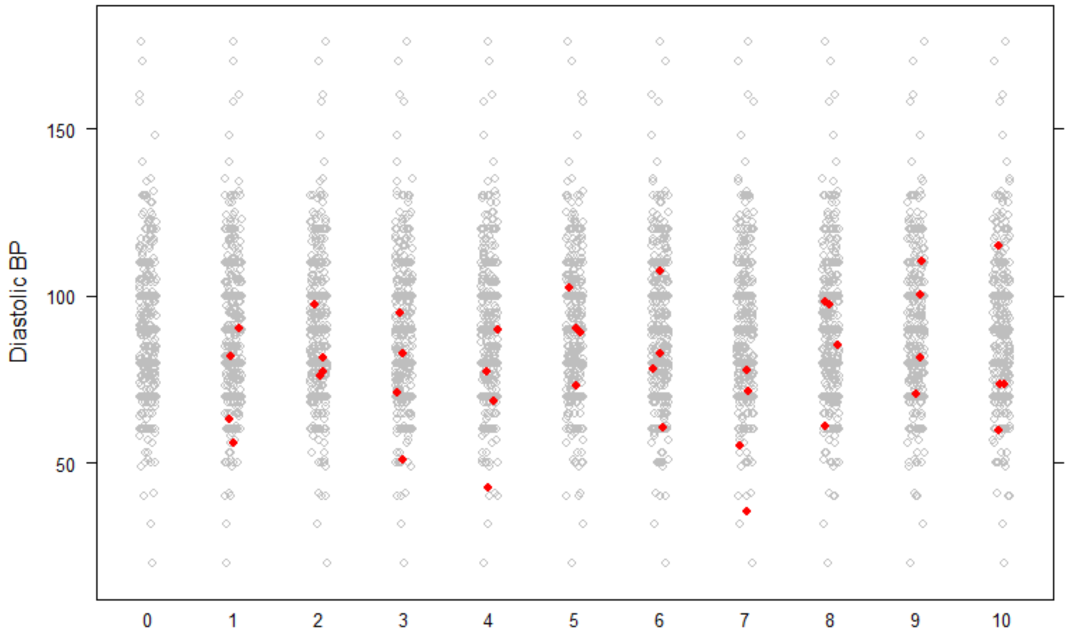


Figure 8.4: Strip plot showing the observed and imputed values of diastolic BP, for both incomplete data, labelled 0, and the 10 imputed data sets, where imputed values are displayed as red points.

were imputed into each of the levels, where every imputed data set had the highest proportion imputed into level 1, followed by level 3. Imputed data sets 4 and 8 had a higher proportion imputed into the middle level compared to other imputed data sets however, with the proportions more similar across the levels. The distribution of imputed values were generally found to be consistent across the imputed data sets for the other binary and categorical covariates.

Further histograms were plotted showing the overall distribution of the imputed values, against the distribution of the imputed data sets and observed data. Again using the imputation of pre-stroke mobility as an example, these histograms can be seen in Figure 8.6. The left histogram gives the imputed values only, the middle shows the proportions in each level over all the imputed data sets, and the right gives the distribution of pre-stroke mobility in the original observed data. Figure 8.6 shows that the distribution of the imputed values differs from the data as a whole, but the distributions of the overall imputed data and the observed data are near identical. Similar results were found for the majority of the covariates, where there were minimal differences in shape between the imputed and observed data.

Finally, trace plots were used to assess the convergence of the imputations and overall showed good convergence. Figure 8.7 shows the trace plots for the imputed values of pre-stroke mobility, smoking status and alcohol consumption in descending order. These show good convergence overall. The BP measures had the additional consideration of having quadratic functional form so we also present the trace plot for the BP measures, where Figure 8.8 gives the trace plot for quadratic systolic BP, and both linear and diastolic BP. Again, these trace plots show good convergence.

The imputation diagnostics as a whole showed that overall the imputations had good convergence, and in general the between imputation variance was not excessive. This suggests that the imputation procedure was satisfactory and the MAR assumption was reasonable to make.

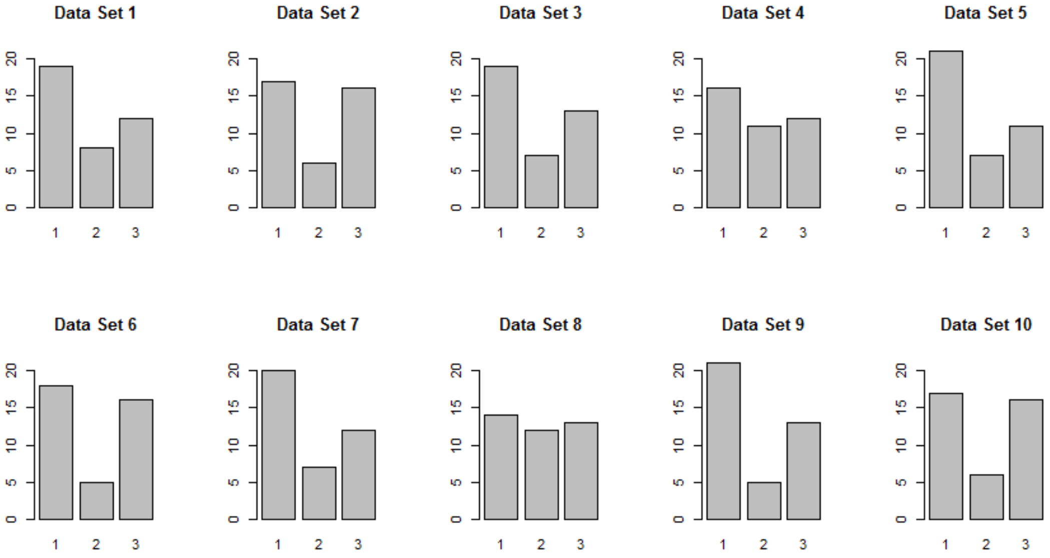


Figure 8.5: Histograms showing the distribution of the imputed values for pre-stroke mobility for each of the 10 imputed data sets.

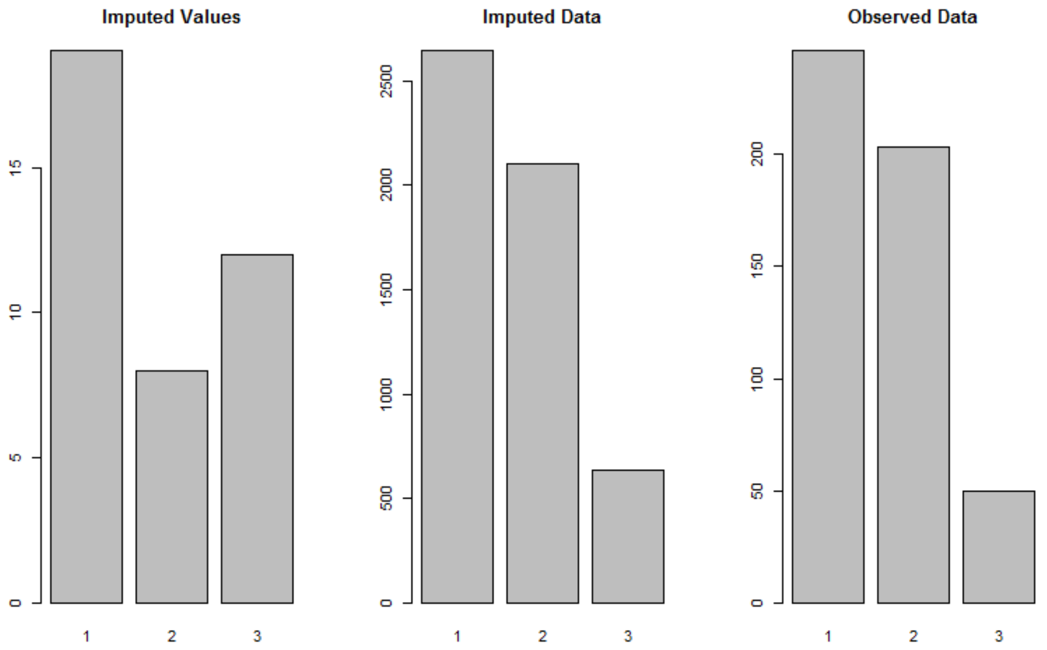


Figure 8.6: Histograms showing the overall distributions of the imputed values, imputed data and observed data for pre-stroke mobility.

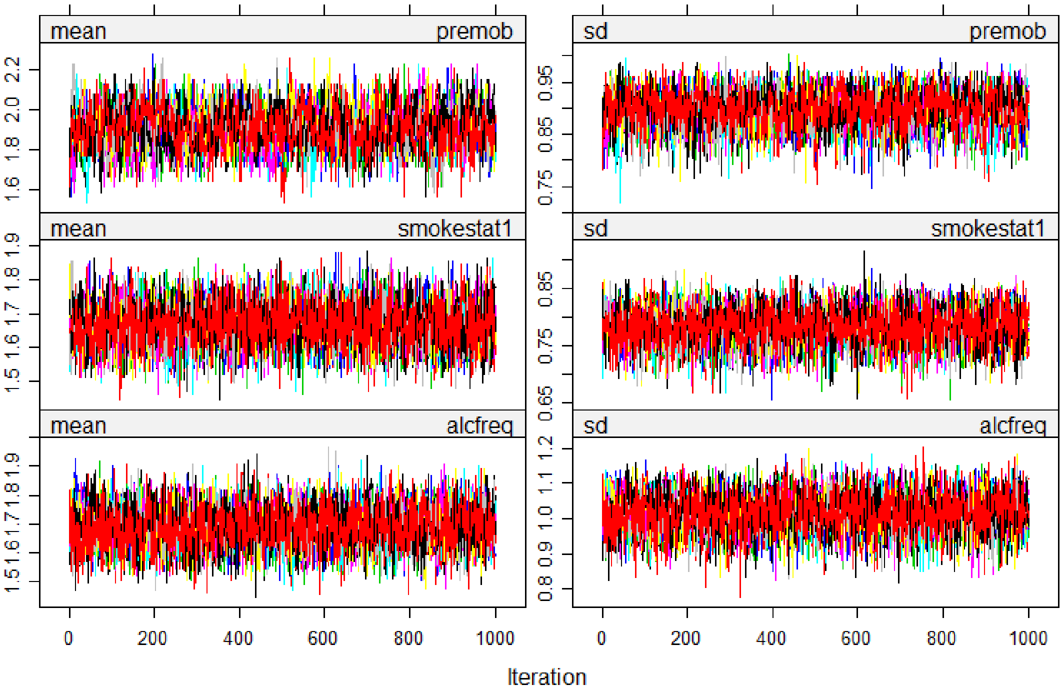


Figure 8.7: Trace plots for imputations of pre-stroke mobility, smoking status and alcohol consumption to assess their convergence.

