UNIVERSITY OF EAST ANGLIA

Stroke mortality and morbidity in United Kingdom

Padma CHUTOO

A thesis presented for the degree of
Doctor of Philosophy

School of Computing Sciences

December, 2021

“ *Our greatest glory is not in never falling, but in getting up every time we fall.*”

– Confucius

Abstract

Stroke is a severe, debilitating, and highly prevalent disease that remains a leading cause of mortality and rising healthcare costs. In the UK, there are approximately 113,000 strokes annually (that is, one every 3 minutes 27 seconds) accounting for around 53,000 deaths (Rothwell et al., 2004). Strokes can be ischaemic or haemorrhagic. About 85% of all strokes are ischaemic and 15% are haemorrhagic (Murphy & Werring, 2020). Transient Ischaemic attack (TIA), often referred to as a "mini-stroke" is regarded as a warning sign for future strokes. The risk of premature death and disability is quite high among stroke survivors. Despite significant progress in prevention, treatments, and rehabilitation, there is still great capacity for further improvements, which in turn could reduce the large economic burden of stroke.

The aim of the research is to investigate the impact of ischaemic stroke and TIA events on the long-term survival of affected patients in the UK and to estimate the influence of various risk factors (demographical, lifestyle factors, co-morbidities, and treatments) on the hazards of all-cause mortality of stroke survivors.

Electronic medical records from 1986 to 2016 were extracted from The Health Improvement Network (THIN) database and were used to develop survival models for IS and TIA. Weibull Double-Cox models adapted from Begun et al. (2019) with frailty (random effect) term of general practices were used in modelling the survival.

The study found a much higher risk of all-cause mortality among IS survivors and TIA patients compared to their age, sex, and general practice matched controls. Moreover, in both studies, aspirin prescription was found to be associated with positive long-term survival prospects in cases and to be more effective relative to other antiplatelet treatment options. These important findings are of interest to healthcare professionals, actuaries, and other stakeholders and are beneficial for stroke sufferers.

Acknowledgements

I am extremely grateful to my main supervisor Professor Elena Kulinskaya for her invaluable support, guidance and encouragement throughout this research. I would like to express my thanks to my second supervisor Professor Nicholas Steel for his helpful comments on drafts and precious advice.

My sincere thanks go to Dr. Ilyas Bakbergenuly for extracting the relevant data and the design for the programming functions for longevity models. I would like to thank Dr. Lisanne A Horvat-Gitsels for sharing her excellent statistical skills and resources.

I would like to express my sincere gratitude to the team mates of the Big Health and Actuarial Data team who shared the journey with me.

Finally, I would like to thank my family and friends for their love, patience, and support throughout this process.

Contents

Abstract ii

Acknowledgements iii

1 Introduction 1
- 1.1 Definition of stroke . . . 1
- 1.2 Stroke Epidemiology . . . 1
- 1.3 Rationale of the thesis . . . 2
  - 1.3.1 Impact of TIA and stroke on the individual level . . . 3
  - 1.3.2 Impact on the Healthcare sector . . . 3
  - 1.3.3 Relevance to retirement planning . . . 4
  - 1.3.4 Relevance to the actuarial sector . . . 6
- 1.4 Aims and Objectives . . . 6
- 1.5 Thesis Outline . . . 7

2 Background and Literature Review 9
- 2.1 Ischaemic Stroke and TIA . . . 9
- 2.2 Intra-cerebral haemorrhage . . . 11
- 2.3 Subarachnoid haemorrhage (SAH) . . . 11
- 2.4 Classification of stroke subtypes . . . 11
- 2.5 Incidence and Prevalence . . . 13
  - 2.5.1 Stroke . . . 13
  - 2.5.2 Epidemiology of TIA . . . 13
  - 2.5.3 Global Stroke Burden . . . 13
- 2.6 Management of acute stroke and TIA . . . 14
  - 2.6.1 Stroke and Health Services . . . 14
    - Thrombolysis . . . 15
    - Thrombectomy . . . 15
    - Carotid Endarterectomy Procedure . . . 16
    - Management of ICH . . . 16
    - Management of TIA . . . 17
    - NICE guidelines : Treatment management . . . 17
  - 2.6.2 Stroke: tools and criteria . . . 19

2.7 Incidence of recurrent stroke . . . 20
2.8 Functional outcomes after Stroke . . . 20
2.8.1 Medical complications after stroke . . . 20
2.8.2 Predicting recovery . . . 23
2.8.3 Specific neurologic deficits . . . 24
2.9 Risk factors of stroke and stroke mortality. . . . 24
2.9.1 Non-Modifiable risk factors . . . 24
2.9.2 Modifiable risk factors . . . 25
2.9.3 Stroke Severity . . . 27
2.10 Stroke Treatments . . . 28
2.10.1 Secondary Prevention . . . 28
2.10.2 Stroke units . . . 28
2.10.3 Drug prescriptions . . . 28
2.11 Relative efficacy of antiplatelet interventions in stroke. . . . 29
2.11.1 Limitation of studies : antiplatelets efficacy . . . 32
2.12 Review of Survival analysis after an Ischaemic stroke . . . 32
2.13 Review of survival analysis after TIA . . . 40
2.14 Limitations of survival studies after IS and TIA. . . . 52
2.15 Life expectancy after TIA and stroke . . . 52
2.16 Conclusion . . . 53

3 Methods **55**
3.1 Introduction . . . 55
3.2 Survival Analysis . . . 55
3.3 Censoring and Truncation . . . 58
3.3.1 Right Censoring . . . 58
3.3.2 Left Censoring . . . 59
3.3.3 Interval Censoring . . . 59
3.4 Assumption of uninformative censoring . . . 59
3.5 Parametric survival models . . . 60
3.5.1 The Exponential distribution . . . 60
3.5.2 The Gamma distribution . . . 61
3.5.3 The Weibull distribution . . . 61
3.5.4 Extreme value Distribution . . . 62
3.5.5 The log-normal distribution . . . 62
3.5.6 The log-logistic distribution . . . 62
3.5.7 Other useful survival distributions . . . 62
3.6 Non-Parametric methods of Survival Analysis . . . 63
3.6.1 Kaplan-Meier Estimator . . . 63

3.6.2 Kaplan-Meier Estimator as the Maximum Likelihood Estimator 64
3.7 Greenwood's Formula 65
3.8 Nelson-Aalen Estimator 67
3.8.1 Comparison between two survival functions: The Log-rank test and the Wilcoxon test 67
Log Rank Test: 68
Wilcoxon Test: 69
3.9 The Cox's model 70
3.9.1 Cox's Proportional Hazards (CPH) models 70
3.10 Fitting the Proportional Hazards Model 71
3.10.1 Confidence Interval for estimated coefficients 73
3.10.2 Frailty in Survival model 74
3.10.3 Ties or grouped observations 76
3.11 Proportional Hazard Regression Diagnostics 77
3.11.1 Martingale Residuals 77
3.11.2 Deviance residuals 78
3.11.3 Delta-beta residuals 79
3.11.4 Schoenfeld residuals 79
3.12 Assessing overall goodness of fit 79
3.12.1 Wald Test 80
3.12.2 Likelihood Ratio Test 80
3.12.3 Harell concordance 81
3.12.4 AIC and BIC 81
3.12.5 Model Selection 82
3.12.6 Verifying the proportionality assumption 83
3.12.7 Tackling time-dependent effect 85
3.13 Shape and scale: Weibull model 86
3.14 The Weibull Double-Cox model with shared frailty 87
3.14.1 Point estimation of the parameters 88
3.14.2 Confidence intervals for the parameters 89
3.15 Actuarial Translation of the survival models 90
3.16 Missingness 91
3.16.1 Approaches to missing data 93
Complete-case analysis 94
Maximum likelihood 94
Imputation 94
3.17 Multiple Imputation 95
3.17.1 Rubin's Rules 98

3.17.2 How many imputations are required? . . . 99
3.17.3 Variable selection after Multiple Imputation. . . . 100
3.18 Chapter summary . . . 101

4 Data Selection and Description **103**
4.1 The Health Improvement Network (THIN) database . . . 103
4.2 Structure of the THIN database . . . 103
4.2.1 Practice file . . . 104
4.2.2 Main files . . . 104
4.2.3 Linked files . . . 104
4.3 Data Extraction . . . 106
4.3.1 Selection criteria . . . 106
Data flags used in THIN database . . . 106
Practice Inclusion Criteria . . . 107
Patients Inclusion Criteria . . . 107
Selection of cases . . . 109
Further Exclusion criteria: Some medical conditions . . . 109
Selection of controls . . . 109
4.3.2 Variables of interest . . . 109
4.4 Dataset . . . 110
4.5 Strength and Limitations of THIN Database . . . 110
4.6 Data description . . . 111
4.6.1 Description of variables . . . 114
Demography, Socio-economic and Social deprivation factors . . . 114
Lifestyle factors . . . 117
Treatments . . . 117
Medical Conditions . . . 117
Medical Tests and their use for diagnoses . . . 118
4.6.2 Asthma . . . 119
4.6.3 Atrial Fibrillation . . . 119
4.6.4 Hypercholesterolemia (Dyslipidemia) . . . 119
4.6.5 Chronic obstructive pulmonary disease (COPD) . . . 120
4.6.6 Hypertension . . . 120
4.6.7 Diabetes . . . 120
4.6.8 Heart Failure . . . 120
4.6.9 Peripheral artery disease (PAD) . . . 121
4.6.10 Hypothyroidism . . . 121
4.6.11 Chronic Kidney Disease, CKD . . . 121
4.6.12 Cardio-Vascular Diseases (CVD) . . . 121

4.6.13 Minor Cancer . . . 122
4.7 Exploratory data analysis . . . 122
4.8 Chapter Summary . . . 126

5 Survival analysis after Transient Ischaemic attack 127
5.1 Study Design . . . 127
5.1.1 Selection criteria . . . 127
5.2 Description of the full-case dataset . . . 129
5.2.1 Univariate analysis . . . 134
5.3 Model development and diagnostic tests . . . 135
5.3.1 Weibull Double-Cox model. . . . 138
5.3.2 Overview of Full - Case model: TIA . . . 140
5.4 Multiple Imputation . . . 146
5.4.1 Exploring Pattern of Missingness. . . . 146
5.4.2 Multiple imputation model . . . 148
5.5 Final models: Results from data from the imputed model. . . . 152
5.5.1 Time-invariant effects . . . 153
5.5.2 Time-varying effects . . . 155
5.6 Morbidity . . . 157
5.7 Life Expectancy model for TIA . . . 157
5.8 Discussion . . . 167
5.9 Conclusions . . . 169

6 Survival analysis after ischaemic stroke 171
6.1 Study Design . . . 171
6.2 Univariate analysis of predictors . . . 174
6.3 Model Development and Diagnostic tests . . . 175
6.4 Double-Cox Weibull model . . . 176
6.5 Overview of full-case of survival after ischaemic stroke . . . 178
6.6 Multiple Imputation Procedures . . . 179
6.7 Results of the survival models on the data from imputed dataset . . . 182
6.7.1 Time-invariant effects of the model (6.4) . . . 187
6.7.2 Time-varying effects of the model (6.4) . . . 187
6.8 Life Expectancy model for IS . . . 189
6.9 Discussion . . . 196
6.10 Conclusions . . . 198

7 Conclusions 199
7.1 Introduction . . . 199

7.2 Main findings . . . 199
7.3 Survival study after TIA . . . 199
7.4 Survival analysis after an IS . . . 200
7.5 Study Strengths . . . 201
7.6 Study Limitations . . . 202
7.7 Implications . . . 203
7.8 General conclusions . . . 205

A Readcodes **207**

B Supplementary materials : Survival model after TIA **209**
B.1 Tables . . . 209
B.2 Figures . . . 222

C Supplementary materials : Survival model after IS **223**
C.1 Tables . . . 223

Bibliography **231**

List of Figures

1.1 Trends in life expectancy at birth, males and females, England, 1981 to 2017, projections and forecasts from 2018 to 2023. 5

2.1 CT of normal brain versus CT of Ischaemic Stroke. 9
2.2 CT scans showing (A) intracerebral haemorrhages (arrows) and (B) subarachnoid haemorrhages (arrows). 11
2.3 An Endovascular therapy. 16
2.4 A Carotid Endarterectomy Procedure. 17
2.5 Treatment pathways for TIA and acute Stroke adapted from NICE guidelines . . . 18
2.6 Barthel Index items 21

3.1 An illustrative example of the Survival and Hazard curves of the study population during 16 years of follow-up 57
3.2 Log-logistic hazard function with shape 2 and scale 1. 63
3.3 Illustration of the effects of the different shapes and scales on the Weibull hazards . 87
3.4 Illustration of the Multiple Imputation process. 96

4.1 Structure of The Health Improvement Network (THIN) database 105
4.2 Townsend Deprivation Score for England and Wales from Office for National Statistics, 2011 Census. 116
4.3 Incidence of prior risk factors and medication use over time in patients with first-ever IS by gender. 123
4.4 Incidence of prior risk factors and medication use over time in patients with first-ever IS by gender. 124
4.5 (a) Proportion of pre-morbid prescription of selected drugs to IS patients by calendar year. (b) 1 month, 3 months and one-year all-cause mortality for first-ever Ischaemic stroke patients 1990-2016. 125

5.1 Data extraction in the TIA dataset 128
5.2 Kaplan-Meier plot after first-ever TIA event. 132
5.3 Unadjusted Kaplan-Meier plot by age groups at TIA event and case-control status. 133
5.4 Percentage of cases and controls lost due to follow-up by social deprivation, age category at entry and TIA diagnosis. 135
5.5 Schoenfeld residuals against the transformed time. 137

5.6 Graphical comparison of the fit of several distributions: . . . 138
5.7 Comparison of the baseline cumulative hazard functions estimated using semi-parametric (magenta) and parametric (Weibull model, grey) methods. Age, sex and case-controls groups. . . . 139
5.8 Plot of the residuals between the parametric and semi-parametric estimates of the baseline hazards pooled across the strata. . . . 140
5.9 Comparison of fit of estimated survival to empirical across different strata. . . . 145
5.10 Missingness pattern in TIA dataset. . . . 146
5.12 Comparison of distribution of imputed variables with distribution of full case. . . . 150
5.13 Forest plot of the adjusted hazard ratios for the factors having scale effect only. . . 160
5.14 Cummulative hazard curves of the different birth cohorts for male cases and controls. 161
5.15 Cummulative hazard curves of the effect of heart failure on male cases and controls. 162
5.16 Hazard ratio curves with 95% confidence intervals for patients aged 39-60 years at entry (top panel) and aged 61-70 years at entry (bottom panel) different antiplatelets. 163
5.17 Hazard ratio curves with 95% confidence intervals for patients aged 71-76 years at entry(top panel) and aged 76 and above at entry(bottom panel) years at entry on different antiplatelets. . . . 164
5.18 Comparison of empirical and estimated survival plots for different strata. . . . 165

6.1 Data extraction in the IS dataset. . . . 172
6.2 Schoenfeld residuals based on variable stroke diagnosis. . . . 175
6.3 Graphical comparison of the cumulative hazards of several survival distributions with respective AIC: *Exponential, Weibull, Logistic, log-logistic, Log-Normal and Gompertz compared to KM cummulative hazards function.* . . . 176
6.4 Comparison of the baseline cumulative hazard functions estimated using semi-parametric (magenta) and parametric (Weibull model, grey) methods. Age, sex and case-controls groups. . . . 177
6.5 Plot of the residuals between the parametric and semi-parametric estimates of the baseline hazards pooled across the strata. . . . 178
6.6 Comparison of distribution of imputed data and full case data distribution in IS data. 180
6.7 Comparison of fit of estimated survival to empirical across different strata. . . . 186
6.8 Hazard ratios of all-cause mortality curves with 95% confidence intervals for patients aged $39-60$ and $61-70$ years at entry treated with different antiplatelets. . . 191
6.9 Hazard ratios of all-cause mortality curves with 95% confidence intervals for patients aged $71-76$ and $77+$ years at entry treated with different antiplatelets. . . . 192
6.10 Comparison of empirical and estimated survival plots for different strata. . . . 193
6.11 Prevalence of medical conditions in patients by age of entry. . . . 194

B.1 The missing pattern in the TIA dataset and the association of missing records. . . . 222

List of Tables

1.1 Major risk factors for stroke 2

2.1 Tools and criteria for stroke 19
2.2 Example of disability measures: the Modified Rankin Scale 22
2.3 Important meta-analyses and trials of antiplatelets therapy in stroke and TIA patients. 31
2.4 Review of studies on survival after IS stroke. 34
2.5 Review of studies on survival after TIA or minor stroke. 42

3.1 Number of deaths and survivors in two risk groups at time t_i 68

4.1 Patient Flag code used in THIN database 107
4.2 Registration Status Flag used in THIN database 108
4.3 Description of variables used in the study 112
4.3 Description of variables used in the study 113
4.4 Description of 15 Mosaic groups based on the Experian Mosaic (2014) 115

5.1 The distribution of variables with missing values in the TIA dataset 127
5.2 Baseline characteristics of TIA dataset 129
5.3 Univariate analysis of predictors for the TIA cohort. 134
5.4 Parameter estimates of the Weibull double-Cox model for the full-case model. . . . 142
5.5 Comparing the distributions in complete case and completed dataset. 149
5.6 Final model : Description, parameter estimates and confidence intervals for the Weibull Double-Cox model with frailty terms. 152
5.7 Adjusted estimated hazard ratios of all-cause mortality by age at diagnosis and case-control status (from the Weibull Double-Cox model on the imputed data). 156
5.8 Adjusted estimated hazard ratios of all-cause mortality by hypertension status and case-control status (from the Weibull Double-Cox model on the imputed data) . . . 156
5.9 Vascular events at follow-up of the TIA cohort. 157
5.10 Calculation of life expectancies of individuals in the TIA dataset at different ages. 166

6.1 The distribution of variables with missing values in the IS dataset before imputation procedures. 173
6.2 Univariate analysis of predictors of the IS cohort 174
6.3 Baseline characteristics of IS dataset. 181

6.4 Final model on pooled data from imputed models: Description, parameter estimates and confidence intervals for the Double -Cox Weibull distribution model with frailty terms. . . . 183
6.5 Adjusted estimated hazard ratios of all-cause mortality by age at diagnosis and case-control status (from the Weibull Double-Cox model on the imputed data.) . . . 187
6.6 Calculation of life expectancies of individuals in the IS dataset at different ages. . 195

A.1 Readcodes for TIA diagnosis . . . 207
A.2 Readcodes for IS diagnosis . . . 208

B.1 Characteristics of the cases and controls: Original dataset. . . . 209
B.2 Cases and controls lost to follow-up by age category, social deprivation and TIA diagnosis. . . . 211
B.3 Description and coding of variables in the study . . . 211
B.4 Results of the Grambsch and Therneau (1994)'s test of the assumption of proportional hazards . . . 213
B.5 Table explaining the patterns of missingness between variables. . . . 213
B.6 Correlation matrix jointly missing variables. . . . 214
B.7 Multiplicative hazard model : test for time-invariant effects (Kolmogorov-Smirnov test) . . . 214
B.8 Multiplicative hazard model : test for time-invariant effects (Cramer von Mises test.) 215
B.9 Life expectancy model for the TIA database. . . . 216
B.10 Estimated coefficients of full-case model and model on the imputed data and their significance. . . . 217
B.11 Estimated coefficients of full-case model and model on the imputed data and their significance: Survival model for Life expectancy. . . . 219
B.12 Hazard ratios over different years associated with uptake of different types of antiplatelets for TIA patients by different age groups. . . . 220
B.13 Hazard ratios over different years associated with uptake of different types of antiplatelets for TIA controls by different age groups. . . . 221

C.1 Results of the Grambsch and Therneau (1994)'s test of the assumption of proportional hazards. . . . 223
C.2 Estimated coefficients and associated significance level of full-case model and data on the imputed data for survival after IS. . . . 224
C.3 Hazard ratios over different years associated with uptake of different types of antiplatelets for patients aged 39−60 years at entry (*IS dataset).* . . . 226
C.4 Hazard ratios over different years associated with uptake of different types of antiplatelets for patients aged 61−70 years at entry (*IS dataset).* . . . 226

C.5 Hazard ratios over different years associated with uptake of different types of antiplatelets for patients aged 71−76 years at entry (*IS dataset).* 226
C.6 Hazard ratios over different years associated with uptake of different types of antiplatelets for patients aged 77+ years at entry (*IS dataset).* 227
C.7 Life expectancy model using the IS dataset fitted to a Weibull double-Cox model. . 228
C.8 Estimated coefficients of full-case model and model on the imputed data and their significance: Survival model for Life expectancy. 229

List of Abbreviations

ADL	Activities of Daily Living
AF	Atrial Fibrillation
AMI	Acute Myocardial infarction
BMI	Body Mass Index
BP	Blood Pressure
CT	Computed Tomography
CHD	Cardiovascular Disease
DBP	Diastolic Blood Pressure
ECG	Electrocardiography
FRS	Framingham Risk Score
HF	Heart Failure
HR	Hazard Ratio
ICD	International Classification of Disease
ICH	Intracerebral haemorrhage
INR	International Normalised Ratio
IS	Ischaemic Stroke
KM	Kaplan-Meier
MI	Myocardial Infarction
mRS	Modified Rankin Score
NHS	National Health Services UK
NIHSS	National Institutes of Health Stroke Scale
OR	Odds ratio
OSCP	Oxfordshire Stroke Community Project
SAH	Subarachnoid haemorrhage
SBP	Systolic Blood Pressure
SMR	Standardized Mortality Ratio
SU	Stroke Units
TIA	Transient ischaemic attack
THIN	The Health Improvement Network
APL	Antiplatelets
DAPT	Dual Antiplatelet Therapy Network
NICE	The National Institute for Health and Care Excellence

Publications

- Chutoo, P. (2020). Survival analysis after a first ischaemic stroke event: A case-control study in the adult population of England. *Methodology, 2,1*.
- Chutoo, P., Kulinskaya, E., Bakbergenuly, I., Steel, N., Pchejetski, D., & Brown, B. (2022). Long-term survival after a first transient ischaemic attack in England: A retrospective matched cohort study. *Journal of Stroke and Cerebrovascular Diseases, 31(9)*, 106663.

To my dad, a double stroke survivor . . .

1 Introduction

The initial chapter endeavours to investigate varying definitions of stroke types and provides a general background and justification for the research. The research aims and objectives are then listed.

1.1 Definition of stroke

The World Health Organisation (WHO, 1988) defines stroke as a clinical condition characterised by rapidly growing clinical indications of localised impairment of brain function that lasts longer than 24 hours or results in death, with no evident cause other than vascular origin. Stroke can be ischaemic or haemorrhagic. An ischaemic stroke is caused by the blood supply to the brain being cut off due to blockage which is the most common type of stroke , occurring in $80-85\%$ of cases (Gomes et al., 2014). A haemorrhagic stroke is caused by a bleeding in or around the brain due to rupture of important blood vessels. A transient ischaemic attack (TIA) is traditionally defined as appearance of stroke symptoms and signs that resolve completely within 24 hours.

1.2 Stroke Epidemiology

In 2010, the American Heart Association reported stroke to be the second leading cause of death and the leading cause of long-term disability worldwide (Lloyd-Jones et al., 2010). The global prevalence of stroke was 33 million, with 16.9 million people having a first stroke. They further observed that stroke accounted for 11.13% of total deaths worldwide. Saka et al. (2009) reported that worldwide, stroke consumes 2% to 4% of health care resources. They estimated that stroke represented a financial burden of £7 billion per year to the UK economy, with direct costs to the National Health Service in the UK of £2.8 billion. Acute Ischaemic stroke accounts for 83% of all strokes in the western world, Intracerebral haemorrhage (ICH), 10% and Subarachnoid haemorrhage (SAH) 3% (Lloyd-Jones et al., 2010). The distribution of the different sub-types of stroke may not be necessarily similar for different parts of the world. As evidenced by the Global Burden of Disease Study 2010, the proportion of strokes due to ICH is higher (up to 39%) in certain Asian countries such as China and the Japan (Krishnamurthi et al., 2013). ICH and ischaemic stroke share a number of risk factors outlined in Table 1.1.

Stroke is the leading cause of mortality and disability worldwide, in both developed and developing countries. According to the Global Burden of Disease survey, stroke is the third leading cause of disability-adjusted life years (DALYs) lost worldwide. The number of years lost due to illness,

TABLE 1.1: Major risk factors for stroke

Risk type	Acute Ischaemic stroke	Intracerebral Haemorrhage
Non-modifiable	Increasing age, family history of stroke, prior stroke/TIA, male gender.	Increasing age, male sex, previous ICH, intracranial aneurysms, cerebral small vessel disease, arteriovenous malformations.
Modifiable	Hypertension, diabetes mellitus, smoking, hypercholesterolaemia, ischaemic heart disease, carotid artery stenosis.	Hypertension, anticoagulant therapy, high alcohol intake, smoking, diabetes mellitus, illicit drug use.

Source:Manning et al. (2016)

disability, or early death is measured in DALYs (Murray & Lopez, 1997). Feigin and Krishnamurthi (2011) estimated that number of deaths and DALYS lost attributable to stroke in 2010 were 5.9 million and 102 million, respectively.

They also pointed out that strokes in the 20 to 64 years age group, now make up nearly a third of the total number of strokes compared with a quarter of strokes in 1990. According to them, " Stroke should no longer be regarded as a disease of old age " because stroke has traditionally been regarded as a condition affecting elderly people, but the proportion of younger people affected by stroke is increasing and this is likely to continue. Unless effective prevention techniques are implemented, the number of cases of disability, disease, and premature mortality caused by stroke is expected to more than treble by 2030 (Mukherjee & Patil, 2011).

1.3 Rationale of the thesis

Stroke is a disease that may occur at any age and it is often due to a combination of numerous risk factors. The risk of mortality or recurrent stroke is highest within the first year. The consequences of stroke are related to several factors such as stroke severity, treatments and pre-existing medical conditions (e.g. diabetes, hypertension), lifestyle and demographical factors. One of the comorbidities highly associated with stroke is atrial fibrillation, a type of irregular heartbeat. Around 1.2 million people in the UK have atrial fibrillation (AF) (NHS Choices, 2017). Having an AF implies that blood clots are more likely to be formed in the heart which thereby increases the risk of stroke. For the past decades, stroke incidence has decreased in high income countries (Feigin et al., 2009) but the absolute number of cases is expected to increase in the future due to a growing number of people and elderly in the population (Béjot et al., 2019; Gorelick, 2019; Rothwell et al., 2004). Survival after stroke has improved during the last decades (Bhatnagar et al., 2015; Boysen et al., 2009; Lee et al., 2011; Rothwell et al., 2004). Nevertheless, many stroke survivors still have an impaired prognosis when compared to a healthy person. Stroke frequently necessitates lengthy hospitalisation

and rehabilitation and reduced quality of life. Stroke is linked to mental and emotional repercussions such as dementia, depression, and subsequent medical difficulties, in addition to the physical consequences. Stroke and transient ischemic attack (TIA) have a significant impact on individuals and society, hence primary prevention is critical to lowering their occurrence and associated costs. Stroke is a huge financial burden on society, accounting for 2 to 4% of overall healthcare costs in developed countries (Ingeman et al., 2011). Most studies have focused on prognosis such as recurrence and functional status after stroke, rather than long-term mortality. Most research have been done on specific target groups, hence not large scale enough to be generalisable to wider population. Primary care data constitutes a rich source of population-based longitudinal data on patients. Data on stroke diagnoses and treatment are collected along their routine care journey. Hence, there is a need for survival modelling based on large-scale population-based primary care data.

The following study will benefit the patients, caregivers, health authorities, clinicians, actuaries and government in terms of identification of risks, healthcare planning, resource planning and retirement planning, pricing of annuities and financial planning.

1.3.1 Impact of TIA and stroke on the individual level

TIA is often considered a "minor stroke" and stroke is considered a " disease of the old ", i.e. to be happening only in older individuals. The symptoms of TIA and stroke are often poorly recognised. Stroke patients may be left with disability and experience physical, emotional or cognitive impairment such as fatigue, memory, or concentration problems. The quality of life may be seriously affected. This can also consequently become a financial strain on the individuals and family life may be affected. Many stroke survivors may continue to require care from caregivers, and family members for activities of daily living like bathing or toileting. This may further impact the psychosocial well-being and a decline in social life with ensued high level of depression (Doran et al., 2014). The hidden syndromes if not managed may also result in subsequent vascular events or even premature deaths. Awareness is indispensable. A survival model estimating the mortality risk after diagnosis of TIA and stroke can help identify which types of interventions and factors are beneficial for survival after the event.

1.3.2 Impact on the Healthcare sector

Healthcare professionals need to understand factors associated with survival improvements or those associated with hazardous effects. This knowledge could help to understand the risks associated with health factors and mitigate the burden associated with stroke. Stroke is the common cause of disability in the adult population. Approximately, two-thirds of stroke survivors are left with a disability (Dworzynski et al., 2013). Some of them are left with permanent disabilities often requiring rehabilitation and long-term care. The reduced quality of life could be a consequence of physical, mental, or emotional impairment. The health care costs are estimated to be £8 billion

annually with £3 billion direct costs to the NHS (Patel et al., 2020). The societal economic burden is made up of direct costs such as lost work days and informal care, as well as indirect costs such as substantial costs connected to home or hospital-based care and rehabilitation. The current high stroke burdens are expected to increase as the ageing population and obesity, diabetes, and physical inactivity become more common.

Early management can help to reduce the NHS healthcare costs. The post-stroke hospitalisation is also much higher in stroke patients than in the matched non-stroke cohort (Public Health England, 2018). Public Health England (2018) warns that the average age for individuals suffering from stroke is decreasing, with more than one-third of strokes occurring in adults between 40 and 69. The middle-aged stroke survivors are at their working-age (Daniel et al., 2009) and may also face a considerable risk of recurrent stroke and other adverse vascular events.

It is important to investigate the impact of the use and adherence to secondary prevention medication. Survival models after stroke can help to identify different characteristics associated with improvement in mortality and how the risk varies across sub-populations. Treatment effects may vary by population. A survival study can help to assess the effectiveness of different medical interventions at different ages and morbidity profiles which can inform clinical guidelines. This can benefit the local authorities and healthcare professionals, caregivers and families providing a strategic avenue for resource allocation and management. In the light of the above considerations, there is a strong case to argue that the accurate survival prognosis after stroke can impact the healthcare sector.

It is important that doctors and healthcare professionals are made aware of the mortality risk in different age groups after a medical diagnosis of stroke. Besides, there is a strong need to understand the mortality risk given the different risk factors to be able to curb the harmful effects. Thirdly, it is of great significance to assess the effectiveness of medical interventions in view of better medical management. Clinical trials being highly selective often exclude elderly patients, hence the effectiveness of medical therapies after a stroke needs a closer inspection, a more "real world" application. Survival models after stroke incorporating risk factors, clinically relevant interactions, medical therapies with diverse age groups, comorbidities, and lifestyle choices can help answer these questions.

1.3.3 Relevance to retirement planning

Defined Benefits and Defined Contribution funds(there is a recent increase towards this option) are two types of retirement funds in the United Kingdom. Guaranteed annuities, lump payouts, and With Profit annuities are available to members. Insurers and annuity providers, on the other hand, are concerned about longevity risk. Individuals who are insured may live longer than expected.

The emergence of enhanced (impaired) life annuities based on the insured's smoking status, general health, geographic location, and pre-retirement occupation has been lately noted in the market (HannoverRE, 2018).

The investment and longevity risk in Defined Benefit funds and guaranteed annuities are borne by the life insurer, which pays pensions for longer than projected. People may not fully appreciate longevity risk or consider its repercussions while planning their retirement income due to uncertainty and underestimation of life expectancy. There have been major improvements in UK life expectancy. Public Health England (2017) produced life expectancy forecasts based on fitting a model to the trends in life expectancy from 1981 to 2017 (prior to the COVID-19 pandemic). The graph in Figure 1.1 shows the improvement in life expectancy.