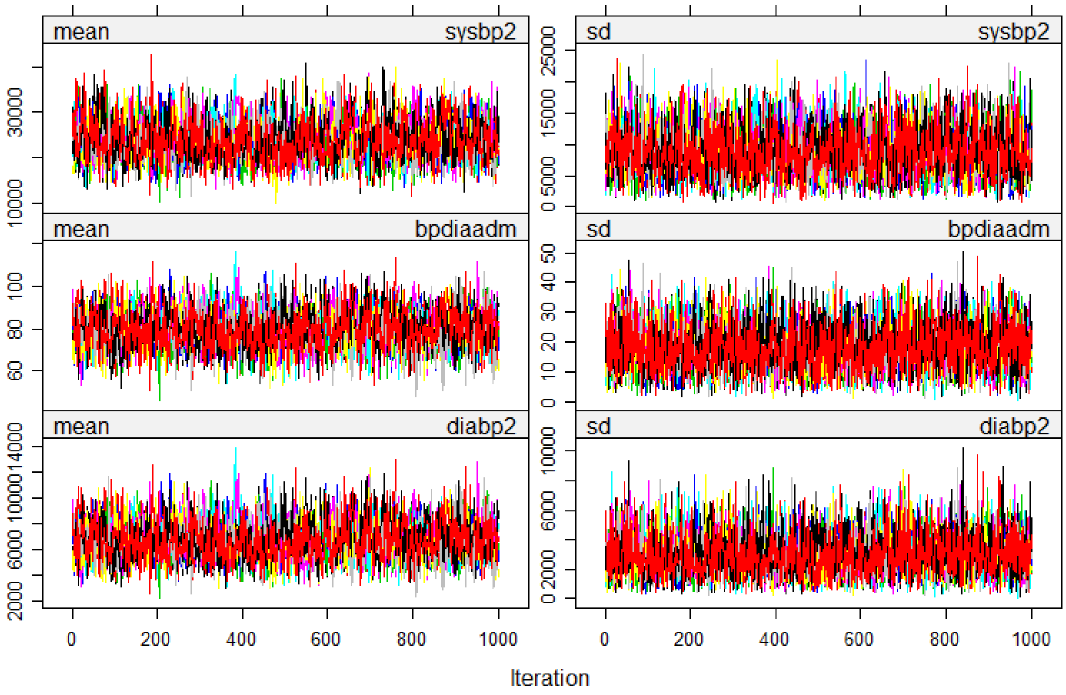


Figure 8.8: Trace plots for imputations of quadratic systolic BP, and linear and quadratic diastolic BP at hospital admission to assess their convergence.

8.5 Model Building

Following the imputation procedure accounting for a change in hazard due to time-dependent covariate effects, the analysis stage was again carried out by using backwards elimination and the Wald test for model selection. The fully adjusted piecewise-proportional hazards model was fitted to each of the imputed data sets, where time-dependent covariate effects were incorporated for age, pre-stroke mobility, worst consciousness level and lesion type shown in CT scan. The interaction between hospital and no CT scan meant during backwards elimination we had to ensure hospital was not excluded while the lesion type shown in CT scan variable remained in the model. Similarly, we had to ensure the linear terms for the BP measures stayed in the model whilst the quadratic terms remained as significant variables.

Implementation of the backwards elimination procedure involved fitting the piecewise-proportional hazards model to each of the imputed data sets, at each stage, to the new set of variables, before combining the estimates using Rubin's rules and excluding the least important covariate. The process was repeated until only variables significant at the 5% level remained, along side any main effects for interactions or quadratic terms.

The combined estimates of the final model from the model building procedure were used to calculate the hazard ratios, and their corresponding 95% confidence intervals and p-values. The results of the model building are presented in Tables 8.2 and 8.3, with a visualisation of the hazard ratios presented in Figure 8.9. For variables with different effects on survival depending on the time post-stroke, this change in effect occurs at 7 days post-stroke. The notation (T $\leq$ 7) and (T $>$ 7) refers to the effects which are dependent upon time, referring to the effect between 0 and 7 days post-stroke and the effect after 7 days, respectively.

Table 8.2 shows that between 0 and 7 days post-stroke, higher ages reduced hazard of death, however, from 7 days onwards post-stroke, older patients had an increased hazard of death. After 7 days, for every 10-year increase in age at time of stroke, there was a 64% increase in hazard of death. Patients with worse pre-stroke mobility had a significantly increased hazard of death after 7 days post-stroke; those needing help had the worst chance of survival with a 165% increase in risk of death compared to those able to walk 200 metres outdoors.

It is also shown in Table 8.2 that patients with diabetes mellitus had a 44% higher hazard of death. Systolic BP has a quadratic form, meaning patients with extremely low or high systolic BP values had the highest hazard of death, where a systolic BP value of 181mmHg gave the lowest estimated hazard.

For worst consciousness level within the first 24 hours post-stroke, any reduction in consciousness level gave an increased hazard of death. Table 8.2 and Figure 8.9 show that the hazard ratios were highest within the first week post-stroke, with hazard ratios 3.6, 6.6 and 7.5 times higher for drowsy, stupor and coma patients

Table 8.2: Results of the piecewise-proportional modelling procedure on the multiply imputed data, showing the pooled estimates of the hazard ratios (HR), along with the corresponding 95% confidence intervals (CI) and *p*-values for: age, pre-stroke mobility, diabetes, systolic BP, and worst consciousness level.

Variable	**HR**	**95% CI**	***p*-value**
Age (T $\leq$ 7) (10years)	0.739	(0.610,0.895)	0.002
Age (T $>$ 7) (10years)	1.636	(1.419,1.886)	<0.0001
Pre-stroke Mobility			
200m Outdoors	*1.00*		
Indoors (T $\leq$ 7)	1.089	(0.636,1.866)	0.755
Indoors (T $>$ 7)	1.517	(1.130,2.038)	0.006
Needs Help (T $\leq$ 7)	0.526	(0.272,1.014)	0.055
Needs Help (T $>$ 7)	2.651	(1.591,4.417)	0.0002
Diabetes Mellitus			
No	*1.00*		
Yes	1.436	(1.035,1.993)	0.03
Systolic BP			
Linear (10mmHg)	0.673	(0.517,0.876)	0.003
Quadratic (100mmHg^2)	1.011	(1.002,1.019)	0.011
Worst Conscious Level			
Alert	*1.00*		
Drowsy (T $\leq$ 7)	3.661	(1.820,7.362)	0.0003
Drowsy (T $>$ 7)	1.829	(1.256,2.662)	0.002
Stupor (T $\leq$ 7)	6.622	(3.399,12.902)	<0.0001
Stupor (T $>$ 7)	1.454	(0.828,2.554)	0.193
Coma (T $\leq$ 7)	7.502	(4.471,12.590)	<0.0001
Coma (T $>$ 7)	3.127	(2.030,4.819)	<0.0001

respectively compared to alert patients. The hazard ratios reduced after 7 days giving an 83% increase in hazard of death for drowsy patients, a 45% increase for stupor and a hazard ratio of 3.1 for unconscious patients.

No lesion was specified as the baseline for the CT scan results, so any change in hazard of death is compared to patients with no lesion showing on the CT scan. Considering the CT scan results, Table 8.3 and Figure 8.9 show that patients diagnosed with a primary intracerebral haemorrhage (PICH) or haemorrhagic infarction (HCI) had an increased risk of death. The hazard ratios were higher within the first week post stroke, reducing from triple to double for HCI after 7 days, and PICH had a hazard ratio of 3.5 initially, reducing to a 43% increase in risk of death after 7 days compared to no lesion. Conversely, cerebral infarction (CI) had a lower hazard ratio in the first week post-stroke. Initially, CI corresponded to a 20% decrease in risk of death compared to patients with no lesion, however post 7 days this increased to a 35% increase in hazard of death.

Figure 8.9 highlights that, of all the CT scan results, patients that did not have a CT scan had the highest hazard ratios overall, and the hazard of death was higher for patients at Hospital 1 compared to Hospital 2. No scan at Hospital 1 resulted in a hazard ratio of 10.5 initially, but this reduced after 7 days to doubling the risk of death compared to no lesion. The estimated hazard ratio for no scan at Hospital 2 in the first 7 days post-stroke was 6.3, but again after 7 days this reduced to give a 66% increase in risk of death compared to patients with no lesion.

The results overall suggest that when adjusted for other baseline covariates, the patients most at risk of death in the first week post-stroke are those which were stupor or comatose in the first 24 hours post-stroke, and those which did not receive a CT scan. For patients that have survived beyond 7 days post-stroke, the hazard of these patients is reduced going forward, however they are still at a higher risk of death compared to the baseline. After 7 days post-stroke, increased age and needing help in terms of pre-stroke mobility now also give some of the largest hazards for death.

Table 8.3: Results of the piecewise-proportional modelling procedure on the multiply imputed data, showing the pooled estimates of the hazard ratios (HR), along with the corresponding 95% confidence intervals (CI) and *p*-values for: hospital and lesion type shown in CT scan.