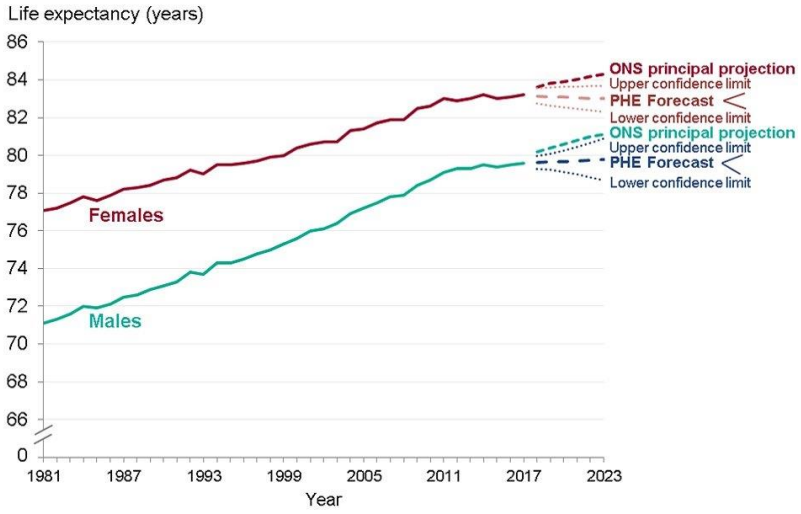


FIGURE 1.1: Trends in life expectancy at birth, males and females, England, 1981 to 2017, projections and forecasts from 2018 to 2023.

Source: PHE analysis of ONS data (2018)

However, recent mortality has been volatile especially with the ongoing COVID-19 pandemic causing mortality shocks. The predictability of the longevity expectation can prove to be challenging. Using data from the Office for National Statistics, Aburto, Kashyap, et al. (2021) shows in comparison to 2019, women's and men's life expectancy at birth decreased by 0.9 and 1.2 years, respectively in 2020. COVID-19 pandemic put a halt to the longevity improvements across many countries; Life expectancy at birth fell in 24 of the 26 nations, including most European countries, Chile, and the United States (Aburto, Schöley, et al., 2021).

It is vital to provide accurate information on a variety of items so that individuals may better understand their longevity risk and take appropriate decisions. It matters to study their longevity risk near retirement ages. The survival model can hence inform how certain socio-demographic factors,

medical conditions, treatments, and lifestyle factors can impact survival prospects at retirement. Stroke is highly prevalent among patients older than 65 years, and improved knowledge about survival prospects can be helpful for insured individuals to plan for retirement and to decide on the best option given their estimated risk of mortality.

1.3.4 Relevance to the actuarial sector

The life insurance sector may experience unexpected loss of profit due to overestimation or underestimation of base risk. If the risk is underestimated, the loss could be in the form of income offered to consumers until death. An unexpected rise in longevity could be the result of shifting population lifestyles or medical advancements. When projecting life expectancy, it is critical for insurance companies to understand the risk factors that influence longevity risk. Hence, appropriate survival models can test if there are survival prospects related to lifestyle choices or treatments. These models can provide insights into properly calibrating the life expectancy model and hence minimizing the basis risk. It is important to find new reliable sources of stroke data with a longer length of follow-up and more coverage of patients representative of the UK population. The mortality risk associated with stroke impacts the actuarial and insurance sectors in the UK. Survival models can hence help actuaries to design improved annuities products tailored to customer's risk.

1.4 Aims and Objectives

Aims

The main aim of the research is to study the impact of stroke on the survival of stroke patients in the UK while adjusting for various risk factors (demographical data, treatments, comorbidities, and lifestyle factors) on the hazards of all-cause mortality after IS stroke and TIA. The treatment effects of different medical therapies will also be explored. Survival analysis will be used to model the longevity of patients after stroke and TIA.

Objectives :

1. To establish the list of risk factors and covariates affecting the longevity of stroke survivors and TIA patients based on the review of the medical literature on the relevant medical condition.

2. To use Cox proportional hazards regression or parametric survival models as appropriate to estimate survival benefits or hazards of these factors and their interactions and find through backward elimination the subset of variables to be adjusted for in the full case analysis.

3. To deal with missing values in the database using multi-level multiple imputation and model checks.

4. To determine the effects of stroke and TIA on the longevity of patients.
5. To estimate the effect of various risk factors such as pre-existing medical conditions, socio-demographic and lifestyle factors and their interactions. To estimate the protective or harmful effects of the medical therapies on survival.
6. To translate the results of the survival models for stroke and TIA into life expectancies.
7. To examine years gained or lost due to medical treatments, comorbidities, and lifestyle factors for TIA patients and stroke survivors.
8. To disseminate the findings to inform the patients, healthcare professionals, and actuaries on the effective health interventions and financial planning.

1.5 Thesis Outline

Chapter 2 is the literature review of the TIA and IS. First, the different risk factors and risk management are discussed. The results of important randomized control trials, meta-analyses, and survival models are reviewed. The limitations of existing research are presented.

Chapter 3 consists of the description of the statistical methods, assumptions of the survival models, relevant statistical tests, and the process of model building. It also takes into consideration the multiple imputation techniques which were used to deal with missing records.

Chapter 4 is a review of the THIN database. It also presents the inclusion-exclusion criteria and the extracted variables are presented.

Chapter 5 presents the developed survival model to estimate the hazards of all-cause mortality associated with a history of TIA. The analysis procedures are explained and the findings are discussed and compared to other studies.

Chapter 6 presents the developed survival model to estimate the hazards of all-cause mortality associated with a history of IS. The analysis procedures are explained and the findings are discussed and compared to previous findings of the literature.

Chapter 7 summarises the research's findings. The strengths and limitations of the research are discussed, followed by the implications of the study.

2 Background and Literature Review

The Chapter starts with some definitions of stroke types and provides a general background. It then gives an overview of the existing literature. The limitations of the existing studies are explored and the gaps in research are then identified. The first section explores the risk assessment and the management of stroke and TIA in UK clinical practice. The following section then provides an overview of the key studies on survival following TIA or stroke and the patients' survival prospects are discussed to identify the gaps in stroke literature.

2.1 Ischaemic Stroke and TIA

Ischaemic stroke is caused by obstruction of a cerebral artery, stopping the blood supply to the brain, and causing ischaemia (decrease in blood supply causing tissue death) in the affected vascular territory. Such occlusions may be due to thrombosis, typically on a background of atherosclerotic arterial plaques, or thrombi emboli from distant sites, most commonly from the left chamber of the heart, or from large arteries such as the aorta or carotid arteries (G. J. Hankey, 2007).

In **TIA**, the obstruction to blood flow is **transient (temporary)**; with the restoration of blood flow to the affected area preventing the development of permanently damaged brain tissue.

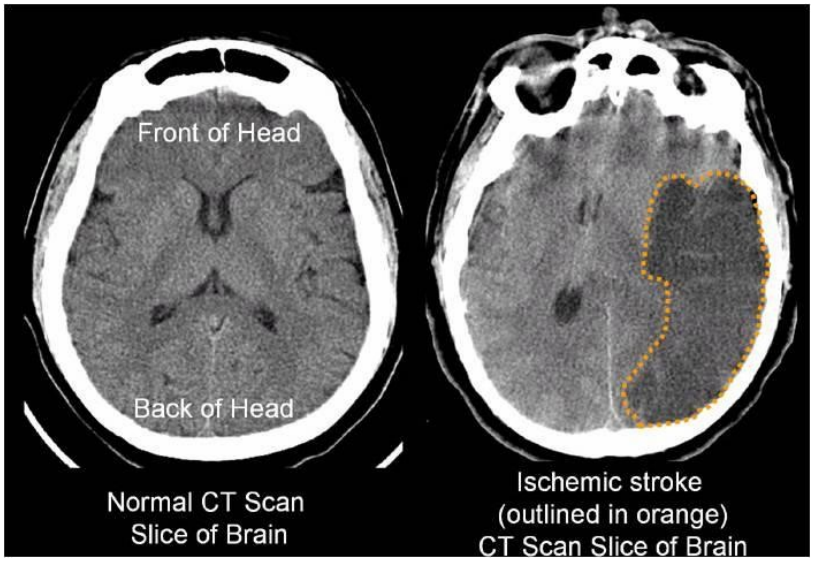


FIGURE 2.1: CT of normal brain versus CT of Ischaemic Stroke.

Source: NE 103 Cerebrovascular Accidents

"Transient" implies brief; "ischaemic" implies the lack of blood flow, and "attack" implies a suddenness and limited time duration of a discrete event. The transient ischaemic attack was first defined in 1975 by the Ad-Hoc Committee on Cerebrovascular Disease as "cerebral dysfunction of ischaemic nature lasting no longer than 24 hours with the tendency to recur" (Millikan et al., 1975). World Health Organisation (WHO, 1988) defined transient ischaemic attack as "rapidly developed clinical signs of focal or global disturbance of cerebral function, lasting less than 24 hours, with no apparent non-vascular cause". Due to controversy in the definition of TIA which was based on an arbitrary time cut-off and which was causing confusion, the American Heart Association and American Stroke Association (Easton et al., 2009) consider the above definition to be outdated.

Conventionally, it was assumed that TIA left no damage to brain tissues (infarction) and that symptoms would completely stop within the 24 hours window. Any symptoms lasting longer than 24 hours up to 7 days were classified as ischaemic stroke. Albers et al. (2002)'s study highlighted the inconsistency in the concept of the traditional definition of TIA and attempted a redefinition. They defined TIA as "a brief episode of neurological dysfunction caused by focal brain or retinal ischemia, with clinical symptoms lasting less than an hour and without neuro-imaging evidence of acute infarction". The Stroke Council of the American Heart Association/American Stroke Association (AHA/ASA) (Easton et al., 2009) removed time as a factor and recommended their current definition of transient ischaemic attack as: "a transient episode of neurological dysfunction caused by focal brain, spinal cord or retinal ischaemia, without acute infarction." Computed tomography (CT) and later magnetic resonance imaging (MRI) questioned the accepted paradigm on how TIAs are assessed by demonstrating that clinically transient episodes are not always temporary at the tissue level. This new definition is "tissue-based" rather than a time-based definition with the diagnosis being based on brain imaging. Despite the revised tissue definition, significant organizations like the World Health Organization (WHO) and the National Institute for Health and Care Excellence (NICE) continue to use the 24-hour threshold (Turner, 2016). The lack of routine brain imaging could be explained by restrictions based on the availability, viability, and price of MRI for use in emergency care of TIA in many clinics and medical centres. Furthermore, about 10% of patients are either unable to tolerate MRI or it is contraindicated, which restricts its general use.

For the given study context, the electronic medical records from the THIN database were used in the UK primary care and hence the definitions used for strokes and TIA were based on NICE guidelines (NICE, 2020b). According to NICE (2020b), ischaemic stroke is defined "as an episode of neurological dysfunction caused by focal cerebral, spinal, or retinal cell death due to infarction following vascular occlusion or stenosis and TIA as "a transient (less than 24 hours) neurological dysfunction caused by focal brain, spinal cord, or retinal ischemia, without evidence of acute infarction." Hence, TIA diagnosis is based solely on clinical diagnosis and not necessarily on brain imaging.

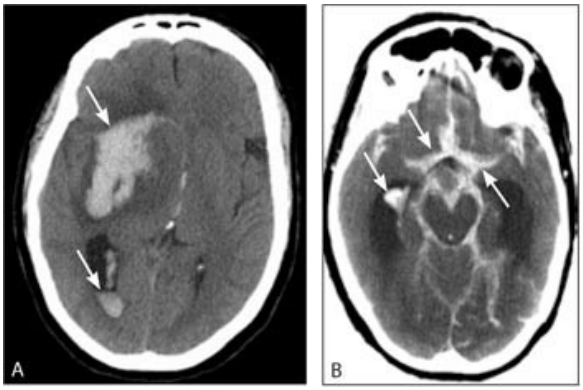


FIGURE 2.2: CT scans showing (A) intracerebral haemorrhages (arrows) and (B) subarachnoid haemorrhages (arrows).

Note that acute haemorrhage appears hyperdense (white) on a CT scan.
Source: Kenneth & Cheng (2009) Acute Stroke Diagnosis

2.2 Intra-cerebral haemorrhage

Intracerebral haemorrhage (ICH) consists of three distinct phases: Initial haemorrhage, haematoma expansion (collection of blood outside of a blood vessel), and peri-haematoma oedema (causes the brain to swell within the skull, leading to increased pressure, reduced conscious level and impaired neurological function). According to Manning et al. (2016), the initial haemorrhage phase is caused by sudden rupture of small intra-cerebral arteries, damaged by hypertension or amyloid angiopathy. Amyloid angiopathy is a blood vessel disorder caused by abnormal amyloid (build-up of abnormal proteins) deposits in the blood vessel walls of the brain which can cause the blood vessel to become weak and rupture resulting in intracranial bleeding.

Hypertension is by far the leading cause of ICH and accounts for approximately two-thirds of cases (Mayer & Rincon, 2005).

2.3 Subarachnoid haemorrhage (SAH)

Head trauma is the most common cause of SAH. SAH is also commonly due to a brain aneurysm which is a swelling of an artery in the brain that can rupture and bleed into the space between the brain and the skull. The main risk factors contributing to SAH are hypertension, cigarette smoking, excessive alcohol use, illicit drugs use or familial history of brain aneurysms.

2.4 Classification of stroke subtypes

Stroke leads to an unexpected beginning of various clinical features, characterised by loss of function through the nature of signs. The symptoms depend on the anatomical site and size of the

affected area. A widely accepted clinical classification for ischaemic stroke is the Oxfordshire Classification of Stroke Project classification (OCSP), initially described by Bamford et al. (1991) and summarised as:

1. **Total Anterior Circulation Stroke (TACS)**: Account for 20% of ischaemic strokes, and is caused by blockage of large vessels, leading to large volume infarcts.

2. **Partial Anterior Circulation Stroke (PACS):** Account for 35% of ischaemic strokes, and is typically caused by occlusions of branches of the middle cerebral artery. Characterised by the presence of two, of the three features of a TACS.

3. **Lacunar Syndrome(LACS):** Account for 20% of ischaemic strokes and is caused by occlusions of small perforating vessels. Characterised by pure motor or sensory loss, or ataxic hemiparesis.

4. **Posterior Circulation Stroke (POCS):** Account for 25% of ischaemic strokes and is caused by occlusions to vessels in the posterior circulation. Characterised by brainstem or cerebellar dysfunction, leading to more complex clinical presentations.

An alternative classification system for ischaemic stroke Adams Jr et al. (1993), the TOAST classification (Trial of Org 10172 in Acute Stroke Treatment). The system classifies stroke subtypes according to clinical features in conjunction with brain imaging, cardiac investigations and carotid artery imaging as follows :

1. **Large artery atherosclerosis :** Patients have clinical and brain imaging findings of significant stenosis or occlusion of a major brain artery, or branch cortical artery presumed secondary to atherosclerosis.

2. **Cardio-embolism:** Includes patients with arterial occlusions due to emboli from the heart. At least one cardio-embolic source must be identified.

3. **Small vessel occlusion :** Includes patients whose strokes are labelled as lacunar infarcts, and do not have higher cortical dysfunction, or potential cardiac sources of emboli, or large vessel disease. Patients have either normal brain imaging or small subcortical infarcts

4. **Stroke of other determined aetiology:** Includes patients with rare causes of stroke.

5. **Stroke of undetermined aetiology:** The underlying cause cannot be identified.

2.5 Incidence and Prevalence

2.5.1 Stroke

Approximately 113,000 people suffer a stroke in the UK annually and there are currently around 1 million stroke survivors (Rothwell et al., 2004). According to Bamford et al. (1991), there are 98,000 ischaemic strokes per 110,000 first ever-strokes annually. These numbers are expected to rise due to ageing population. Stroke incidence is projected to rise in the UK by 60% annually between 2015 and 2035 and the societal costs associated is expected to treble by then (King et al., 2020). The increasing number of survivors can also be accounted by improved treatment. Stroke does not only affect individuals and their close ones but there are substantial associated health and social cost implications.

2.5.2 Epidemiology of TIA

A first-ever TIA affects approximately 50 people per 100,000 of the UK population each year. Approximately 15% of ischaemic strokes are preceded by a TIA (Rothwell et al., 2007). It is estimated that 12% to 30% of patients report a history of TIA shortly before their stroke and approximately a quarter of them occur within hours before the stroke (Hackam et al., 2009; Rothwell et al., 2007). TIA represents a true medical emergency. If preventive treatment for TIA is started early (within 24 hours), the 90-day risk of stroke can be reduced by 80%. It is also estimated that around 10,000 recurrent strokes can be prevented every year in the UK through early initiation of treatment (Giles & Rothwell, 2007).