Variable	**HR**	**95% CI**	***p*-value**
Hospital			
Hospital 1	*1.00*		
Hospital 2	0.836	(0.619,1.129)	0.242
CT Scan: Lesion Type			
No Lesion	*1.00*		
CI ($T \leq 7$)	0.797	(0.294,2.157)	0.655
CI ($T > 7$)	1.352	(0.961,1.902)	0.083
HCI ($T \leq 7$)	3.380	(0.679,16.824)	0.137
HCI ($T > 7$)	2.041	(1.043,3.993)	0.037
PICH ($T \leq 7$)	3.458	(1.394,8.577)	0.007
PICH ($T > 7$)	1.428	(0.850,2.397)	0.178
No Scan @ Centre A ($T \leq 7$)	10.461	(4.282,25.553)	<0.0001
No Scan @ Centre A ($T > 7$)	2.075	(1.063,4.051)	0.032
No Scan @ Centre B ($T \leq 7$)	6.349	(2.583,15.602)	0.0001
No Scan @ Centre B ($T > 7$)	1.663	(1.077,2.568)	0.022

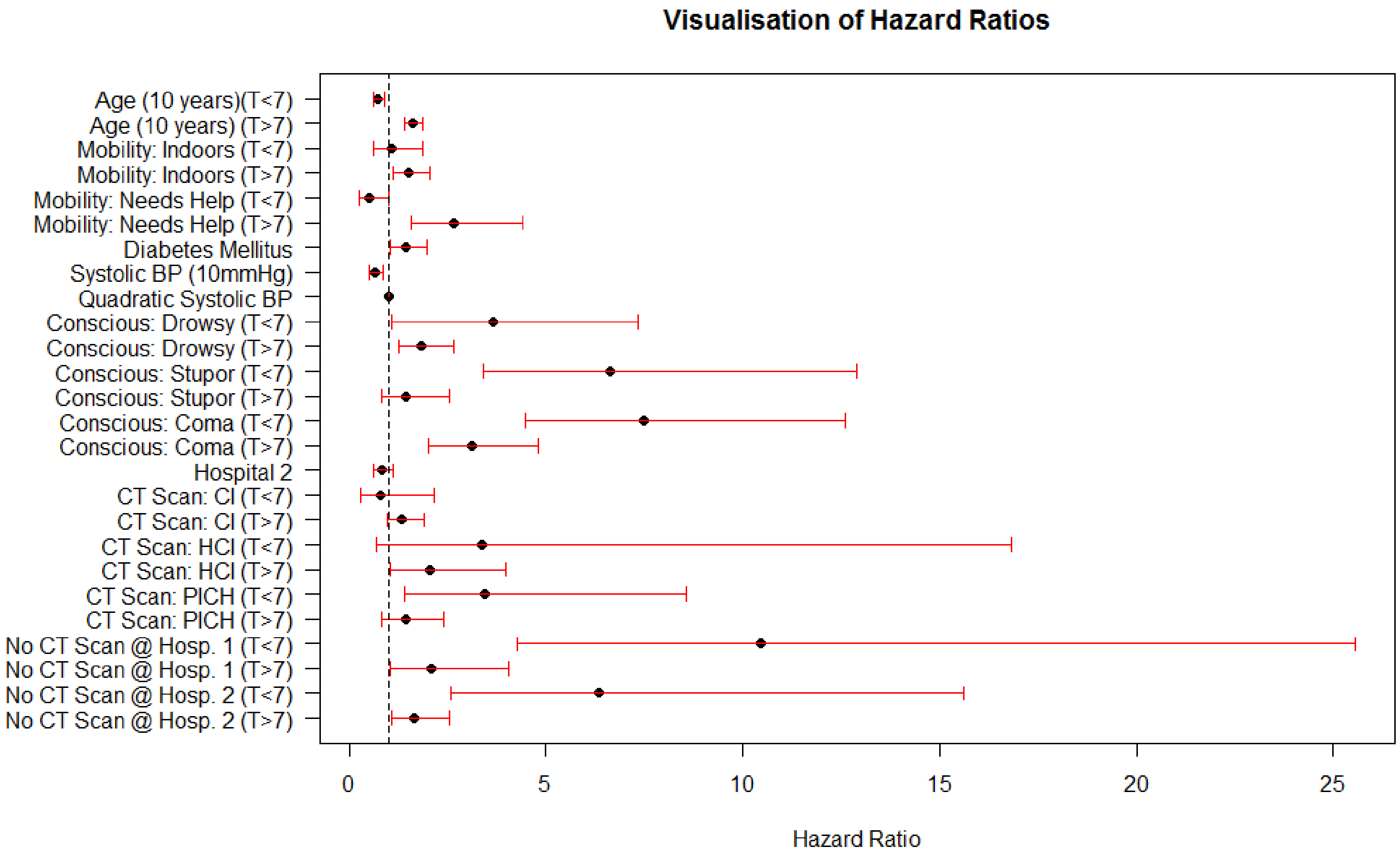


Figure 8.9: Visualisation of the hazard ratios and corresponding confidence intervals for each of the variables in the piecewise-proportional hazards model.

8.6 Model Validation

The key focus of model validation for the piecewise-proportional hazards was to check that the proportional hazards assumption was now satisfied after application of methods to handle the initial violation of this assumption. In order to assess the proportional hazards assumption in a piecewise-proportional hazards model, the methods outlined in Chapter 7 were applied. For variables with time-dependent covariate effects, the proportional hazards assumption of the two coefficients were assessed within each time interval separately. For variables with a constant coefficient over the full follow-up time, the proportional hazards assumption was assessed across the whole follow-up time.

The scaled Schoenfeld residuals from the model fitted to each imputed data set were regressed against the centered Kaplan-Meier function of time, using the regression models defined in Section 7.5, and where the residuals were scaled using the techniques outlined in Section 7.4.1.

Table 8.4 gives the combined results of the linear regression modelling of the residuals against time, where the coefficients and variance estimates were combined using Rubin's rules, and the Wald test was used to assess if the combined coefficients were significantly non-zero. Table 8.4 gives the coefficients and corresponding p-values of these regression models.

From the results in Table 8.4, it is clear the use of the piecewise-proportional hazards models to include time-dependent covariate effects has resolved the issue around violation of the proportional hazards assumption, where the p-values are all larger than 0.05. These results show that the time-dependent coefficients are constant and do not violate the proportional hazards assumption within their corresponding time-intervals. Variables fitted in the piecewise-proportional hazards model with time-constant coefficients also satisfy the proportional hazards assumption.

As an additional assessment of the proportional hazards assumption, the scaled

Schoenfeld residuals can be plotted against time to visualise the nature and extent of any non-proportional hazards. Given the change in effect of some covariates on survival, the scaled Schoenfeld residuals were plotted separately for each time interval for those covariates with time-dependent effects, resulting in two plots for each level of these covariates. For covariates assumed to have constant effect over time, a single plot of the residuals were produced for the full follow-up period.

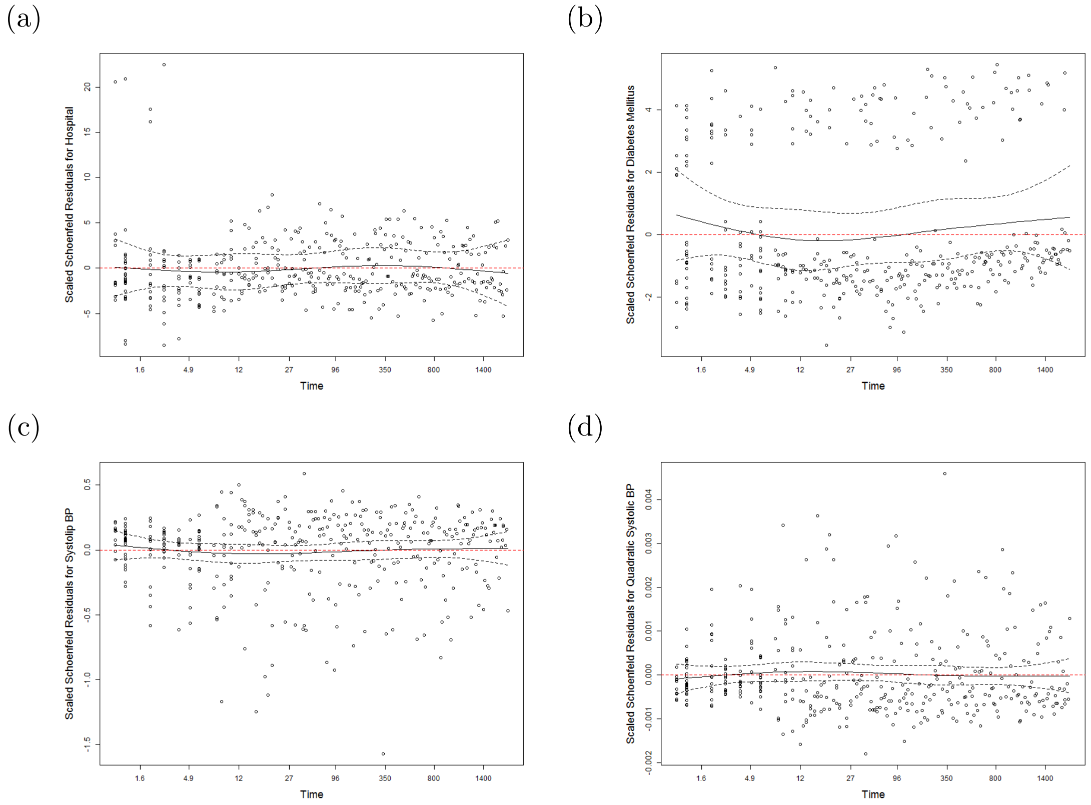


Figure 8.10: Combined plots of scaled Schoenfeld residuals against time (days) with smooth curve to visualise extent of violation of proportional hazards for variables with a constant effect over time: (a) Hospital; (b) Diabetes; (c) Systolic BP; (d) Quadratic Systolic BP.