According to Turner (2016), estimating the incidence and prevalence of TIA is complicated by the change in the definition of TIA (time versus tissue-based definition) and the new proposed classification has not been universally adopted.

2.5.3 Global Stroke Burden

As previously stated, stroke is unquestionably a devastating disease, as it is the world's second-biggest cause of death, comprising approximately 10% of all deaths and being responsible for 5.5 million deaths each year, with 44 million disability-adjusted life-years (DALYs) lost (Mukherjee & Patil, 2011). Mukherjee and Patil (2011) also reported that in 2010 alone, there were 16.9 million strokes worldwide, of which 70 per cent occurred in low- and middle-income countries and the worrying tendency is likely to rise over the next 20 years. More than 85 per cent of the global stroke mortality hails from low-income and middle-income countries, especially from Africa and Asia. Undoubtedly, more than ever before, there is urgency for a worldwide emphasis on reducing the mortality and morbidity rates of cardiovascular disease and stroke.

The main barriers in many countries are documented as the dearth of infrastructure, insufficient systems of care, lack of effective plans to address cardiovascular risk factors, financial struggle and lack of trained health care workers (Kim & Johnston, 2011). According to Mukherjee and Patil (2011), the economic repercussions have also been very severe; in 2005, it was estimated that the losses to gross domestic product (GDP) due to vascular diseases were nearly $1 billion in China and India. Healthcare systems should be adequately funded, staffed, and structured in order to combat the stroke epidemic. The use of intravenous alteplase (IV tPA) following an acute ischaemic stroke, for example, has been shown to enhance outcomes. Patients treated with intravenous venous recombinant tissue plasminogen activator (tPA) within the first three hours of the onset of stroke had improved three-month functional outcomes compared to the control group, according to a clinical trial conducted by the National Institute of Neurological Disorders and Stroke (NINDS) rt-PA Stroke Study Group in 1995 (TNIoNDa, 1995).

Thrombolytic therapy is now widely used in suitable patients presenting early (<4.5 hours) from stroke onset and is of proven success in clinical practice. (Tanny et al., 2013)

According to NHS Choices (2017), alteplase is recommended soon after stroke onset, being the most effective treatment, however, it should not be used in the window of 4.5 hours that has elapsed. Before the alteplase therapy, it is recommended that a brain scan is carried out to rule out the possibility of haemorrhagic stroke as alteplase may then cause more harm than good as it may make bleeding worse. Nevertheless, the proper administration of IV tPA to suitable patients entails the organisation and availability of infrastructure. Though some countries like Brazil, Argentina, China and India have been effective at developing the systems for administration of IV tPA, low-income countries still face obstacles because of the scarcity of medical services and human resources and the medical services being geographically inaccessible at times. Another reason is that the IV tPA is exorbitantly costly.

Overall, the main risk factor is hypertension which is held responsible for approximately 54% of the global stroke burden. For the low-income and middle-income countries, stroke affects mostly the working-age (41 to 65 years) adults. Targeting risk factors of stroke at a primary care level, and focussing on healthy lifestyles, could substantially improve the global stroke burden.

2.6 Management of acute stroke and TIA

2.6.1 Stroke and Health Services

Stroke constitutes a major health crisis in the UK. The majority of stroke survivors are left with significant morbidity after stroke event(s). Eventually, the care for those patients requires 4–6% of the total NHS budget (Saka et al., 2009). Though there has been a recent indication of a decline in age-specific mortality from stroke, this does not necessarily translate to a decrease in morbidity or

any drop in need for services. In fact, with the population ageing, the incidence drop might be offset by the old-age-related health problems and required care.

An analysis of World Health Organisation (WHO) data shows UK general mortality is broadly similar to that in other Western European countries (Krishnamurthi et al., 2013). However, in terms of stroke case fatality, the UK tends to have higher numbers when compared to Western European countries (Sarti et al., 2000).

Thrombolysis

The first management for patients with ischaemic stroke requires rapid specialist assessment for the consideration of thrombolytic therapy, the administration of drugs called "lytics" or "clot busters" to dissolve blood clots that have acutely (suddenly) obstructed major arteries or veins, potentially causing serious or life-threatening consequences. It is often referred to as "thrombolysis". It is of utmost importance to initiate the therapy as soon as possible to prevent permanent damage. The main aim is to reestablish the normal cerebral blood flow to the affected area.

Thrombectomy

Mechanical thrombectomy (removal of the occlusive clot by endovascular means) is also often referred to as the "endovascular therapy for acute ischaemic stroke", or "endovascular clot removal". Only a small proportion of severe ischaemic strokes can be treated by this emergency procedure. It is only effective at treating stroke if a blood clot is lodged in a large artery in the brain. The sooner the procedure, the more effective the treatment. It involves inserting a catheter into an artery often in the groin and removing the clot using the device or suction, (NHS Choices, 2017).

More recently, mechanical thrombectomy was shown to be successful in clinical practice. It has beneficial outcomes in reducing death and disability and in achieving the normal flow of the obstructed vessel in acute ischaemic stroke. Findings originated from a trial of 459 patients followed up for over two years in the Netherlands (Van den Berg et al., 2017). They found that participants who received thrombectomy scored a median of three (slight disability) on the Modified Rankin Score (mRS) scale compared with four (moderate disability) in the standard group care. Thrombectomy was hence associated with an improved likelihood of a better functional score by two years (OR = 1.68, C.I: 1.21 to 2.30).

National guidelines were updated as a result of these findings. NICE (2008) guidelines endorsed that thrombolysis for acute ischaemic stroke should be administered only within a well-organised stroke service and should be given within the 4.5 hours window. Proper management should focus on the restoration of homeostasis (including BP control, and maintenance of hydration, nutrition, and normoglycaemia), prompt administration of antiplatelet agents, prevention of complications, and prompt early specialist rehabilitation. Bracard et al. (2016) guidelines recommended that the

mechanical clot removal should be performed by experienced clinicians with appropriate facilities. About 9,000 patients per year may be eligible for thrombectomy and it has been shown to reduce the chance of disability after stroke (Bracard et al., 2016).
In a pooled analysis from the Mechanical Embolus Removal in Cerebral Ischemia (MERCI) and different other trials, Nogueira et al. (2009) found that final revascularisation, baseline National Institute of Health Stroke Scale, age and systolic blood pressure were associated with good outcomes and survival at 90 days ($P < 0.0018$ for all). There are a few centres where thrombectomy is available in the UK, but there are not enough trained professionals for the service to be rolled out across the UK. Since the treatment is not fully commissioned yet, almost a third of hospitals have no access to thrombectomy either on-site or by referring to another hospital. However, due to the limited number of such interventions in stroke care, the current study will not study the factor.

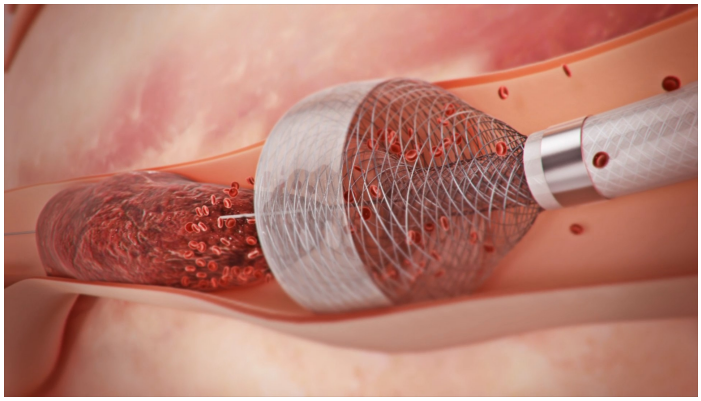


FIGURE 2.3: An Endovascular therapy.

Source: *Daviddatwood.com*

Carotid Endarterectomy Procedure

Narrowing of an artery in the neck called the carotid artery can also cause an ischaemic stroke resulting from a build-up of fatty plaques. If the carotid stenosis is particularly severe, surgery is recommended to unblock the artery. This is done by a surgical technique called a Carotid Endarterectomy Procedure.

It involves the surgeon making an incision and removing the fatty deposits. However, due to the limited number of such interventions in stroke care, the current survival study will not study the given factor.

Management of ICH

The management of patients suffering from acute ICH involves mainly medication, specialised stroke units where the principal aim of the acute treatment is the regularisation of the blood clotting, re-establishment of homeostasis, and the control of blood pressure and complications. Present

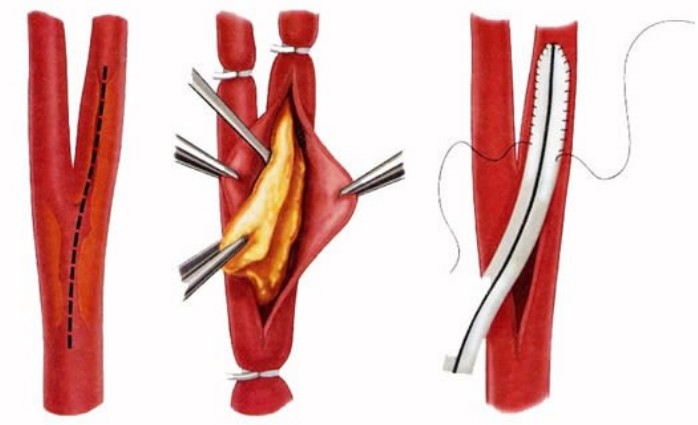


FIGURE 2.4: A Carotid Endarterectomy Procedure.

Source: coastalsurgery.com

guidelines mention that initial monitoring and management of ICH patients should take place in an intensive care unit or dedicated stroke unit with physician and nursing neuroscience acute care expertise (Manning et al., 2016).

Management of TIA

If a patient is suspected to have TIA, the management starts with a first assessment to identify the risk level of stroke, together with a quick start of antiplatelet therapy. The $ABCD_2$ score is a simple scoring system that can be used to stratify those patients who need urgent specialist assessment (within 24 hours) and those who need assessment within one week. Key steps in the specialist assessment include confirmation of TIA diagnosis; carotid imaging, and referral for surgery (<2 weeks) if indicated; identification of underlying atrial fibrillation (AF); prompt treatment of risk factors including hypertension and high cholesterol (NICE, 2008).

NICE guidelines : Treatment management

The pathway from National Institute for Health and Care Excellence (NICE) guidelines (NICE, 2020b) depicts the pathway of the management of stroke and TIA. The guidelines were updated in March 2017 and took into consideration of the following:

- Referral to specialist assessment and subsequent imaging in people with suspected TIA.
- Use of pharmacological or mechanical methods for clearing blood clots.
- Early antihypertensive treatment in Haemorrhagic stroke.
- Decompressive hemicraniectomy (surgery relieving massive brain swelling) in people older than 60 years.

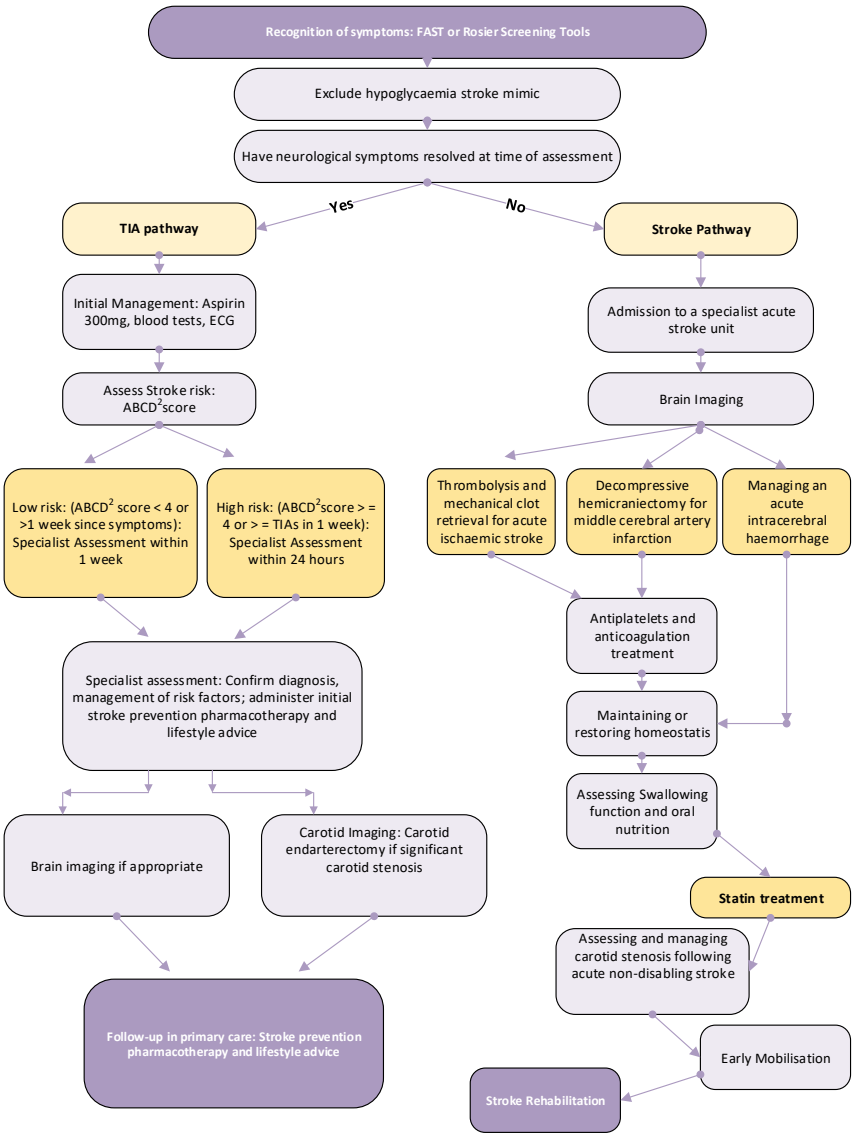


FIGURE 2.5: Treatment pathways for TIA and acute Stroke adapted from NICE guidelines

Source : NICE (2020b)

2.6.2 Stroke: tools and criteria

The Table 2.1 below shows different tools and criteria used in primary care for identifying stroke.

TABLE 2.1: Tools and criteria for stroke

Tools	Criteria
ABCD and ABCD2	Prognostic score to identify people at high risk of stroke after a TIA. It is calculated based on: **A** : Age ($\geq$ 60 years , 1 point) **B** : Blood pressure at presentation ($\geq$ 140/90 mmHg, 1 point) **C** : Clinical features (unilateral weakness, 2 points; speech disturbance without weakness, 1 point) **D**: Duration of symptoms ($\geq$ 60 minutes , 2 points; 10 to 59 minutes, 1 point) The calculation of ABCD2 also includes the presence of diabetes (1 point). Total scores range from 0 (low risk) to 7 (high risk).
FAST	**Face Arm Speech Test**: Used to screen for the diagnosis of stroke or TIA. **Facial weakness** : Can the person smile? Has their mouth or eye drooped? **Arm weakness** : Can the person raise both arms? **Speech problems** : Can the person speak clearly and understand what you say? **Test all three symptoms**.
ROSIER	Recognition of Stroke in the Emergency Room. Scale used to establish the diagnosis of stroke or TIA. Factors assessed include: demographic details, blood pressure and blood glucose concentration; items on loss of consciousness and seizure activity; and physical assessment including facial weakness, arm weakness, leg weakness, speech disturbance and visual field defects.

2.7 Incidence of recurrent stroke

Main studies of stroke have only focused on the incidence of the first stroke but study on the incidence of recurrent stroke is equally vital. According to National Institute of Neurological Disorders (2021), approximately 25% of people who recover from their initial stroke will have another stroke within five years. Recurrent stroke is the main contributor to stroke-related disability and death, with the risk of severe disability or death from stroke increasing with each additional recurrent stroke. The risk of a recurrent stroke is greatest right after a stroke; however, this risk will usually decrease with time. Within 30 days of their first stroke, about 3% of stroke patients will experience another stroke, and one-third of recurrent strokes occur within two years of the first stroke (National Institute of Neurological Disorders, 2021).

The Oxfordshire Stroke Community Project showed that the risk of suffering a recurrence within five years of a first stroke was 30% (95% CI: 20 to 39%)(Burn et al., 1994). The WHO MONICA project involved a younger population; aged 35-64 years. The team of researchers found that the recurrence of stroke was related to factors such as age and ethnicity (Truelsen et al., 2003).