Figure 8.10 presents the plots for the covariates with a constant effect over the full follow-up period. Unlike the `cox.zph` function, the residuals scaled by the methods in Section 7.4.1 are not transformed by the coefficient, and thus to interpret these plots, the smooth can be compared to the zero line to check for non-proportional hazards. Figure 8.10 clearly shows that the smooths on each of these plots are close to the zero line with the zero line within the confidence intervals of the smooth. Figure 8.10(b) has a wide confidence interval, but this is likely due to the large amount of missing data that has been imputed for diabetes and any between-imputation variance due to this.

The remaining plots of the scaled Schoenfeld residuals, given in Figures 8.11, 8.12, 8.13, 8.14, and 8.15, give separate plots for each time interval to examine the extent of non-proportional hazards for each coefficient separately.

Firstly, the residuals for age are given in Figure 8.11, where Figure 8.11(a) shows a slight issue around the end of the first time-period. However, given the zero line is still within the confidence interval, and the p-value in Table 8.4 is greater than 0.05, it can be concluded that the proportional hazards assumption has not been violated significantly. Figure 8.11(b) indicates that the effect of age after 7 days post-stroke also does not violate the proportional hazards assumption.

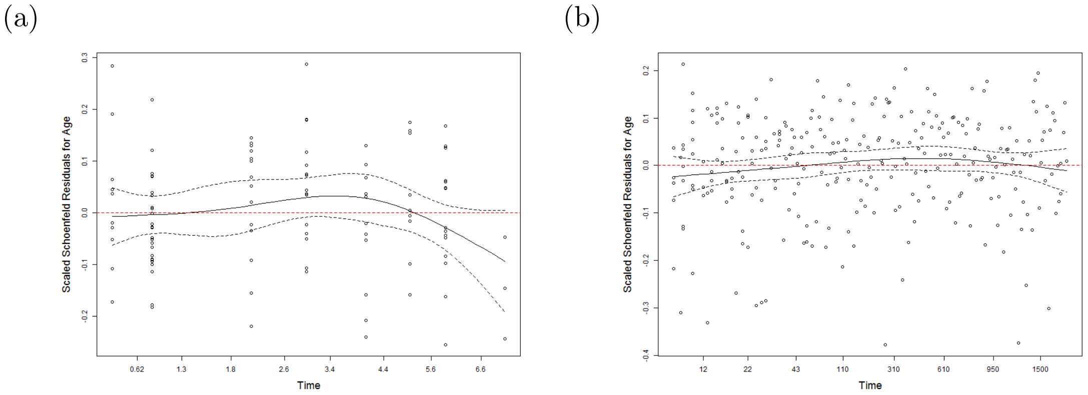


Figure 8.11: Combined plots of scaled Schoenfeld residuals against time (days) with smooth curve to visualise extent of violation of proportional hazards for each of the time periods for Age: (a) T $\leq$ 7; (b) T $>$ 7.

Considering the effects of pre-stroke mobility, Figure 8.12 shows that the proportional hazards is satisfied for the effects of both levels within each of the time intervals, where the smooth fits are close to the zero line, with the zero line within the confidence interval of the smooth on each of these plots.

Figure 8.13 and 8.14 give the plots of the scaled Schoenfeld residuals for the effects of lesion types shown in CT scan. The results for PICH in Figure 8.13(e) are similar to that of age before 7 days post-stroke, however, again the zero line remains within the confidence interval of the smooth, and the formal test results also indicated this possible violation is not significant. The remaining plots in Figure 8.13 show smooth curves all close to the zero line with minimum curvature.

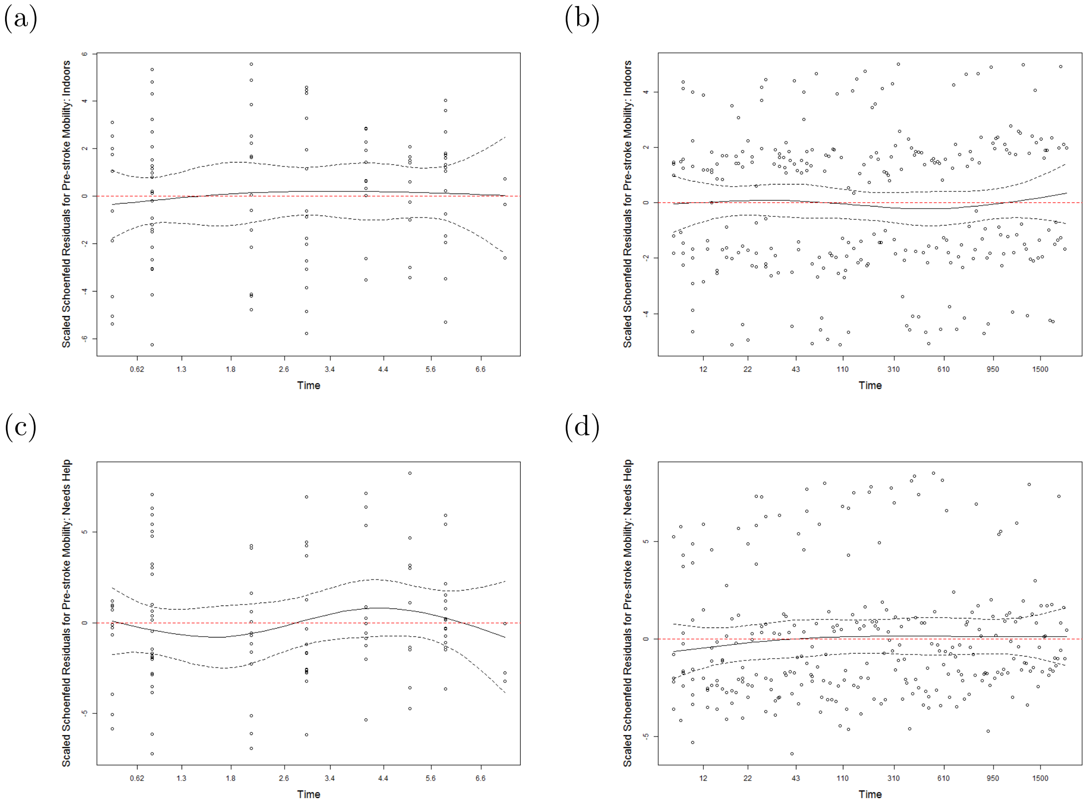


Figure 8.12: Combined plots of scaled Schoenfeld residuals against time (days) with smooth curve to visualise extent of violation of proportional hazards for variables with a constant effect over time for each of the time period for Pre-stroke Mobility: (a) Indoors (T $\leq$ 7); (b) Indoors (T $>$ 7); (c) Needs Help (T $\leq$ 7); (d) Needs Help (T $>$ 7).

Looking at the residual plots for no CT scan at the two hospitals, Figure 8.14 reiterates the test results in Table 8.4, indicating the proportional hazards assumption is not violated by the effect of no scan, interacted with hospital, for both time intervals of before and after 7 days post-stroke.

Finally, considering the effects of the levels of worst consciousness level, Figure 8.15 shows that there is some non-linearity in the effect prior to 7 days post-stroke, see Figures 8.15(a), 8.15(c) and 8.15(e), however the zero line is again contained within the confidence interval of the smooths and the test results in Table 8.4