2.8 Functional outcomes after Stroke

2.8.1 Medical complications after stroke

In Europe, strokes cause numerous cases of long-term adult disability (Truelsen et al., 2006) and the prevalence of disability due to stroke is anticipated to increase globally due to ageing population. Survivors of strokes are left with different degrees of residual incapacity that can endure for months and years (Johnston et al., 1999).

Post-effects of stroke can be motor or non-motor impairment (Sommerfeld et al., 2004). Motor impairment affects the control of the movement of the face, arm, and leg of one side of the body. Residual motor disability affects about 80% of patients to varying levels. Non-motor impairments include cognitive decline (including memory, executive functioning, attention, concentration, and alertness), low mood, and impaired communication abilities. This can largely affect the interaction, motivation, and continuation of normal activities. According to Edwardson et al. (2016), medical complications may result frequently from acute ischaemic stroke and include pneumonia, the need for intubation and mechanical ventilation, gastrointestinal bleeding, congestive heart failure, cardiac arrest, deep vein thrombosis, pulmonary embolism, and urinary tract infections.

According to Tei et al. (2000) and Siegler and Martin-Schild (2011), there is an association between early neurologic deterioration during the acute phase of ischaemic stroke and an increased risk of morbidity and mortality. Pohjasvaara et al. (2002) explains that stroke outcome may entail post-stroke depression. This is often attributable to post-stroke disability and cognitive impairment. It

is essential to monitor the stroke outcome on patient function, independence, and quality of life. Care needs and rehabilitation depend on the degree of disability and handicap post-stroke. Several clinical measures have been developed for clinical practice.

FIGURE 2.6: Barthel Index items

Source : Mahoney et al. (1965)

Barthel Index of Activities of Daily Living

Instructions: Choose the scoring point for the statement that most closely corresponds to the patient's current level of ability for each of the following 10 items. Record actual, not potential, functioning. Information can be obtained from the patient's self-report, from a separate party who is familiar with the patient's abilities (such as a relative), or from observation. Refer to the Guidelines section on the following page for detailed information on scoring and interpretation.

The Barthel Index

Bowels
0 = incontinent (or needs to be given enemata)
1 = occasional accident (once/week)
2 = continent
Patient's Score:_____

Bladder
0 = incontinent, or catheterized and unable to manage
1 = occasional accident (max. once per 24 hours)
2 = continent (for over 7 days)
Patient's Score: _____

Grooming
0 = needs help with personal care
1 = independent face/hair/teeth/shaving implements provided)
Patient's Score: _____

Toilet use
0=dependent
1=needs some help, but can do something alone
2=independent (on and off, dressing, wiping)
Patient's Score: _____

Feeding
0 = unable
1 = needs help cutting, spreading butter, etc.
2= independent (food provided within reach)
Patient's Score: _____

Transfer
0 = unable - no sitting balance
1 = major help (one or two people, physical), can sit
2 = minor help (verbal or physical)
3 = independent
Patient's Score:_____

Mobility
0 = immobile
1 = wheelchair independent, including corners, etc.
2 = walks with help of one person (verbal or physical)
3 = independent (but may use any aid, e.g., stick)
Patient's Score:_____

Dressing
0 = dependent
1 = needs help, but can do about half unaided
2 = independent (including buttons, zips, laces, etc.)
Patient's Score:_____

Stairs
0 = unable
1 = needs help (verbal, physical, carryin
2 = independent up and down
Patient's Score:_____
Bathing
0 = dependent
1 = independent (or in shower)

Patient's Score:_____

Total Score:_____

Scoring:
Sum the patient's scores for each item. Total possible scores range from 0-20 with lower scores indicating increased disability. If used to measure improvement after rehabilitation, changes of more than two points in the total score reflect a probable genuine change, and change on one item from fully dependent to independent is also likely to be reliable.

The Barthel Activities of Daily Living Index (Mahoney et al., 1965) is the most regularly used measure of disability. This gives a disability score from 0 (severe disability) to 20 (independent) and can be sub-divided into groups (Figure 2.6).

Another example of a scale that includes both disability and handicap is the Modified Rankin scale with six grades, from no symptoms to severe disability (Group et al., 1988). It can also be collapsed down to two levels. For instance, in the Oxford Community Stroke Project it was reduced to 'functionally independent' (grades 0 to 2) and 'functionally dependent' (grades 3 to 5)(Table 2.2).

TABLE 2.2: Example of disability measures: the Modified Rankin Scale

Source : Group et al. (1988)

Modified Rankin Scale	
0	No symptoms
1	Minor symptoms which do not interfere with lifestyle
2	Minor handicap: symptoms which lead to some restriction in lifestyle but do not interfere with the patient's,capacity to look after themselves
3	Moderate handicap: symptoms which significantly restrict lifestyle and prevent totally independent existence
4	Moderately severe handicap: symptoms which clearly prevent independent existence though not needing, constant attention
5	Severe handicap: totally dependent requiring constant attention night and day

2.8.2 Predicting recovery

Several studies, G. Hankey et al. (2007) and Jorgensen et al. (1995) amongst others explain that most of the recovery after stroke will happen in the first 3 to 6 months but some patients may still improve by the 18 months. Jorgensen et al. (1995) conducted a prospective study in Denmark on more than 1100 patients after acute stroke and found that patients who sustained mild disability tended to recover within two months and those who had moderate disability recovered within three months. Those with severe disabilities who recovered did so within four months, and those with the most severe disability within five months from onset. Neurological recovery would normally be resolved on average two weeks earlier than functional recovery.

A number of studies have tried to link survival after stroke with the functional outcome at certain time points. Kissela et al. (2009) used functional outcome at three months after stroke to predict survival at four years. Along the same lines, Slot et al. (2008) have suggested that long-term survival can be predicted by the functional outcome at 6 months; their studies were based on 7710 patients with ischaemic stroke, followed up for 19 years. They urged that early intervention should be warranted so as to have positive effects on survival.

Douiri et al. (2013) emphasised that cognitive impairment is one of the long-term impacts of stroke. They highlighted the importance of the effective preventive intervention and strategic planning that are needed in health systems, to identify and manage stroke survivors with cognitive impairments. A retrospective open cohort analysis of patients who had their first TIA and matched controls was conducted by Moran et al. (2015). Data was extracted from The Health Improvement Network (THIN) database and the results show that transient ischaemic stroke patients are more likely, compared with controls, to consult a GP for fatigue, psychological and cognitive impairments. The findings suggested that patients required more therapy beyond stroke prevention.

Y. Wang et al. (2003) developed a model that had good sensitivity and specificity for the prediction of survival for patients with acute IS. Their model included eight clinical predictors including dysphagia (swallowing problems) and urinary incontinence. Patients with urinary incontinence and dysphagia had a 7 and 9 times (respectively) greater risk of dying within 1 year after stroke compared to those who had no dysphagia and who were continent.
Weimar et al. (2002) collected 1754 patients' prospective records from the German Stroke Database to develop a prognostic model for functional independence. Functional independence was defined as a Barthel Index less or equal to 95 after 100 days. Their resulting models correctly classified more than 80% of the patients and found the following prognostic factors for functional independence : the patient's age, right and left arm paresis, gender, past stroke, diabetes, fever, and neurological problems, as well as the NIH Stroke Scale at admission and the Rankin Scale 48-72 hours later.

Age, right and left arm paresis at admission, NIH Stroke Scale at admission, Rankin scale 48-72 hours later, gender, prior stroke, diabetes, fever, neurological complications. The extent of recovery

might be relevant in explaining survival outcomes post-stroke but due to unavailability or poor recording of functional status after stroke, we were, unfortunately, unable to adjust the factor in our model.

2.8.3 Specific neurologic deficits

It can be challenging to predict recovery from neurological deficits. It requires an experienced neurologist to carefully examine and review appropriate neuro-imaging. The degree of improvement and the time course may differ for specific deficits, but undeniably, mild deficits improve more rapidly and more completely than severe deficits. Some of the deficits include: arm and hand weakness, leg weakness and ambulation, aphasia (a condition that affects the brain and leads to problems using language correctly, dysphagia (trouble in swallowing or paralysis of the throat muscles), sensory loss, visuo-spatial neglect (inability of a person to process and perceive visual stimuli)& depression and anxiety.

2.9 Risk factors of stroke and stroke mortality.

The numerous risk factors for stroke are discussed further. Some risk factors can be changed or medically treated, they are called modifiable risk factors, while some risk factors there is no control are called non-modifiable risk factors.

2.9.1 Non-Modifiable risk factors

Age

Age is the single major risk factor for stroke. After the age of 55, the risk of having stroke doubles every decade. According to Feigin et al. (2009), one in five women and one in six men will have a stroke after the age of 75.

In the United Kingdom, one among every four strokes occurs in people under the age of 65, and the number of adults aged 20 to 64 who had a stroke increased by 25% internationally between 1990 and 2010 (Stroke Association, 2016). The overall mortality after stroke rises steeply with age (Di Carlo et al., 2006; Fernandes et al., 2012; Jeng et al., 2008; Modrego et al., 2004) due to worsening of health with increasing age.

The mean age for stroke mostly cited in the literature was above 65 years (Garcıéa-Rodrıéguez et al., 2013; Weimar et al., 2002) amongst others. In terms of projection of short-term mortality, a study of comparison between Auckland and Perth populations, revealed that the risk of dying for the oldest patients (>= 85 years) within 30 days after stroke is reported to be 25% to 50% (Bonita et al., 1994) while the risk of dying in one year after stroke is reported to be ranging from 50% to more than 90% (Dennis et al., 1993). In other studies, we found that advancing age was associated with increased long-term mortality (>= 4 years) (Bates et al., 2014; Carter et al., 2007). Age has

certainly an influence on stroke and should therefore be considered an important factor for survival analysis.

Gender

Numerous studies have explored the influence of male or female sex on the incidence and survival after stroke. Most studies report that sex has no influence on survival (Barrett et al., 2007; Carter et al., 2007; Wolfe et al., 2005). Some studies have shown that survival for stroke was in favour of the females (Rutten-Jacobs et al., 2013). L. Goldstein et al. (2011), Mogensen et al. (2013), and Sacco et al. (1997) found that at any given age, males had a higher age-specific stroke incidence than women. Chang et al. (2010) who conducted a hospital-based study in Taiwan and found that men had 1.52 times more risk than females. In a study about the survival after one year after haemorrhagic stroke using the THIN database, Garcıéa-Rodrıéguez et al. (2013) reported that the 30-days fatality for men with ICH was two-fold that of women.

However, other studies revealed that women had more severe strokes (Andersen et al., 2005). In the Danish MONICA project, women had a higher risk for death than men after one year after stroke (Brønnum-Hansen et al., 2001). The disparity across different studies may be partly explained by the differences in the study design, populations and follow-up time.

Family history of stroke

The family history of stroke is often considered to be a predictor of stroke. Liao et al. (1997) found a higher risk of stroke among individuals with familial history compared to those without familial history of stroke. Increasing awareness of familial stroke may benefit preventing stroke. However, family history is poorly coded in the THIN database.

Ethnicity

Ethnicity is an important risk factor for the development of cerebrovascular diseases including stroke. Brown et al. (2013) study on 29,115 white patients and 1326 black patients found the black race was a risk factor for stroke, with stroke rates 1.6% for Whites and 2.5% for Blacks ($p = 0.009$). According to British Heart Foundation (2020), in the UK, south asian individuals are more likely to develop coronary heart disease than white Europeans and African or African Caribbean patients are more likely to develop a stroke than any other ethnic group. There is uncertainty about whether genetics plays a role in these differences.

2.9.2 Modifiable risk factors

Hypertension

Hypertension, also known as high blood pressure (BP) is one of the major risk factors for stroke and TIA, accounting for more than half of the ischaemic strokes. Stress on blood vessels, caused by high blood pressure can cause atherosclerosis (narrowing and hardening of blood vessels) which may result in an obstruction leading to ischaemic stroke or TIA or causes a blood vessel in the brain

to burst which leads to bleeding and haemorrhagic stroke. Globally, the overall prevalence of high blood pressure in adults aged 25 and over was around 40% in 2008 (Lawes et al., 2008). Numerous studies have found a consistent association between increasing BP and stroke. The higher the BP, the higher risk for stroke (Lawes et al., 2004).
The leading causes of hypertension are lifestyle, including smoking, alcohol intake, inactivity and diet. In most cases of hypertension, patients will have no symptoms. Furthermore, as people get older, high blood pressure becomes more common, and persons of south-asian and african-caribbean origin are more at risk than white people (British Heart Foundation, 2020). Anti-hypertensive medicines have been found in studies to successfully lower the risk of stroke. Anti-hypertensive medicines (such as angiotensin-converting enzyme (ACE) inhibitors, diuretics, and β-blockers), as well as lifestyle adjustments, are indicated for individuals with diagnosed hypertension in order to lower the risk of stroke and other vascular events.

Atrial fibrillation

Atrial fibrillation (AF) is to do with an irregular heart rhythm. AF is related to an approximately five times higher risk of stroke. Besides, patients with AF have a higher risk of mortality and disability compared to patients without AF. In AF patients, the risk of blood clots in the heart is increased by reduced blood flow due to the irregular heart rhythm. According to Rahman et al. (2014), the global prevalence of AF is 33.5 million and prevalence increases with age.

Aspirin has been found to reduce the risk of stroke in AF patients; however, anticoagulants were found to be more effective (Q. Zhang et al., 2015). According to Yao et al. (2016), anticoagulation medication with warfarin was found to reduce stroke risk more compared to lipid-lowering and antihypertensive drugs. New anticoagulant drugs are now available which are potentially safer (Yao et al., 2016).

High Cholesterol

Dyslipidaemia is characterised by the abnormal amount of lipids in the blood more formally quantified as high levels of total or low-density lipoprotein (LDL) cholesterol or low levels of high-density lipoprotein (HDL) cholesterol. The relationship between cholesterol levels and stroke is not straightforward; Weir et al. (1997) claimed that high levels of total cholesterol increase the risk for ischaemic stroke but protect against haemorrhagic stroke. There are contradictions in epidemiological studies regarding the association between lipid levels and stroke risk. Nevertheless, lipid-lowering drugs have been found to be effective at reducing stroke incidence. High cholesterol can be explained by lifestyle factors (especially diet) and also caused by familial hypercholesterolemia (a genetic condition).

Diabetes Mellitus

Patients with diabetes mellitus (DM), especially DM type II, have an increased susceptibility to

generalised atherosclerosis and often a clustering of other cardiovascular risk factors, such as hypertension, obesity, and increased levels of serum lipids. Previous studies have shown that diabetes independently increases the risk of stroke but there is a lack of evidence that intensive glycaemic control reduces the risk of stroke (Vasudevan et al., 2006). Compared to non-diabetic patients, diabetic patients have higher hazards of vascular events and deaths after stroke (Eriksson et al., 2012).

Smoking

Smoking causes 4 to 5 million premature deaths worldwide each year and the leading effects of smoking are cardio-vascular diseases in approximately 1.7 million people (McBride, 1992). The estimated risk of stroke among smokers is raised between 2 and 5 fold in comparison to non-smokers after adjustment for other risk factors (Hardie et al., 2003).

Alcohol consumption

Alcohol abuse can lead to a number of medical complications, including stroke. However, the relation between stroke risk and the amount of alcohol consumed is not straightforward (Hillborn, 1998). Mukamal et al. (2005)'s study on alcohol consumption and risk of stroke indicated that there was a positive relationship between heavy alcohol consumption and relative risk of stroke. However, they also found that light or moderate alcohol consumption may be protective against haemorrhagic stroke.

2.9.3 Stroke Severity

Stroke severity, i.e the amount of neurological deficits together with age, are perhaps the most consistent and powerful predictors of stroke survival. Most commonly used stroke severity scales in the literature is the NIHSS total score at admission (Barrett et al., 2007; Chang et al., 2010; Chang et al., 2006; Palm et al., 2014; Rutten-Jacobs et al., 2013; Tseng & Chang, 2006; Weimar et al., 2002) to quote a few.

Other studies on stroke used different stroke scales. The Glasgow Coma Score was used by Wolfe et al. (2005) for stroke patients in South London Register. Andersen et al. (2005) and Mogensen et al. (2013) used the Scandinavian Stroke Severity (SSS) in Denmark and reported that the severity of stroke has a strong influence on the stroke survival.