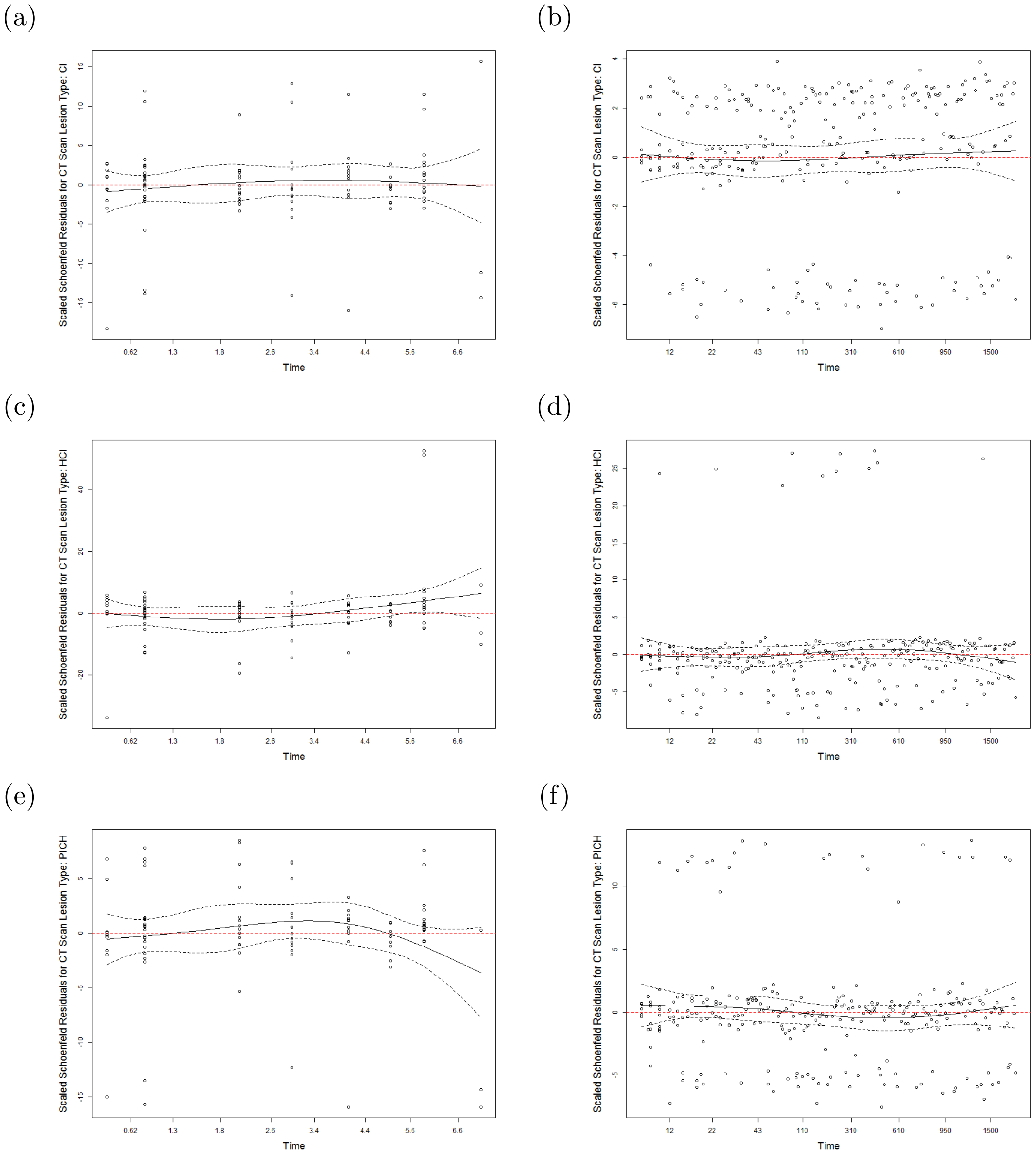


Figure 8.13: Combined plots of scaled Schoenfeld residuals against time (days) with smooth curve to visualise extent of violation of proportional hazards for variables with a constant effect over time for each of the time period for Lesion Type shown in CT Scan: (a) CI (T $\leq$ 7); (b) CI (T $>$ 7); (c) HCI (T $\leq$ 7); (d) HCI (T $>$ 7); (e) PICH (T $\leq$ 7); (f) PICH (T $>$ 7).

showed that the proportional hazards assumption was not significantly violated for these effects. For the effects of consciousness after 7 days post-stroke, Figures 8.15(b) and 8.15(f) show constant effects, however there is some curvature of the smooth in Figure 8.15(d) for the effect of stupor. Given the test results in Figure 8.4, and the zero line being contained within the confidence interval of the smooth in Figure 8.15(d), however, this does not suggest a significant violation of the proportional hazards assumption.

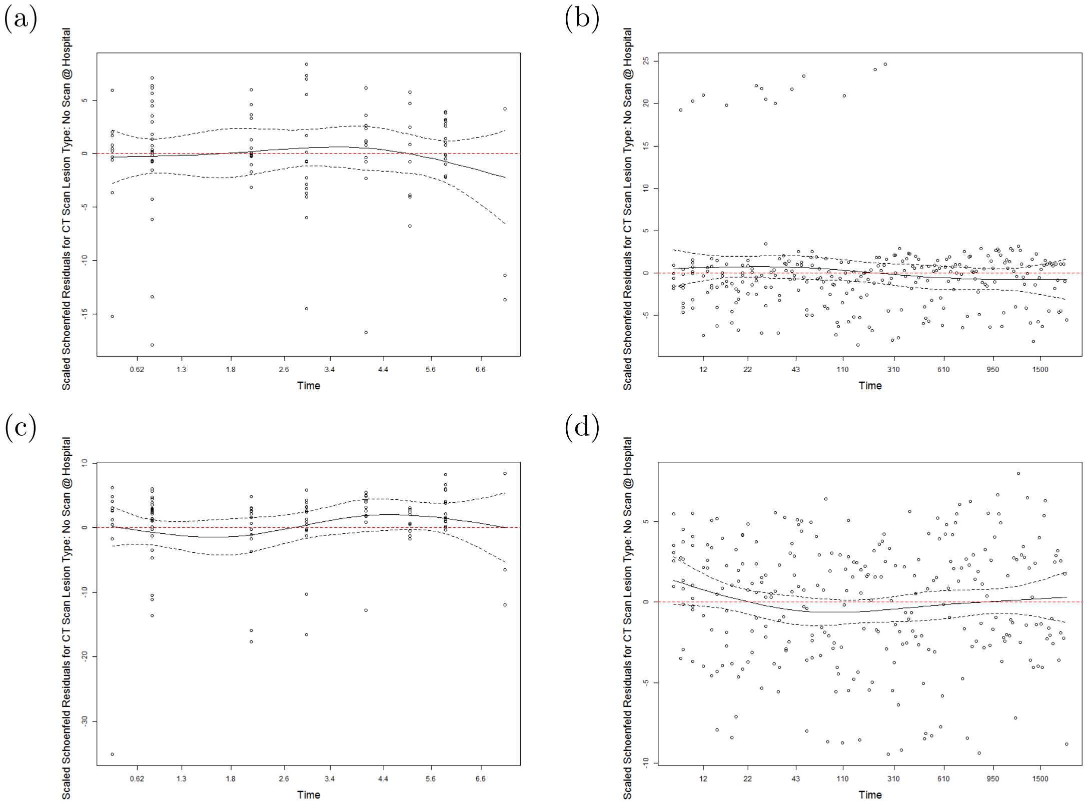


Figure 8.14: Combined plots of scaled Schoenfeld residuals against time (days) with smooth curve to visualise extent of violation of proportional hazards for variables with a constant effect over time for each of the time period for Lesion Type shown in CT Scan: (a) No Scan at Hospital 1 (T $\leq$ 7); (b) No Scan at Hospital 1 (T $>$ 7); (c) No Scan at Hospital 2 (T $\leq$ 7); (d) No Scan at Hospital 2 (T $>$ 7).

Overall, the plots of the scaled Schoenfeld residuals with smoothed curves, and the results of the residuals regressed against time, indicate that the proportional hazards has not been violated by any of the covariate effects within the piecewise-proportional hazards model, so we can conclude that the model fit is adequate in terms of the proportional hazards assumption.

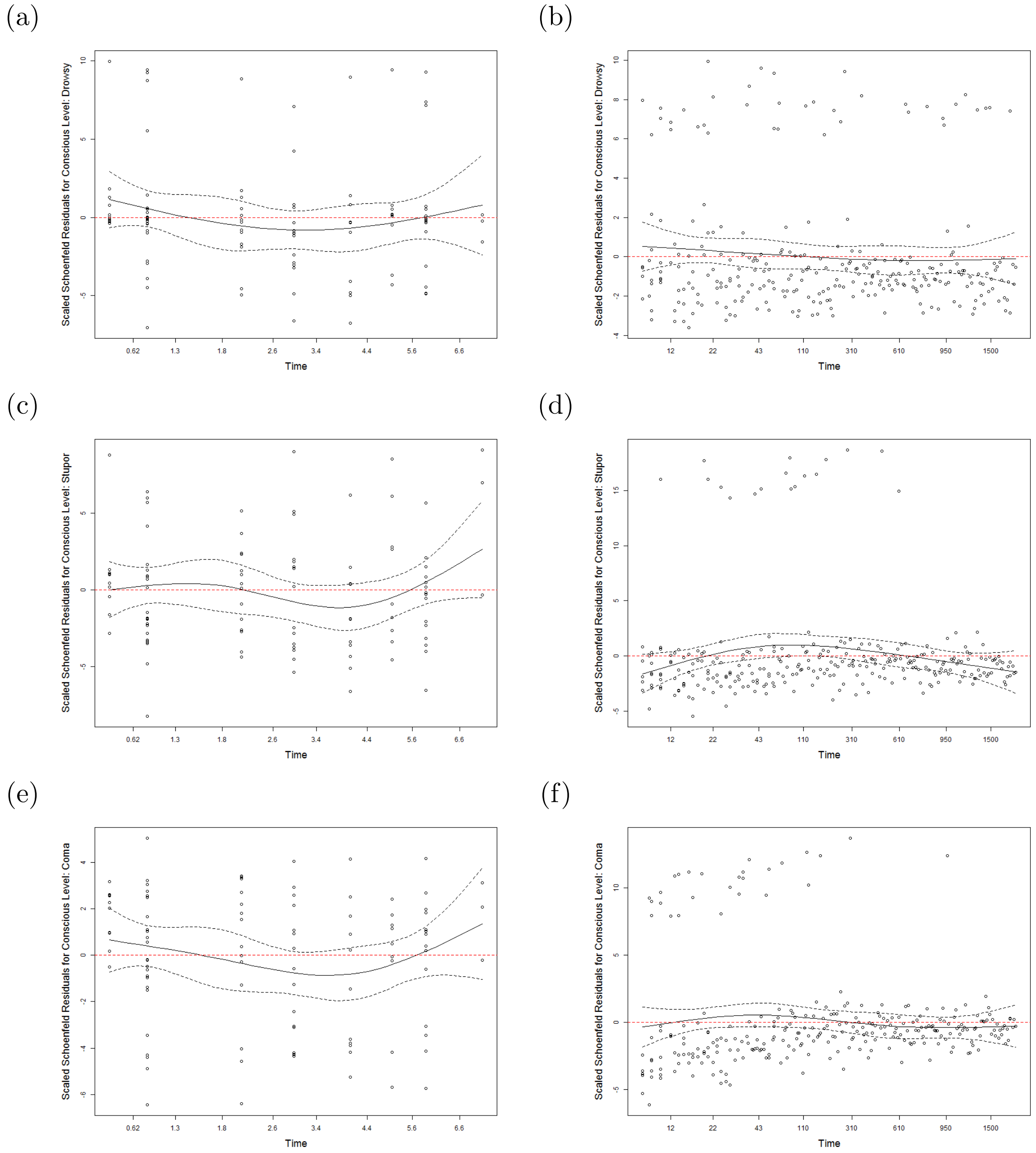


Figure 8.15: Combined plots of scaled Schoenfeld residuals against time (days) with smooth curve to visualise extent of violation of proportional hazards for variables with a constant effect over time for each of the time period for Worst Consciousness Level: (a) Drowsy ($T \leq 7$); (b) Drowsy ($T > 7$); (c) Stupor ($T \leq 7$); (d) Stupor ($T > 7$); (e) Coma ($T \leq 7$); (f) Coma ($T > 7$).

Table 8.4: Results of the formal test of the proportional hazards assumption, showing the pooled coefficient estimates of the scaled Schoenfeld residuals regressed against time, and their corresponding p-values, for each covariate effect in the adjusted piecewise-proportional hazards model.