Several studies use the stroke subtypes as the predictors of survival and showed that stroke subtypes indeed had a strong significance (Carter et al., 2007; Chang et al., 2010; De Jong et al., 2003; Rutten-Jacobs et al., 2013). Two commonly used stroke subtype classifications are the TOAST and the Oxford Community Stroke Project (also known as the Bamford Classification) as explained in this chapter earlier in Section 2.4.

2.10 Stroke Treatments

Primary prevention is arguably the most effective approach to reducing the burden of stroke over time. Promotion of lifestyle changes such as increased physical activity, healthy diet, and smoking cessation are important, as are recent improvements in the treatment of associated conditions such as hypertension, diabetes, and atrial fibrillation.

2.10.1 Secondary Prevention

In patients with prior stroke, secondary prevention aims to reduce the risk of recurrent events and death. This includes both lifestyle changes and the use of medications Interventions (Secondary prevention) in people who have had a transient ischaemic attack or strokes) commonly used are summarised below:

- **Lifestyle advice**: Diet; smoking; alcohol; weight; exercise, where appropriate blood pressure reduction.
- **Antiplatelet agents**: aspirin and other agents such as dipyridamole and clopidogrel are also used.
- **Anticoagulation**: For patients with atrial fibrillation and those with certain types of valvular heart disease, and for those where the stroke was considered cardio-embolic in origin.
- **Statins** : For patients with known coronary heart disease.
- **Carotid endarterectomy**: For patients with significant (>70%) carotid artery stenosis on the same side.

2.10.2 Stroke units

Earlier studies have shown that patients with stroke have a better outcome when treated at a stroke unit (SU) compared to a general ward. The elements of a successful SU are a multidisciplinary team focused on stroke care, early management and mobilisation, treatment, rehabilitation, and continuous education of the staff (Rønning & Guldvog, 1998).

2.10.3 Drug prescriptions

Antihypertensive medication

Treatment of hypertension should be a high focus in both primary and secondary prevention. Both Systolic Blood Pressure (SBP) and Diastolic Blood Pressure (DBP) levels are linked to the risk of stroke; for example, lowering SBP by 10 mmHg reduces the risk of a first-time stroke by one-third (Perry Jr et al., 2000). The use of antihypertensive drugs such as ACE inhibitors and diuretics in primary prevention reduces the risk of stroke and other vascular events. In the secondary prevention

of stroke, similar effects have been seen. The Perindopril Protection Against Recurrent Stroke Study (PROGRESS) found that combining an ACE inhibitor (Perindopril) with a diuretic (Indapam) reduced the risk of recurrent stroke (PROGRESS, 1999).

Antiplatelet medication

Antiplatelet medications work by targeting platelets to prevent blood clots (thrombus). Antiplatelet medication (e.g. aspirin) has been shown to reduce the risk of stroke and other vascular events by roughly a quarter in patients who have had a previous stroke or TIA (22%) (G. J. Hankey & Warlow, 1999). Aspirin is a common and inexpensive medicine used in secondary stroke prophylaxis, lowering the risk of recurrent stroke while increasing the risk of haemorrhagic stroke. Dipyridamole in combination with other drugs has been shown to be more effective than monotherapy. However, it is less well tolerated by patients. Clopidogrel is comparable to aspirin, but it is used in combination with other medications.

Anticoagulation

Anticoagulation medicines actively prevent blood coagulation and have been shown to be more effective in preventing cardio-embolic stroke than antiplatelet therapy such as aspirin (Members et al., 2012). However, there is a higher chance of bleeding. As a result, anticoagulant medicines should only be given to patients who have a known cardio-embolic condition (e.g. atrial fibrillation). In randomised controlled trials, warfarin reduces the risk of stroke when compared to the placebo group, although it must be closely monitored.

Novel Oral Anticoagulation (NOAC) medicines have been shown to be beneficial in both primary and secondary prophylaxis, and patients treated with NOACS have a lower risk of bleeding and do not require anticoagulation therapy. (Larsen & Lip, 2014).

Lipid-regulating drugs

Lipid concentration such as LDL-cholesterol is strongly associated with the risk of AMI but the relation to stroke is more complex. The knowledge about the benefits of statins as secondary prevention against recurrent stroke is limited. In the Stroke Prevention by Aggressive Reduction in Cholesterol Levels trial (SPARCL) a high dose of atorvastatin in patients with recent stroke or TIA led to a 16 % reduced risk of recurrent stroke after a follow-up time of 4.9 years when compared to a placebo. However, a slight increase in haemorrhagic stroke was reported in the treatment group. Guidelines in secondary prevention recommend the use of statin in patients with prior stroke of known atherosclerotic origin (SPARCL et al., 2006).

2.11 Relative efficacy of antiplatelet interventions in stroke.

Antiplatelets are vital components in the management of non-cardioembolic IS and TIA. The most used antiplatelet options are aspirin, dipyridamole, clopidogrel, or the combinations with aspirin. While most of the RCTs have focused on the occurrence of vascular events, very few have focused

on the mortality outcomes (Table 2.3). NICE guidelines (NICE, 2010) were based on the review of the FASTER trial (Kennedy et al., 2007) which had a small size, hence study may have been underpowered to determine any statistical differences. Most classic trials had short-term follow-up and hence could not provide adequate evidence on the long-term prevention of vascular events. The primary endpoints of most of these trials have been on the occurrence of vascular events or the composite outcomes of IS, MI, vascular death, or re-hospitalisation (the MATCH, FASTER and POINT trials) with very few attempting to study the impact on mortality.

NICE (2021) guidelines recommend modified-release dipyridamole in combination with aspirin as secondary prevention of vascular events for TIA patients. NICE recommends clopidogrel for IS patients (NICE, 2010). The reviews were based on the FASTER and the CHANCE trials. However, both trials could not provide evidence of survival benefits of the antiplatelets due to their shorter follow-up of 90 days. The lack of evidence related to mortality on survival benefits of treatments leaves a lot of uncertainties. Clinical trial outcomes fail to translate into the benefits for patients in routine care (Heneghan et al., 2017). The strict inclusion-exclusion prevents most of the patients to be eligible, especially those who are old and multi-morbid. It is also difficult to ascertain benefits to the different age groups. Furthermore, the population used in the CHANCE trial was not generalisable to the UK population as it was conducted in China where the population risk factors and interventions differ.

TABLE 2.3: Important meta-analyses and trials of antiplatelets therapy in stroke and TIA patients.

Trials / Meta-analyses	**Population**	**Follow-up**	**Interventions**	**Outcomes**
CHARISMA	15,603 (CVD)	28 months	Clopidogrel + aspirin vs. aspirin	Reduction in outcomes(stroke +MI +vascular death) using the dual therapy.
PROFESS	20,322(IS)	2.5 years	Aspirin+Dipyridamole vs. clopidogrel	No significant difference in recurrence outcomes and vascular death but more bleeding risks using dual therapy.
CHANCE	5170 (TIA + IS)	90 days	Clopidogrel + aspirin vs. aspirin	No reduction in all-cause mortality, reduction in primary outcome using dual therapy but more bleeding risk using the dual therapy.
FASTER	392 (IS)	90 days	Clopidogrel + aspirin vs. aspirin	No difference in primary cause outcome(vascular events + death) but more bleeding with DAPT.
POINT	4881 (IS)	90 days	Clopidogrel + aspirin vs. aspirin	Lower risk of composite events in the dual therapy but more bleeding disorders compared to aspirin.
CAST	21,106 (IS)	1 month	Aspirin versus placebo	Fewer deaths in the aspirin group.
ESPRIT	2739 (TIA +IS)	3.5 years	Aspirin+Dipyridamole vs.aspirin	Composite death and bleeding events reduced by dual therapy.
ESPS-2	6602 (TIA+IS)	2.0 years	Aspirin,Dipyridamole,both vs.placebo	Most death reduced by dual therapy.
CAPRIE	19,185 (IS +MI+PAD) patients	1.9 years	Clopidogrel vs. aspirin	Risk reduction of vascular events and vascular deaths when using clopidogrel.
MATCH	7,599 (TIA+ IS)	2.5 years	Clopidogrel + aspirin vs. aspirin	More bleeding events for the dual therapy. No difference in survival benefit.

2.11.1 Limitation of studies : antiplatelets efficacy

The impacts of different antiplatelets on long-term mortality have not been clarified given short follow-up periods. The NICE reliance based on the 2 RCTs (The CHANCE and the FASTER trials) is questionable. So, it remains uncertain which antiplatelet option is the most effective and what is the impact on hazards of the long-term mortality. The uncertainty surrounding this issue therefore needs to be addressed. Hence, one of the aims of the study is to estimate the long-term survival benefits and harms of the different antiplatelet interventions in patients with and without a prior diagnosis of TIA and IS.

2.12 Review of Survival analysis after an Ischaemic stroke

Different survival studies on stroke are tabulated in Table 2.4. The table has details about different populations, follow-up periods, data collection and factors adjusted for in the models and the main outcomes of the studies. There were numerous studies that have examined the mortality after an ischaemic event, though most of them reported only mortality rates of stroke cases without a control group and only a few attempted case-control survival comparisons.
Findings from the MONICA project by Truelsen et al. (2003) showed that stroke mortality differed within different study populations from several countries. The tabulated studies (Table 2.4) reviewed populations from the United States, Chile, Taiwan, Australia and the European countries; UK, Denmark, Netherlands & Italy.
Most studies adjusted for age, previous TIA and diabetes mellitus. However, many of them did not adjust for further factors such as socio-demographic factors and different medical interventions. The current study adjusted for more diverse factors and included all second-order interactions of factors in the model.
The short-term mortality after IS was reported to be 16.1% at 1 month, 35.1% by 1 year and 60.2% at 5 years by Andersen and Olsen (2011). The annual case fatality was 4.8% more for stroke patients compared to the general population of the same age and sex (Hardie et al., 2003). In terms of hazards of mortality, the survival models in the stroke literature estimated the hazard ratio associated with the diagnosis of IS to be ranging from 2 to 5 times compared to controls (Brønnum-Hansen et al., 2001; Carter et al., 2007; Dennis et al., 1993; Hardie et al., 2003).

The follow-up period ranged from 1 year to 11 years. The only study with a long follow-up period was Gresham et al. (1998), with patients followed up for more than 20 years.

Data collection period spanned from 1982 at the earliest Brønnum-Hansen et al. (2001)to year 2018 Hagberg et al. (2019). Most of the studies were from the hospital and register data while the current study used the primary care data with more rich and detailed information on socio-demographic and lifestyle factors and generalisable to the routine clinical care across the UK.

One study included four hospitals in Leeds (Carter et al., 2007) . Two studies used data extracted from the THIN database (Garcıéa-Rodrıéguez et al., 2013; Gonzalez-Perez et al., 2013), however, the focus was more on the survival prospects after haemorrhagic type of stroke. None of the reviewed studies adjusted for any hospital/centre effects for possible survival variations between them. The recorded sample sizes were mostly reasonable, two studies had small sample sizes of n = 251 (Hardie et al., 2003) and n = 360 (Chang et al., 2010).

The majority of studies used the traditional Cox proportional Hazards model but none of them checked for any violations of the proportional hazards assumption. Time-varying effects are possible in the long-term follow-up studies (M. Zhang et al., 2018). Failure to address the time-dependent i.e, violating the Proportional hazards assumption may cause bias in estimates and any important effects might be missed, hence undermining the results of the research. We address this issue in our survival modelling.

TABLE 2.4: Review of studies on survival after IS stroke.

Study	Population	Data Collection Period	Follow-up Time	Sample Size	Factors Adjusted For	General Outcome
Bronnum-Hansen et al. (2001)	Patients aged 25 years or older in Copenhagen (Netherland) with first stroke	1982 to 1991	10 years	4162 with a first-ever stroke	Age group, Gender, Stroke Subtype	Those who had survived their initial stroke by 1 month had an almost 5 fold greater risk of dying within 1 year after the stroke than the general population.
Dennis et al. (1993)	Community-based stroke register (the Oxfordshire Community Stroke Project): Patients registered with stroke	1988 from OCSP	6.5 years	675 patients with a first-ever stroke	Age and Gender	During the first 30 days, 19% patients died, and survivors at 30 days after a first-ever stroke had 2-fold more risk than the general population. Stroke increased the relative risk of dying in younger patients.
Hannerz & Nielsen (2001)	Patients from the Swedish National Hospital Discharge registry	Jan 1989 to Nov 1993	5 years	103,591 patients		The life expectancy ratio in 1983 between stroke survivors and the general population was 0.571 for men and 0.578 for women.
Andersen et al. (2005)	Community-based patients in Copenhagen, Denmark	March 1992 to November 1993	10 years follow up	999 patients	Age, SSS, previous Stroke, other disabling diseases, diabetes, atrial fibrillation, sex	Women had more severe strokes. Adjusting for age, stroke severity, stroke type and risk factors, women had a higher probability of survival at 1 year (HR = 1.47); 5 years (HR = 1.47) and 10 years (HR = 1.49).

Table 2.4 – *Continued from previous page*

Study	Population	Data Collection Period	Follow-up Time	Sample Size	Factors Adjusted For	General Outcome
Andersen & Olsen (2011)	Community-based patients in Copenhagen, Denmark	March 1992 to November 1993	10 years	999 stroke patients	Age, SSS, previous Stroke, other disabling diseases, diabetes, atrial fibrillation, sex	Mortality rate was 16.6 % at 1 month, 35.1 % at 1 year, 60.2% at 5 years and 81.3 % at 10 years. Age, SSS, previous Stroke, other disabling diseases, diabetes, atrial fibrillation & sex were associated with survival at 1 month.
Mogensen et al. (2012)	Community-based patients in Copenhagen, Denmark	March 1992 to November 1993	10 years follow-up	998 patients with first-ever stroke	Age, gender, and stroke severity	Older age, male sex, greater stroke severity, and diabetes were all linked to death following a stroke.
Hoiffmeister et al.(2013)	Hospitalized patients in Chile	2003 to 2007	5 years	51,130 with first-ever stroke	Age, Sex, Geographical regions, and status insurance	The elderly patients (>80 years) HR = 4.07 and hospital admission in the North (HR = 1.14) and South (HR =1.06) were associated with lower survival after stroke. Patients with private insurance have a higher probability of survival after stroke than patients with public insurance.

Table 2.4 – *Continued from previous page*

Study	Population	Data Collection Period	Follow-up Time	Sample Size	Factors Adjusted For	General Outcome
Garcia- Rodriguez et al. (2013)	THIN data (patients aged 20 to 89 years)	Jan 2000 to Dec 2008	5.8 years	1797 ICH & 1340 SAH	Demographics and lifestyles factors, use of anti-platelet	Aspirin use was not associated with an increased risk of ICH but was associated with a decreased risk of SAH compared to no therapy. Warfarin use was associated with a greatly increased risk of ICH and a moderately increased risk of SAH compared to no therapy.
Andersen et al. (2014)	Danish Stroke Register: patients aged >40 years admitted to hospital for stroke in Denmark	2003 to 2012	9.5 years	56,581 patients	Age, income, level of education, sex, stroke severity, subtype, and a cardiovascular risk profile	People aged <65 years with basic education had a slightly higher risk for death than those with the highest education. The survival of patients with low income was reduced by 30% as compared to those with high income.
Belleudi et al. (2016)	Patients admitted to hospital, Lazio, Italy	2011 to 2012	1 year	9958 patients with ischaemic stroke	Age, Education level, Care pathway, Access to rehabilitation, Treatment post-discharge	Mortality was 14.9% in the acute phase and 14.3 % in the post-acute phase among those who survived the acute phase. The adjusted mortality in acute and post-acute phases decreased with an increase in educational level.

Table 2.4 – *Continued from previous page*

Study	Population	Data Collection Period	Follow-up Time	Sample Size	Factors Adjusted For	General Outcome
Hardie et al.(2003)	All individuals with a suspected acute stroke or transient ischaemic attack in the Perth community	Feb-89	10 years	251 patients	Age	Among 1-year survivors of stroke, the average annual case fatality was 4.8%, which was 2.3 times greater than for the general population of the same age and sex.
Wolfe et al.(2005)	Patients in the South London Stroke Register	January 1995 to December 2002	One year	2321 patients	Age and stratification for socio-economic status, type of stroke	After adjusting for various factors, Black ethnicity was a predictor of better survival with a hazard ratio of about 0.7 (compared to white patients). Gender had no effect. Current smoking, untreated atrial fibrillation, untreated diabetes and treated diabetes were associated with reduced survival.