Variable	**Coefficient**	**p-value**
Age (T $\leq$ 7) (10years)	-0.010	0.886
Age (T $>$ 7) (10years)	0.002	0.886
Hospital (Hospital 1)		
Hospital 2	0.004	0.984
Pre-stroke Mobility (200m Outdoors)		
Indoors (T $\leq$ 7)	0.263	0.879
Indoors (T $>$ 7)	-0.001	0.997
Needs Help (T $\leq$ 7)	0.429	0.850
Needs Help (T $>$ 7)	0.042	0.896
Diabetes Mellitus (No)		
Yes	0.006	0.953
Systolic BP (10mmHg)	-0.0001	0.986
Quadratic Systolic BP (100mmHg2)	<0.0001	0.998
CT Scan: Lesion Type (No Lesion)		
CI (T $\leq$ 7)	0.633	0.857
CI (T $>$ 7)	0.019	0.933
HCI (T $\leq$ 7)	2.986	0.767
HCI (T $>$ 7)	0.019	0.964
PICH (T $\leq$ 7)	-0.402	0.889
PICH (T $>$ 7)	-0.045	0.907
No Scan @ Centre A (T $\leq$ 7)	-0.198	0.944
No Scan @ Centre A (T $>$ 7)	-0.131	0.866
No Scan @ Centre B (T $\leq$ 7)	1.404	0.791
No Scan @ Centre B (T $>$ 7)	-0.029	0.926
Worst Conscious Level (Alert)		
Drowsy (T $\leq$ 7)	-0.591	0.825
Drowsy (T $>$ 7)	-0.049	0.885
Stupor (T $\leq$ 7)	0.018	0.993
Stupor (T $>$ 7)	-0.030	0.933
Coma (T $\leq$ 7)	-0.340	0.852
Coma (T $>$ 7)	-0.046	0.896

8.7 Conclusion

This chapter has provided further analyses of the stroke audit data, providing application of the methodological developments made within Chapters 6 and 7. The analyses within this chapter addressed the issues found in the previous analyses in Chapter 4 through incorporating interaction terms, quadratic terms and identifying the time-point at which the change in effect of variables occurred.

The methods outlined in Chapter 6 were applied to the stroke audit data in order to account for the non-proportional hazards, giving a less restrictive approach to predicting the missing values, and ensuring the imputation models were approximately compatible to the piecewise-proportional hazards model.

Following the multiple imputation procedure, the model building procedure involved backwards selection using the Wald test to obtain a parsimonious pooled piecewise-proportional hazards model, identifying an important set of risk factors for survival post-stroke. The model was validated through application of the methods outlined in Chapter 7, where it was concluded the covariate effects within the model did not violate the proportional hazards assumption.

In the context of stroke, the analyses in this chapter identified age, pre-stroke mobility, worst consciousness level, and lesion type shown in CT scan to be important for survival post-stroke in the adjusted setting, where 7-days post-stroke was found to be the changepoint for the effects of these risk factors. Overall older patients were shown to have better survival initially but were at increased risk of death after a week had passed following stroke. Generally worsening consciousness increased hazard of death, and the effect of this was more extreme within the first 7-days post-stroke. Similarly, compared to no lesion, presence of a lesion or not having a CT scan increased risk of death, particularly within the first week post-stroke, with the risk of no scan related to hospital admitted to.

On the other hand, pre-stroke mobility was more important after 7-days, where patients with worse mobility prior to stroke had increased hazard of death. Further, diabetes was found to be important in the adjusted model, where diabetic patients had increased hazard of death, and systolic BP at hospital admission was also shown to be important for survival post-stroke, with more extreme BP values, low or high, resulting in increased risk of death.

This analysis has demonstrated the application of the methodological developments made within Chapters 6 and 7, whilst providing insight in the context of stroke, to aid in a greater understanding of risk factors for survival post-stroke, and help to identify which patients are most at risk, in both the acute phase and for long-term survival.

Chapter 9

Conclusion

This thesis aimed to gain a greater understanding of the differences between stroke patients in terms of survival by developing methodology for handling missing data in survival data with non-proportional hazards, in order to identify important risk factors for time to death post-stroke.

Using the stroke audit data as our motivating data set, the thesis presented an overview of current methods for survival analysis and handling missing data, conducive to conducting an initial analysis to identify the key areas where methodological advancements were needed.

Following this, in Chapter 5, we derived and justified techniques for assessing the proportional hazards assumption of a pooled Cox regression model fitted to multiply imputed data. We presented a formal test and a visualisation technique, where simulation results suggested the test for complete covariate data was comparable to standard practice, but is notably more convenient to implement for a model fitted to multiply imputed data.

The finding of non-proportional hazards motivated the remainder of this thesis, where accounting for non-proportional hazards within an imputation framework was a key focus of this; failure to do so can introduce extra bias into inferences. We extended upon the imputation framework outlined by White and Royston (2009) to develop suitable forms of imputation models for multiple covariate types. Chapter 6 presented the derivation of these imputation model forms, where the theoretical rationale was further reinforced by favourable simulation results.

In depth consideration of the piecewise-proportional hazards model resulted in the detection of issues around assessment of the proportional hazards assumption for this model, where the current recommendations by Therneau et al. (2019) ignore how the disjoint nature of the model can impact the test. Within Chapter 7, we addressed these issues and developed an alternative approach to scaling the Schoenfeld residuals. Under this alternative approach, we presented adaptations of the formal test and visualisation technique to provide suitable methods of assessing the proportional hazards assumption within a survival model with a time-split.

The methodological advancements achieved within this thesis have fulfilled the aims to develop methodology for the multiple imputation of survival data in the presence of missing data, and for assessment of the proportional hazards assumption for survival models, both after multiple imputation, and for a model with a time-split. This has enabled the clinical aim of this thesis to be achieved through application of these methods.

In the context of stroke, we have identified several important risk factors for time to death post-stroke, including age, pre-stroke mobility, worst consciousness level, diabetes mellitus, systolic BP, and lesion type shown in CT scan, where hospital affected the risk relating to no scan. These results have aided in understanding differences between stroke patients in terms of survival, and can enable predictions to be made regarding which patients are most at risk of death following a stroke, both in the acute phase and long-term.

9.1 Further Work

Finally to conclude this thesis, we acknowledge areas of work where the achievements of this thesis could be developed further. Given the opportunity, in terms of the applied context of this thesis, it would be of interest to see the methods developed in this thesis applied to a more recent data set regarding survival of stroke patients.

Methodologically, this work has developed an imputation framework for handling non-proportional hazards, where the time-dependent covariate effects are incorporated through using a model with a time-split and piecewise-constant coefficients. This could be extended to produce a more general imputation framework which considers more general approaches of incorporating time-dependent covariate effects into the analysis model, for example, a linear function of time interacted with the covariate effect. Keogh and Morris (2018) provided methodology for a very general imputation model form, but there is still scope for developments to be made regarding more specific functions of time-dependent effects.

Taking an alternative approach to multiple imputation could be of interest, as opposed to using MICE. For example, the algorithm developed by Bartlett et al. (2015) to be compatible with the substantive model. It would be interesting to see whether the inflation of the Type 1 error of the proportional hazards test in Section 5.4 would be smaller if an imputation model compatible with the survival model were used.

It should also be noted that the stroke audit data set contains longitudinal data on functionality and mood following stroke. Considering this, a possible extension of this work would therefore be to consider if these longitudinal measures should be accounted for within the imputation framework for predicting the missing baseline covariate values.

Additionally, the presence of this longitudinal data provides a possible further extension of this work for the joint modelling of the baseline and longitudinal measures for predicting the survival of stroke patients. This would require consideration of how both the baseline and longitudinal measures could be incorporated into an imputation framework, where missing data is present in both.

Finally, an additional development of this work would be around the diagnostics of survival models following multiple imputation, where this thesis has focussed on assessing the proportional hazards model, but extensions could be made to further assessments of model fit. In particular, considering the functional form of covariates, Lin et al. (1993) suggested the use of cumulative sums of the martingale residuals for assessing the adequacy of fit of a Cox regression model; giving a formal test of functional form. This could be expanded upon to produce a formal test for a pooled model fitted to multiple imputed data.

Bernhardt (2018) made an attempt to introduce model validation techniques for multiply imputed data, however some of their suggestions were somewhat questionable, in particular combining p-values by taking the average. Marshall et al. (2009) present guidelines for combining estimates over multiply imputed data, considering different types of estimates with varying underlying distributions and appropriate methods for combining each. There is scope for the application and development of this with regards to model validation techniques of survival models.

References

Aalen, O. (1978). Nonparametric inference for a family of counting processes. *The Annals of Statistics*, 6(4):701–726.

Aho, K., Harmsen, P., Hatano, S., Marquardsen, J., Smirnov, V. E., and Strasser, T. (1980). Cerebrovascular disease in the community: results of a WHO collaborative study. *Bulletin of the World Health Organization*, 58(1):113.

Allison, P. D. (2001). *Missing Data*, volume 136. Sage publications.

Andersen, K. K. and Olsen, T. S. (2011). One-month to 10-year survival in the Copenhagen stroke study: interactions between stroke severity and other prognostic indicators. *Journal of Stroke and Cerebrovascular Diseases*, 20(2):117–123.

Anderson, J. A. and Senthilselvan, A. (1982). A two-step regression model for hazard function. *Applied Statistics*, 31:44–51.

Andrews, R., Paley, B., Bhasi, C., Hoffman, A., Kavanagh, M., Vestesson, E., et al. (2016). SSNAP 2016 Acute Organisational Audit Report. `https://www.strokeaudit.org/Documents/National/AcuteOrg/2016/2016-AOANationalReport.aspx`. Accessed: December 3, 2018.

Appelros, P., Nydevik, I., and Viitanen, M. (2003). Poor outcome after first-ever stroke: predictors for death, dependency, and recurrent stroke within the first year. *Stroke*, 34(1):122–126.

Azur, M. J., Stuart, E. A., Frangakis, C., and Leaf, P. J. (2011). Multiple imputation by chained equations: What is it and how does it work? *Int J Methods in Psychiatr Res.*, 20:40–49.

Barnard, J. and Rubin, D. B. (1999). Small-sample degrees of freedom with multiple imputation. *Biometrika*, 86(4):948–955.

Bartlett, J. W., Seaman, S. R., White, I. R., Carpenter, J. R., and Alzheimer's Disease Neuroimaging Initiative (2015). Multiple imputation of covariates by fully conditional specification: accommodating the substantive model. *Statistical methods in medical research*, 24(4):462–487.

Bernhardt, P. W. (2018). Model validation and influence diagnostics for regression models with missing covariates. *Statistics in medicine*, 37(8):1325–1342.

Bodner, T. E. (2008). What improves with increased missing data imputations? *Structural Equation Modeling: A Multidisciplinary Journal*, 15(4):651–675.

Bray, B. D., Clid, G. C., James, M. A., Hemingway, H., Paley, L., Stewart, K., et al. (2016). Weekly variation in health care quality by day and time of admission: a nationwide, registry-based, prospective cohort study of acute stroke care. *The Lancet*, 388:170–177.

Breslow, N. (1974). Covariance analysis of censored survival data. *Biometrics*, 30(1):89–99.

Breslow, N. (1984). Methods for identifying mortality risk in longitudinal studies. In Vallin, J., Pollard, H. J., and Heligman, L., editors, *Methodologies for the Collection and Analysis of Mortality Data*, pages 367–391. Ordina Editions.

Carpenter, J. and Bartlett, J. (2016). Missing data. `www.missingdata.org`.

Carpenter, J. R. and Kenward, M. G. (2013). *Missing Data and its Application.* Statistics in Practice. John Wiley & Sons, Ltd, 1st edition.

Clark, T. G., Bradburn, M. J., Love, S. B., and Altman, D. G. (2003). Survival Analysis Part IV: Further concepts and methods in survival analysis. *British journal of cancer*, 89(5):781–786.

Collett, D. (2015). *Modelling Survival Data in Medical Research*. Texts in Statistical Science. Chapman and Hall/CRC Press, third edition.

Collins, T. C., Petersen, N. J., Menke, T. J., Souchek, J., Foster, W., and Ashton, C. M. (2003). Short-term, intermediate-term, and long-term mortality in patients hospitalized for stroke. *Journal of clinical epidemiology*, 56(1):81–87.

Cox, D. R. (1972). Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220.

Crichton, S. L., Bray, B. D., McKevitt, C., Rudd, A. G., and Wolfe, C. D. (2016). Patient outcomes up to 15 years after stroke: survival, disability, quality of life, cognition and mental health. *J Neurol Neurosurg Psychiatry*, 87(10):1091–1098.

Di Carlo, A., Lamassa, M., Franceschini, M., Bovis, F., Cecconi, L., Pournajaf, S., Paravati, S., Biggeri, A., Inzitari, D., Ferro, S., et al. (2018). Impact of acute-phase complications and interventions on 6-month survival after stroke. A prospective observational study. *PloS one*, 13(3):e0194786.

Doehner, W., Schenkel, J., Anker, S. D., Springer, J., and Audebert, H. J. (2012). Overweight and obesity are associated with improved survival, functional outcome, and stroke recurrence after acute stroke or transient ischaemic attack: observations from the TEMPiS trial. *European heart journal*, 34(4):268–277.

Easton, J. F., Stephens, C. R., and Angelova, M. (2014). Risk factors and prediction of very short term versus short/intermediate term post-stroke mortality: a data mining approach. *Computers in biology and medicine*, 54:199–210.

Efron, B. (1977). The efficiency of Cox's likelihood function for censored data. *Journal of the American Statistical Association*, 72(359):557–565.

Feigin, V. L., Krishnamurthi, R. V., Parmar, P., Norrving, B., Mensah, G. A., Bennett, D. A., Barker-Collo, S., Moran, A. E., Sacco, R. L., Truelsen, T., et al. (2015). Update on the global burden of ischemic and hemorrhagic stroke in 1990-2013: the GBD 2013 study. *Neuroepidemiology*, 45(3):161–176.

Feigin, V. L., Norrving, B., and Mensah, G. A. (2017). Global Burden of Stroke. *Circulation Research*, 120(3):439–448.

Global Burden of Disease (GBD) Mortality and Causes of Death Collaborators (2015). Global, regional, and national age-sex specific all-cause and cause-specific mortality for 240 causes of death, 1990-2013: a systematic analysis for the Global Burden of Disease Study 2013. *Lancet*, 385(9963):117–171.

Goldacre, M. J., Duncan, M., Griffith, M., and Rothwell, P. M. (2008). Mortality rates for stroke in England from 1979 to 2004: trends, diagnostic precision, and artifacts. *Stroke*, 39(8):2197–2203.

Gore, S. M., Pocock, S. J., and Kerr, G. R. (1984). Regression models and non-proportional hazards in the anaylsis of breast cancer study. *Applied Statistics*, 18:176–195.

Graham, J., E Olchowski, A., and Gilreath, T. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention science : the official journal of the Society for Prevention Research*, 8:206–13.

Grambsch, P. M. and Therneau, T. M. (1994). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, 81(3):515–526.

Gresham, G. E., Kelly-Hayes, M., Wolf, P. A., Beiser, A. S., Kase, C. S., and D'Agostino, R. B. (1998). Survival and functional status 20 or more years after first stroke: the Framingham Study. *Stroke*, 29(4):793–797.

Hardie, K., Hankey, G. J., Jamrozik, K., Broadhurst, R. J., and Anderson, C. (2003). Ten-year survival after first-ever stroke in the Perth Community Stroke Study. *Stroke*, 34(8):1842–1846.

Harrell, F. E. and Lee, K. L. (1986). Verifying assumptions of the Cox proportional hazards model. In *SUGI II: Proceedings of the Eleventh Annual SA Users Group International Conference*, pages 823–828. SAS Institute, Inc.

Harrell, Jr., F. E. (2006). *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis.* Springer-Verlag, Berlin, Heidelberg.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additiive Models.* Monographs on Statistics and Applied Probability. Chapman and Hall.

Heldner, M. R., Li, L., Lovett, N. G., Kubiak, M. M., Lyons, S., Rothwell, P. M., and Study, O. V. (2018). Long-term prognosis of patients with transient ischemic attack or stroke and symptomatic vascular disease in multiple arterial beds. *Stroke*, 49(7):1639–1646.

Hess, K. R. (1995). Graphical methods for assessing violations of the proportional hazards assumption in Cox regression. *Statistics in medicine*, 14(15):1707–1723.

Hosmer, D. W., Lemeshow, S., and May, S. (2008). *Applied Survival Analysis: Regression Modeling of Time-to-Event Data.* Wiley-Interscience, New York, NY, USA, 2nd edition.

Intercollegiate Stroke Working Party (2012). *National Clinical Guideline for Stroke*, volume 20083. Citeseer.

Johnson, W., Onuma, O., Owolabi, M., and Sachdev, S. (2016). Stroke: a global response is needed. *Bulletin of the World Health Organisation*, 94(9):634–634A.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481.

Kay, R. (1986). Treatment effects in competing-risks analysis of prostate cancer data. *Biometrics*, 42:203–211.

Kendall, M., Cowey, E., Mead, G., Barber, M., McAlpine, C., Stott, D. J., Boyd, K., and Murray, S. A. (2018). Outcomes, experiences and palliative care in major stroke: a multicentre, mixed-method, longitudinal study. *CMaJ*, 190(9):E238–E246.

Keogh, R. H. and Morris, T. P. (2018). Multiple imputation in Cox regression when there are time-varying effects of covariates. *Statistics in medicine*, 37(25):3661–3678.

Khosravi, H., Khosravi, F., Salari, H., and Khosravi, A. (2017). One-year survival and related factors in patients with ischemic stroke. *International Journal of Health Studies*, 2(4).

Kontopantelis, E., White, I. R., Sperrin, M., and Buchan, I. (2017). Outcome-sensitive multiple imputation: a simulation study. *BMC medical research methodology*, 17(1):2.

Kowarik, A. and Templ, M. (2016). Imputation with the R package VIM. *Journal of Statistical Software*, 74(7):1–16.

Lee, E. T. and Wang, J. W. (2003). *Statistical Methods for Survival Data Analysis*, volume Third edition of *Wiley Series in Probability and Mathematical Statistics*. Wiley-Interscience.

Lee, S., Shafe, A. C., and Cowie, M. R. (2011). UK stroke incidence, mortality and cardiovascular risk management 1999–2008: time-trend analysis from the General Practice Research Database. *BMJ open*, 1(2):e000269.

Li, K.-H., Meng, X.-L., Raghunathan, T. E., and Rubin, D. B. (1991a). Significance levels from repeated p-values with multiply-imputed data. *Statistica Sinica*, 1(1):65–92.

Li, K. H., Raghunathan, T. E., and Rubin, D. B. (1991b). Large-sample significance levels from multiply imputed data using moment-based statistics and an F reference distribution. *Journal of the American Statistical Association*, 86(416):1065–1073.

Lin, D. Y., Wei, L.-J., and Ying, Z. (1993). Checking the Cox model with cumulative sums of martingale-based residuals. *Biometrika*, 80(3):557–572.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data.* Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., 2nd edition.

Luengo-Fernandez, R., Leal, J., and Gray, A. (2015). UK research spend in 2008 and 2012: comparing stroke, cancer, coronary heart disease and dementia. *BMJ open*, 5(4):e006648.

Ma, S., Wang, J., Wang, Y., Dai, X., Xu, F., Gao, X., Johnson, J., Xu, N., Leak, R. K., Hu, X., et al. (2018). Diabetes mellitus impairs white matter repair and long-term functional deficits after cerebral ischemia. *Stroke*, 49(10):2453–2463.