Table 2.4 – *Continued from previous page*

Study	Population	Data Collection Period	Follow-up Time	Sample Size	Factors Adjusted For	General Outcome
Carter et al. (2007)	White Europeans recruited from 4 hospitals in Leeds	n.d	7.4 years	545 patients with ischaemic stroke matched with a cohort of 330 age matched healthy controls.	Age group, Gender, Stroke Subtype, Aspirin use, atrial fibrillation, previous stroke	Patients with ischaemic stroke were at more than 3-fold increased risk of death compared with the age-matched control cohort.
Gonzalez-Perez et al. (2013)	THIN data (patients aged 20 to 89 years)	Jan 2000 to Dec 2008	6.5 years	3137 cases of haemorrhagic stroke.	Age, gender	Patients who survived the initial 30 days after an ICH or SAH had a significantly increased risk of death compared with the control population both during the first year of follow-up and 1 year after the index event. In patients aged 20 to 49 years, the excess risk of death was almost 15-fold higher than that of the control population.

Table 2.4 – *Continued from previous page*

Study	Population	Data Collection Period	Follow-up Time	Sample Size	Factors Adjusted For	General Outcome
Rutten- Jacobs et al. (2013)	FUTURE study with patients aged 18 to 50 years, a a prospective cohort study of prognosis after transient ischaemic attack, ischaemic stroke or Haemorhagic stroke admitted to a medical Centre in the Netherlands	Jan 1980 to Nov 2010	11.1 years	959 patients	Age and Gender	At the end of the follow-up, 192 patients had died. For each stroke type, observed 20- year mortality among 30-day survivors exceeded expected mortality in the general population.
Chang et al. (2013)	Acute IS patients hospitalised in Taiwan	1998 to 1999	3 years	360 patients	Age and stroke severity	Age and initial stroke severity were strong predictors for 3-year survival after a first IS.

2.13 Review of survival analysis after TIA

Several studies of survival after TIA are described in Table 2.5. The survival outcomes after a TIA event varied across studies due to different treatments, methodologies, and other sources of variation.

One of the sources of variation was the source of data whether being hospital or community-based. Community services are critical in treating and managing acute and chronic illnesses, as well as assisting people in living independently. Nurses and other allied medical professionals visit patients at home and conduct personal risk assessments once they return home to the community, whereas hospital staff interactions are more regulated and contained. There exist some evidence that there is a "quality- gap" between hospital care and community care (Cumbler, 2015). Saltman et al. (2015) compared the survival of 973 patients with in-hospital stroke care with 28,837 community-onset stroke care. After adjustment of different risk factors in the model, they concluded that the mortality rate of 30 days and 1 year after stroke were similar for both cohorts. Nonetheless, they also noted that hospital patients had higher rates of comorbidities, were less likely to receive thrombolysis, and experienced more severe strokes compared to those in the community. The hospital-onset stroke patients were more likely to be disabled (with a modified Rankin scale of $3-6$) than their counterparts in the community-based (77% versus 65%).

A vast number of the studies reviewed were hospital-based which included: Daffertshofer et al. (2004), De Jong et al. (2003), Edwards et al. (2017a), Eriksson and Olsson (2001), S.-E. Eriksson (2017), Gattellari et al. (2012), Hankey et al. (1991), Johnston et al. (1999), and van Wijk et al. (2005). Several community-based studies included Anderson et al. (2004), Brønnum-Hansen et al. (2001), Burn et al. (1994), Coull et al. (2004), Hankey (2003), Hardie et al. (2003), and Jacob et al. (2020). Some studies used combinations of hospital and community-based Clark et al. (2003) and Carter et al. (2007).

Moreover, another source of variation was the sample size. The sample size plays an important role on the precision of estimates. Many studies are limited by small sample sizes. The sample size of the stuides on TIA can be classified as follows: mmall sample size (n = 100 to 200): Coull et al. (2004), Gresham et al. (1998), and Hagberg et al. (2019), medium sample size (n = 200 to 600): Andersen et al. (2005), Burn et al. (1994), Carter et al. (2007), Eriksson and Olsson (2001), Hankey et al. (1991), and Hardie et al. (2003) and large sample size (n from 900 to 20,000): Daffertshofer et al. (2004) ,Gattellari et al. (2012), Edwards et al. (2017a), De Jong et al. (2003) and Jacob et al. (2020).

The mean age at onset of TIA varied from 64 to 73 years. A vast majority of strokes had a mean age of 72 years. The younger mean ages were: 65 years (Gresham et al., 1998; van Wijk et al., 2005) and 64 years Carter et al. (2007).

Another possible source of variation reviewed was the varied definition of TIA in the different studies. These diverse definitions included:

- Patients with TIA symptoms from 1 hour to 72 hours (Hankey et al., 1991)
- First event De Jong et al. (2003), Dhamoon et al. (2009), Gattellari et al. (2012), Hankey (2003), and Jacob et al. (2020)
- Incident event /TIA at the presentation: Eriksson and Olsson (2001), Johnston et al. (2000), Sacco et al. (1994), and van Wijk et al. (2005)
- TIA patients surviving a certain day: Edwards et al. (2017a) considered TIA patients who did not experience adverse advents in the first 90 days.
- TIA or any stroke: Andersen et al. (2005), Burn et al. (1994), S.-E. Eriksson (2017), and Hardie et al. (2003)

Prospective studies usually have fewer potential sources of bias and confounding than retrospective studies. The majority of the studies were prospectively followed up; the few retrospective studies were Edwards et al. (2017a) and Eriksson and Olsson (2001).

Length of follow-up varied from 3 months to 6 months (Coull et al., 2004; Daffertshofer et al., 2004; Johnston et al., 2000) to the longest with 20 years or more (Anderson et al., 2004; Gresham et al., 1998). The recruitment period also played a great role in terms of reliability. Some studies were performed many years ago and hence have limited applicability. The current clinical management (for instance the initiation of antiplatelets to TIA patients at presentation) was not endorsed in the past. The oldest study of more than $>$ 20 years ago was Hankey et al. (1991) considered TIA diagnosed in years 1976-1986 and the most latest were Hagberg et al. (2019) observed patients diagnosed in 2007-2018; Edwardson et al. (2016) in 2003-2013; Jacob et al. (2020) in 2007-2016. No studies mentioned checking Proportional hazards assumptions.

TABLE 2.5: Review of studies on survival after TIA or minor stroke.

Author Country	Cohort	Sample Size	Age group	Length of Follow-up/ Years of follow-up	Survival Models	Covariates	Findings
Sacco et al. (1982) USA	Patients with initial stroke (Framingham Study). community-based.	$N = 394$	30 – 62 years	Up to 26 years Prospectively	Cox PH models	Age, gender	Survival was better for women than men. After 10 years, the cumulative survival = 35%.
Hankey et al. (1991) **United Kingdom**	First TIA with no prior history of stroke Hospital-based Oxfordshire Study	$N = 469$	≥ 18 years	Mean = 4.1 years 1976 – 1986 Prospectively			32 deaths observed, 5 years survival was 79.5%, 4.5% per year risk of death.

Continued on next page

Table 2.5 – *Continued from previous page*

Author Country	Cohort	Sample Size	Age group	Length of Follow-up/ Years of follow-up	Survival Models	Covariates	Findings
Sacco et al. (1994) **USA**	Patients with cerebral infarction. Population based The Northern Manhattan Study	$N = 323$	Mean age = 70 years	Mean = 3.3 years Prospectively	Cumulative life table and Cox PH Models.	Vascular disease, Heart Failure, Admission glucose, Basilar Syndrome	Cumulative risk of death: 30 days = 8%, 1 year = 22%, 5 years = 45% Basilar Syndrome (RR = 2.0)
Gresham et al. (1998) **USA**	Patients with initial stroke (Framingham Study) community-based	$N = 148$ patients matched to 148 age-sex matched controls.	Mean age = 65.7 years	20 years or more 1972 – 1974 Prospectively	Kaplan-Meier Survival curve		20 years after follow-up stroke survivors experienced greater mortality than age- and sex-matched controls (9.5%).

Continued on next page

Table 2.5 – *Continued from previous page*

Author Country	Cohort	Sample Size	Age group	Length of Follow-up/ Years of follow-up	Survival Models	Covariates	Findings
Johnston et al. (2000) **USA**	First TIA (admission in the emergency department) Hospital-based	$N = 1707$	Mean age = 72 years	Up to 90 days after admission 1997 – 1998 Prospectively	Multivariate Logistic regression, Kaplan-Meier lifetable	DBP, SBP, weakness, History on examination, Age, Diabetes Mellitus	Antiplatelets were associated with low risk of stroke. 2.6% died within 90 days. All HRs were non-significant.
Brønnum-Hansen et al. (2001) **Denmark**	First Stroke Register	$N = 4162$	$\geq$ 25 years Mean age = 68 years	5 years 1982 – 1991	Poisson distribution to estimate standardized Mortality Ratios (SMRs) Cox PH models	sex, Age at onset,	Mortality at 1 year: 41 %, at 5 years: 60%. Non-fatal stroke was associated 5-fold increase in risk of death from 1 month to 1 year after stroke.
Eriksson and Olsson (2001) **Sweden**	Stroke and other stroke subtypes admission to stroke unit. Hospital-based	$N = 339$	Median = 74	Up to 14 years Prospectively	Cox PH	Age, severity of stroke, previous stroke, SBP, Heart Failure, Fasting glucose, Diabetes Mellitus	21.5% patients died in the first year. The risk of death for stroke patients was 4.5 times higher than normal population of same age and gender.

Continued on next page

Table 2.5 – *Continued from previous page*

Author Country	Cohort	Sample Size	Age group	Length of Follow-up/ Years of follow-up	Survival Models	Covariates	Findings
Hartmann et al. (2001) **USA**	Patients with cerebral infarction Population based The Northern Manhattan Study	$N = 980$	Mean = 70 years	Mean = 3 years 1990 – 1992 Prospectively	Kaplan-Meier Survival curve Life-table Mortality Risk	Race, ethnicity, age	Cumulative mortality risk at 1 month: 5%, after 1 year: 16%, after 3 years: 29%, after 5 years: 41%.
Clark et al. (2003) **UK**	First TIA with no prior history of stroke community-based hospital-based	$N = 290$	Mean = 71.9 years	Median = 3.8 years 1981-1986	Kaplan-Meier Survival curve	History of heart disease.	10 years risk of death was 27.8%.

Continued on next page

Table 2.5 – *Continued from previous page*

Author Country	Cohort	Sample Size	Age group	Length of Follow-up/ Years of follow-up	Survival Models	Covariates	Findings
Hankey (2003) **Australia**	First Stroke (acute or TIA) community-based Perth Community Stroke Study (PCSS)	$N = 372$	Mean = 73 years	Up to 5 years 1989 – 1990 Prospectively	Kaplan-Meier product limit estimator Cox Proportional Hazard models	Intermittent claudification, urinary incontinence previous TIA, previous Barthel Index previous TIA, previous Barthel Index previous TIA(HR = 1.9), urinary incontinence (HR = 2.0), previous Barthel index (HR=2.0). Increasing age	First-ever stroke survivors associated with 4.2 times greater risk of dying compared to individuals of same age and sex. 1-year survivors continue to die over the next 4 years at a rate of 10% per year.
Hardie et al. (2003) Australia (PCSS)	First TIA or stroke community-based Prospectively Perth Community Stroke Study	$N = 251$	Mean = 72.7 years	Up to 10 years 1989 – 1990	Kaplan-Meier Survival curve Cox PH models		All patients with first-ever stroke were approx. 3 times greater risk of dying compared to same age/sex in general population.

Continued on next page

Table 2.5 – *Continued from previous page*

Author Country	Cohort	Sample Size	Age group	Length of Follow-up/ Years of follow-up	Survival Models	Covariates	Findings
De Jong et al. (2003) **Netherlands**	First-ever infarction (ischaemic stroke) Hospital-based	$N = 998$	Mean = 71 years	Mean = approx. 1.89 years 1987 – 1994 Prospectively	Kaplan-Meier Survival curve	Diabetes Mellitus, Age, Stroke subtype, initial stroke severity	Diabetes and age were predictors of 1-year mortality. Age: HR = 6.62 (3.88 – 11.29) Diabetes Mellitus: HR = 1.51 (0.99 – 2.32).
Daffertshofer et al. (2004) Germany	TIA patients Hospital-based	$N = 1380$	Mean = 71 years	Up to 6 months	Multivariate logistics regression	Age, modified Rankin scale, cardiogenic source of emboli, visible infarction on MRI	Mortality after TIA at 6 months = 5% Age $>$ 60 years: $OR = 2.8(1.0 - 5.0)$ Visible Infarction on CT scan: OR = 1.8 (1.1 to 3.1).
Anderson et al. (2004) **New Zealand**	First ever or recurrent stroke	$N = 680$ community-based Auckland Region Coronary or Stroke (ARCOS)	Mean = 70	Up to 21 years 1981 – 2002 Prospectively	Kaplan-Meier Survival estimates	Age, sex	Stroke patients had twice the mortality rate than the general population.

Continued on next page

Table 2.5 – *Continued from previous page*

Author Country	Cohort	Sample Size	Age group	Length of Follow-up/ Years of follow-up	Survival Models	Covariates	Findings
Coull et al. (2004) UK	First TIA (no prior history of stroke) community-based & hospital-based Oxford Community Project (OXCP)	$n = 87$ TIA and $n = 87$ minor stroke	Mean = 75 years	Up to 3 months 2002 – 2003 Prospectively	Kaplan-Meier Survival estimates		Outcome: risk of stroke Early risk of stroke after TIA/minor stroke is high: 8%, 11.5% and 17.3% at 1 week, one month and 3 months.
van Wijk et al. (2005) Netherlands Dutch TIA Trial (DTT)	Minor stroke or TIA Hospital-based	$N = 2473$	Mean = 65 years	Mean = 10.1 years 1986 – 1989 Prospectively	Cox PH models	Age, Diabetes, Claudification, Previous Vascular Surgery, ECG	After 10 years, 60% patients died Age>65 years: HR=3.33, Diabetes: 2.10, Claudification: HR=1.77, Previous Vascular surgery: HR= 1.94.
Carter et al. (2007) UK	ischaemic stroke patients	$N = 545$	Mean = 64 years hospital-based + community-based	Median = 7.4 years Prospectively	Cox Proportional Hazard models	Age, stroke, AF, previous stroke/TIA, Albumin/creatinine/ hemostatic factors	Patients with ischaemic stroke were at more than 3-fold increased risk of death compared with age-matched controls.

Continued on next page

Table 2.5 – *Continued from previous page*

Author Country	Cohort	Sample Size	Age group	Length of Follow-up/ Years of follow-up	Survival Models	Covariates	Findings
Dhamoon et al. (2009) USA	Patients with cerebral infarction The Northern Manhattan Study	$N = 524$ Population based	Mean = 68.6 years	Up to 5 years Prospectively	Cox Proportional Hazard models Multivariate generalised estimation equation (GFC) regression	Age, sex, race, ethnicity, education level, marital status, insurance status and medical conditions.	Outcome: factors predicting functional status
Gattellari et al. (2012) Australia	TIA patients (primary diagnosis) Hospital-based Program of Research Informing Stroke Management (PRISM)	$N = 22,157$	Median female = 78 years	Median = 4.1 years median male = 73 years 2000 – 2007 Prospectively	Cox Proportional Hazard models	Heart Failure, ischaemic Heart Disease, AF, smoking, diabetes mellitus, Hypertension, carotid stenosis.	10% of TIA patients died within 1 year, 13% lower survival at 5 years, 20% lower survival at 9 years Females higher relative survival (RR = 0.79) Increasing age increasing risk of death

Continued on next page

Table 2.5 – *Continued from previous page*

Author Country	Cohort	Sample Size	Age group	Length of Follow-up/ Years of follow-up	Survival Models	Covariates	Findings
S.-E. Eriksson (2017) Sweden	Stroke and other stroke subtypes admission to stroke unit	n = 288 men and n = 261 women Hospital-based	Median female = 70 years median male=64 years	Mean = 7.8 years 1986 – 2011 Retrospectively	Cox PH models	Age, history of diabetes, fasting glucose, TIA/severity of stroke, hypertension/ antihypertensive treatment, previous MI	Survival curves showed more risk of death for survivors. First line treatment beneficial on survival, Anticoagulant (OR=0.67), Antiplatelet (OR=0.67).
Edwards et al. (2017a) Canada	90 days stroke and other subtypes survivors Hospital-based	N = 26366 (15950 stroke and 10416 TIA)	Mean = 72	Up to 5 years 2003 – 2013 Retrospectively	Cox PH models		Stroke survivors had higher mortality/morbidity. HR at 1 year: 0.9 at 3 years: 1.3, at 5 years : 1.4.
Hagberg et al. (2019) Norway	First-ever stroke or TIA	n = 195 patients RCT CAST trial	Mean = 71.6 years	Up to 7 years 2007 – 2018 Prospectively	Cox PH models	Age, modified Rankin scale, hypercholesterolemia, AF, BMI.	Lower age: HR=1.08 Higher BMI: HR=0.91 Age and BMI were independent predictors for long term survival.