Mant, J., Wade, D., and Winner, S. (2004). *Health care needs assessment: Stroke. In: Stevens A, Raftery J, Mant J, Simpson S (eds) (2004) Health care needs assessment: the epidemiological based needs assessment reviews.* Oxford: Radcliffe Medical Press, second edition.

Marshall, A., Altman, D. G., Holder, R. L., and Royston, P. (2009). Combining estimates of interest in prognostic modelling studies after multiple imputation: current practice and guidelines. *BMC Medical Research Methodology*, 9(1):57.

McKnight, P. E., McKnight, K. M., Sidani, S., and Figueredo, A. J. (2007). *Missing Data: A Gentle Introduction*. Methodology in the Social Sciences. The Guildford Press.

Meng, X.-L. and Rubin, D. B. (1992). Performing likelihood ratio tests with multiply-imputed data sets. *Biometrika*, 79(1):103–111.

Mogensen, U. B., Olsen, T. S., Andersen, K. K., and Gerds, T. A. (2013). Cause-specific mortality after stroke: relation to age, sex, stroke severity, and risk factors in a 10-year follow-up study. *Journal of stroke and cerebrovascular diseases*, 22(7):e59–e65.

Molenberghs, G., Fitzmaurice, G., Kenward, M., Tsiatis, A., and Verbeke, G. (2014). *Handbook of Missing Data Methodology*. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. CRC Press.

Molenberghs, G. and Kenward, M. G. (2007). *Missing Data in Clinical Studies*. Statistics in Practice. John Wiley & Sons, Ltd.

Moons, K. G., Donders, R. A., Stijnen, T., and Harrell Jr, F. E. (2006). Using the outcome for imputation of missing predictor values was preferred. *Journal of clinical epidemiology*, 59(10):1092–1101.

Mozaffarian, D., Benjamin, E. J., Go, A. S., Arnett, D. K., Blaha, M. J., Cushman, M., Das, S. R., de Ferranti, S., Després, J. P., Fullerton, H. J., Howard, V. J., Huffman, M. D., Isasi, C. R., Jiménez, M. C., Judd, S. E., Kissela, B. M., Lichtman, J. H., Lisabeth, L. D., Liu, S., Mackey, R. H., Magid, D. J., McGuire, D. K., Mohler, E. R., Moy, C. S., Muntner, P., Mussolino, M. E., Nasir, K., Neumar, R. W., Nichol, G., Palaniappan, L., Pandey, D. K., Reeves, M. J., Rodriguez, C. J., Rosamond, W., Sorlie, P. D., Stein, J., Towfighi, A., Turan, T. N., Virani, S. S., Woo, D., Yeh, R. W., Turner, M. B., and on behalf of the American Heart Association Statistics Committee and Stroke Statistics Subcommittee (2016). Heart disease and stroke statistics-2016 update a report from the American Heart Association. *Circulation*, 133(4):e38–e48.

Mudzi, W., Stewart, A., and Musenge, E. (2012). Case fatality of patients with stroke over a 12-month period post stroke. *South African Medical Journal*, 102(9):765–767.

Nakibuuka, J., Sajatovic, M., Nankabirwa, J., Ssendikadiwa, C., Furlan, A. J., Katabira, E., Kayima, J., Kalema, N., Byakika-Tusiime, J., and Ddumba, E. (2015). Early mortality and functional outcome after acute stroke in Uganda: prospective study with 30 day follow-up. *Springerplus*, 4(1):450.

National Audit Office (2010). Department of Health: Progress in Improving Stroke Care. `https://www.nao.org.uk/report/department-of-health-progress-in-improving-stroke-care/`. Accessed: October 10, 2018.

Ng'andu, N. H. (1997). An empirical comparison of statistical tests for assessing the proportional hazards assumption of Cox's model. *Statistics in medicine*, 16(6):611–626.

Norrving, B., Leys, D., Brainin, M., and Davis, S. (2013). Stroke definition in the ICD-11 at the WHO. *World Neurol*, 28(3).

Olsen, T. S., Dehlendorff, C., and Andersen, K. K. (2007). Sex-related time-dependent variations in post-stroke survival–evidence of a female stroke survival advantage. *Neuroepidemiology*, 29(3-4):218–225.

Parry-Jones, A. R., Paley, L., Bray, B. D., Hoffman, A. M., James, M., Cloud, G. C., Tyrrell, P. J., Rudd, A. G., and SSNAP Collaborative Group (2016). Care-limiting decisions in acute stroke and association with survival: analyses of UK national quality register data. *International Journal of Stroke*, 11(3):321–331.

Petty, G. W., Brown Jr, R. D., Whisnant, J. P., Sicks, J. D., O'Fallon, W. M., and Wiebers, D. O. (2000). Ischemic stroke subtypes: a population-based study of functional outcome, survival, and recurrence. *Stroke*, 31(5):1062–1068.

R Core Team (2019). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.

Reiter, J. P. (2007). Small-sample degrees of freedom for multi-component significance tests with multiple imputation for missing data. *Biometrika*, 94(2):502–508.

Robinson, R. G. and Jorge, R. E. (2015). Post-stroke depression: a review. *American Journal of Psychiatry*, 173(3):221–231.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys.* Wiley series in probability and mathematical statistics. John Wiley and Sons, Inc.

Rudd, A. and Wolfe, C. (2002). Aetiology and pathology of stroke. *Hospital Pharmacist*, 9:32–36.

Sacco, R. L., Kasner, S. E., Broderick, J. P., Caplan, L. R., Connors, J., Culebras, A., Elkind, M. S., George, M. G., Hamdan, A. D., Higashida, R. T., et al. (2013). An updated definition of stroke for the 21st century: a statement for healthcare professionals from the American Heart Association/American Stroke Association. *Stroke*, 44(7):2064–2089.

Saka, o., McGuire, A., and Wolfe, C. (2009). Cost of Stroke in the United Kingdom. *Age and Ageing*, 38:27–32.

Schoenfeld, D. (1982). Partial residuals for the proportional hazards regression model. *Biometrika*, 69(1):239–241.

Simon, R. and Altman, D. G. (1994). Statistical aspects of prognostic factor studies in oncology. *British journal of cancer*, 69(6):979–985.

Sterne, J. A., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., Wood, A. M., and Carpenter, J. R. (2009). Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *Bmj*, 338:b2393.

Stroke Association (2016). A new era for stroke: Our campaign for a new national stroke strategy. `https://www.stroke.org.uk/get-involved/campaigning/new-era-for-stroke`. Accessed: October 10, 2018.

Stroke Association (2018). State of the Nation: Stroke Statitsitcs. Online Brochure. `https://www.stroke.org.uk/system/files/sotn_2018.pdf`. Accessed: October 9, 2018.

Tableman, M. and Kim, J. S. (2005). *Survival Analysis Using S: Analysis of Time-to-Event Data*. Texts in Statistical Science. Chapman and Hall.

Templ, M. and Filzmoser, P. (2008). Visualization of missing values using the R-package VIM. *Research report cs-2008-1, Department of Statistics and Probability Therory, Vienna University of Technology.*

Therneau, T., Crowson, C., and Atkinson, E. (2019). *Using Time Dependent Covariates and Time Dependent Coefficients in the Cox Model*. Mayo Clinic.

Therneau, T. and Grambsch, P. (2013). *Modeling Survival Data: Extending the Cox Model*. Statistics for Biology and Health. Springer New York.

Therneau, T. M. (2015a). *A Package for Survival Analysis in S*. version 2.38.

Therneau, T. M. (2015b). *A Package for Survival Analysis in S*. version 2.38.

Therneau, T. M., Grambsch, P. M., and Fleming, T. R. (1990). Martingale-based residuals for survival models. *Biometrika*, 77(1):147–160.

van Buuren, S. (2012). *Flexible Imputation of Missing Data*. Interdisciplinary Statistics Series. Chapman and Hall/CRC Press.

van Buuren, S. and Groothuis-Oudshoorn, K. (2011). mice: Multiple Imputation by Chained Equations in R. *Journal of Statistical Software*, 45(3):1–67.

Von Hippel, P. T. (2009). How to impute interactions, squares, and other transformed variables. *Sociological methodology*, 39(1):265–291.

Weimar, C., Ziegler, A., König, I. R., Diener, H.-C., and on behalf of the German Stroke Study Collaborators (2002). Predicting functional outcome and survival after acute ischemic stroke. *Journal of Neurology*, 249(7):888–895.

White, I. R., Daniel, R., and Royston, P. (2010). Avoiding bias due to perfect prediction in multiple imputation of incomplete categorical variables. *Computational statistics & data analysis*, 54(10):2267–2275.

White, I. R. and Royston, P. (2009). Imputing missing covariate values for the Cox model. *Stat Med*, 28(15):1982–1998.

White, I. R., Royston, P., and Wood, A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine*, 30(1):377–399.

Winnett, A. and Sasieni, P. (2001). Miscellanea. A note on scaled Schoenfeld residuals for the proportional hazards model. *Biometrika*, 88(2):565–571.

Winovich, D. T., Longstreth Jr, W. T., Arnold, A. M., Varadhan, R., Zeki Al Hazzouri, A., Cushman, M., Newman, A. B., and Odden, M. C. (2017). Factors associated with ischemic stroke survival and recovery in older adults. *Stroke*, 48(7):1818–1826.

Wolfe, C. D. A. (2000). The impact of stroke. *British Medical Bulletin*, 56(2):275–286.

Wood, A. M., White, I. R., and Royston, P. (2008). How should variable selection be performed with multiply imputed data? *Statistics in medicine*, 27(17):3227–3246.

World Health Organisation (2018). The top 10 causes of death. `http://www.who.int/en/news-room/fact-sheets/detail/the-top-10-causes-of-death`. Accessed: November 15, 2019.

Zhou, M. (2001). Understanding the Cox regression models with time-change covariates. *The American Statistician*, 55(2):153–155.