Continued on next page

Table 2.5 – *Continued from previous page*

Author Country	Cohort	Sample Size	Age group	Length of Follow-up/ Years of follow-up	Survival Models	Covariates	Findings
Jacob et al. (2020) Germany	Initial TIA diagnosis community-based Prospectively	$n = 19,824$	Mean = 68.8 years	Up to 10 years 2007-2016	Cox Proportional Hazard models	Age, gender, Diabetes, Hypertension, heart disease, AF, diuretics, beta blockers, heparin, antiplatelets,	Age 51-60 years positively associated with death (HR = 1.88), Hypertension (HR = 1.95), heart disease (HR = 1.59), AF(HR = 1.4), diuretics (HR = 0.73), Anticoagulants (HR = 0.49), antiplatelets (HR = 0.89).

2.14 Limitations of survival studies after IS and TIA.

Despite the debilitating consequences of stroke and TIA, the available data on the survival following these neurological events is quite sparse. The different stroke subtypes, source populations, different periods of recruitment, and the population type make the comparison of different reported risk estimates quite tricky. The existing studies were limited by differences in settings whereby some were carried out in acute care settings, stroke units, or clinical trial settings or settings which were not comparable to modern stroke treatment. Hospital-based studies are biased due to changes related to admission and referrals, hence distorting longitudinal trends. MONICA study by Truelsen et al. (2003) included younger than 75 years, thereby excluding half of all strokes. The small sample sizes of some identified studies could have given rise to inaccurate risk estimates.

Only approximately half of people whose stroke was preceded by a transient ischemic attack sought medical assistance for the transient ischaemic attack, according to the Oxfordshire Community Stroke Project by (Dennis et al., 1993). In the same way, some people who have minor strokes may never seek medical help (Kolominsky-Rabas et al., 2001). Furthermore, some studies' findings could not be generalised due to their special selection or strict inclusion-exclusion criteria. One example could be the selection of hospitalised stroke patients causing an underestimation of the survival rates by excluding the patients who did not get the care from the hospital setting. Furthermore, the different lengths of follow-up and varying degrees of adjustment may have given way to potential confounding factors. There was obviously a scarcity of long-term follow-up research. The question of how the diagnosis of TIA and IS impacts long-term mortality had to be addressed. In this thesis, the issue was addressed by investigating the long-term hazards of all-cause mortality following IS and TIA events while adjusting for a range of confounders and adjusting for any general practice effects. Our study was based on a large-scale population, was longitudinal and involved modern stroke care which was reflective of the routine care in the UK.

2.15 Life expectancy after TIA and stroke

In a statistical context, life expectancy is the average of all possible survival times (Robert et al., 2021). Amongst the stroke survivors, not only is rehabilitation a lifelong endeavour, but the sad fact is that the total life expectancy of the survivors can also be affected. The lower life expectancy can be accounted for by the lower quality of life of the stroke survivors, hence suggesting the need for not only better stroke prevention but also better stroke rehabilitation.

Medication's mixed effects, anxiety, financial constraints, and the impact on social life can all have an impact on one's quality of life. Stroke has been found in studies to significantly reduce a patient's ability to walk, do everyday chores, and care for personal requirements such as bathing, eating, and clothing. Interventions such as physiotherapy, occupational therapy, and speech and language therapy can enable patients to recuperate some of their "freedom", (Medicover, 2021).

Life tables have been commonly used by government, actuarial, and scientific institutions to determine the annual mortality rates (m) for each age x based on the general population. The impact of stroke on life expectancy is critical for stroke patients and their caregivers' counselling and financial planning.
In a hospital context, doctors frequently discuss cancer survival and relapse rates. They also consider other illnesses' mortality risks when making medical judgments. Stroke has received limited attention with regards to the life expectancy estimation until recently, Shavelle et al. (2019) 's study which provided the long-term life expectancy after stroke by age, sex, and severity of disability measured by the Modified Rankin Scale (mRs). Their results were based on a systematic review of 11 studies comprising 35,000 individuals. They showed how severity of disability can have a major impact on the long-term survival and suggested rehabilitation programs. The web-based survival tool was also made available. Hannerz and Nielsen (2001) also estimated age- and sex-specific life expectancies among individuals who have survived the acute phase (1 month) of a stroke within the study period from 1989 to 1993. Their results showed the improvement in life expectancy of stroke survivors (1989 versus 1993), 22.9% increase in men and 12.9% for women. Despite the improvement with time, the life expectancy of survivors was lower compared to the general population.

2.16 Conclusion

Numerous studies have looked at the survival prospects following a TIA or ischaemic stroke by estimating case fatality, short-term mortality, and long-term mortality. In terms of adjusted hazards of mortality associated with IS, Brønnum-Hansen et al. (2001), Carter et al. (2007), and Hardie et al. (2003) estimated HR ranging from 3 to 5. With regards to TIA, Edwards et al. (2017a), Gattellari et al. (2012), and Jacob et al. (2020) estimated HR ranging from 2 to 4. Because of their lack of adjustment for a number of risk factors and the fact that they were largely hospital-based or register-based, these estimates are likely to overestimate the mortality risk associated with IS and TIA. Our research, which used primary care data and was controlled for a range of confounders, could result in lower hazard estimates.

The survival following TIA and stroke was modelled using Cox's PH regression in the literature. The assumption checks, on the other hand, were rarely noted. When PH assumptions are violated, the findings can be biased. It is critical to double-check the PH assumption. If any violations of PH assumption are discovered, our study will conduct the relevant statistical tests and take the necessary steps to address them.

Age, sex, and stroke severity were mostly adjusted across most studies, as were clinical variables, ethnicity, functional level (at 3 months, 6 months, or 1 year), and sometimes CT results. The majority, however, considered gender and age. Some also adjusted for socioeconomic status, diabetes, hypertension, hypercholesterolaemia, atrial fibrillation, and drug therapy. The effects of

drugs were shown to have a mixed influence on survival chances. Eriksson and Olsson (2001) estimated HR=0.67 for both anticoagulants and antiplatelets and Jacob et al. (2020) estimated HR=0.67 and HR=0.89 for antiplatelets.
Male generally had worse survival prospects (Andersen et al., 2005; Gattellari et al., 2012; Hannerz & Nielsen, 2001; Mogensen et al., 2013). Wolfe et al. (2005) found no gender effect. The absence of long-term survival studies after TIA and stroke in UK primary care, as well as research on the efficacy of pharmaceutical therapies, were recognised as research gaps. The elderly and those with comorbidities are frequently excluded from medicinal trials. Investigating the effects of antiplatelet treatments in a “real-world" situation would be interesting. More research towards more effective therapeutic choices is essential.

Given the paucity of adequate studies and the limitations of the existing ones, large-scale population research is important. Hence, our study intends to fill in the knowledge gaps identified. The next chapters endeavour to achieve the aims and objectives of the study.

3 Methods

The chapter presents statistical models that were used to estimate the adjusted hazards of all-cause mortality after stroke. Some important diagnostic tests are also explained. The Multiple Imputation technique is described.

3.1 Introduction

Survival data is a term used for describing data that measures the time to a certain event. In survival analysis, the event may be death, the occurrence of disease (or complication) such as time to an epileptic seizure, the time it takes for a patient to respond to therapy, or time from response until disease relapse (Hanagal, 2011). Models of survival analysis examine the relationship between the survival times and two or more predictors, usually termed as *covariates* (Fox & Weinsberg, 2011).

Time to an event is a positive real-valued random variable having continuous distribution. It is necessary to define the starting time point, say 0, from which times are measured. The time variable T is always positive and commonly skewed (with a large number of events observed either at the beginning or at the end of follow-up).

When we measure age, the starting time point may be the date of birth. For studying the occurrence of a disease, the time scale is the known duration of the disease. In our study, the time variable is the time since diagnosis of stroke or TIA. The follow-up of stroke patients starts on the date of the first occurrence of the stroke diagnosis and ends on the date of death, transfer out of the practice, or the closing date of follow-up (9^{th} Jan 2017). The time for the controls is measured from the time of entry to the death date, transfer date, or end of the study end date.

3.2 Survival Analysis

Let T be a non-negative continuous random variable, representing the time until the event of interest. We denote :

$$F(t) = P(T \leq t) : \text{distribution function;}$$
$$f(t) : \text{probability density function.}$$

For survival data, we consider rather

$$S(t) : \text{survival function,}$$

$H(t)$: cumulative hazard function,
$h(t)$: hazard function.

The survival function is defined at any time t, as the probability of surviving beyond time, t,

$$S(t) = P(T > t) = 1 - F(t). \tag{3.1}$$

The survival function is the probability that the event of interest will happen after time t. It is a decreasing function, taking values in [0,1]. It equals 1 at t = 0 and 0 at $t = \infty$. Hazard function (or hazard rate) is defined as :

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}. \tag{3.2}$$

The *hazard function* represents the instantaneous failure rate and is defined as the probability that the event occurs in a specific time interval Δt. The hazard function, $h(t)$ measures the instantaneous risk of dying right after time t given the individual is alive at time t. It is a positive function (not necessarily increasing or decreasing). The hazard function $h(t)$ can have many different shapes and is therefore a useful tool to summarize survival data.

Cumulative hazard function, $H(t)$ is defined as :

$$H(t) = \int_0^t h(u)\, \mathrm{d}u. \tag{3.3}$$

$H(t)$ is an increasing function, taking values in $[0, \infty]$. It is related to $S(t)$ as:

$$H(t) = -\log S(t). \tag{3.4}$$

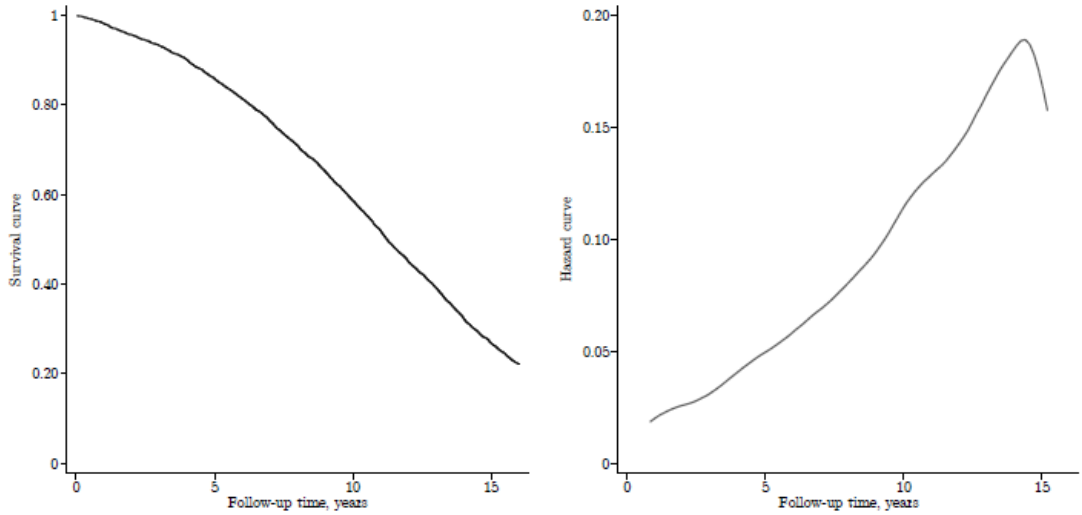
Hence,

$$S(t) = \exp(-H(t)). \tag{3.5}$$

The survival function is a monotonic non-increasing function of time. Typical examples of the hazards and the survival functions are given in Figure 3.5. The survival function is regarded as the most favourable tool to incorporate the *risk* of the event and *time* at which risk occurs. Statistical estimators have been developed to find the time-to-event outcomes and to estimate the survival and the hazard functions. In the later section, we will present the Kaplan-Meier estimator (Kaplan & Meier, 1958) and the Cox regression (Cox, 1972) which are the two most famous methods used in epidemiology and medical research.[1]

[1] The introductory papers of these two methods are listed among the 100 most-cited papers of all time and occupy the first two positions in the field of statistics (Van Noorden et al., 2014).

FIGURE 3.1: An illustrative example of the Survival and Hazard curves of the study population during 16 years of follow-up



From the Equation (3.4) and Equation (3.5), we can further express the hazard function ($h(t)$) of Equation 3.2 as :

$$h(t) = \frac{1}{P(T \geq t)} \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t)}{\Delta t} \tag{3.6}$$

$$= \frac{f(t)}{S(t)} \tag{3.7}$$

$$= \frac{-d}{dt} \log S(t) \tag{3.8}$$

$$= \frac{d}{dt} H(t). \tag{3.9}$$

Since $h(t)$ is also equal to the negative of the derivative of $\log S(t)$, we have the useful relationship (D. W. Hosmer & Lemeshow, 1999) and (Kalbfleisch & Prentice, 2011):

$$S(t) = \exp\left\{ -\int_0^t h(u)\, \mathrm{d}u \right\}. \tag{3.10}$$

Hence, the survival function, $S(t)$, expressed in terms of the hazard function is given by :

$$S(t) = e^{-H(t)}. \tag{3.11}$$

3.3 Censoring and Truncation

According to Hanagal (2011), in biomedical applications data, the 'time to event' may not be observed for all the individuals in the study population (sample) due to the period of study being finite. This phenomenon when we fail to observe the endpoints of the event of interest is called *censoring*. The combination of censoring and differential follow-up creates some unusual difficulties in the analysis of such data that cannot be handled properly by the standard statistical methods and survival analysis is the appropriate framework developed for this purpose.
A common feature of epidemiological or medical data is that they contain either censored or truncated observations. Censored data arises when an individual's time to event is known to occur only in a certain period of time. Well-known censoring schemes are right censoring, where all that is known is that the individual is still alive up to a given time; left censoring, when all events of interest prior to the study's start date; or interval censoring, in which the only information available is that the event occurred within a certain time interval. Left truncation, in which only those who live long enough are included in the sample, and right truncation, where the sample only includes people who have experienced the event by a certain date are two truncation schemes.

3.3.1 Right Censoring

Only a lower bound for the time of interest is known:

$$\begin{gathered} T = \text{survival time}, \\ C = \text{censoring time}, \\ \Rightarrow \text{Data} : (Y, \delta) \text{ with}, \\ Y = \min(T, C) , \\ \delta = \mathrm{I}(T \leq C). \end{gathered}$$

There are three types of right censoring:

- Type I right censoring: All subjects are followed for a fixed amount of time. All censored subjects have the same censoring time.
- Type II right censoring: All subjects start to be followed up at the same time and follow-up continues until r individuals have experienced the event of interest (r is some pre-determined integer).
  The n-r censored items all have a censoring time equal to the failure time of the r^{th} item.
- Random right censoring: The study itself continues until a fixed time point but subjects enter and leave the study at different times. Censoring is a random variable. e.g. random right censoring in a cancer clinical trial.

3.3.2 Left Censoring

Some subjects have already experienced the event of interest at the time they enter in the trial. Only an upper bound for the time of interest is known

$$\begin{gathered}\text{Data} : (Y_\ell, \delta_\ell) \text{ with} \\ Y_\ell = \max(T, C_\ell), \\ \delta_\ell = I(T > C_\ell), \\ C_\ell = \text{censoring time.}\end{gathered}$$

A malaria trial is an example of left censoring. Malaria is monitored in children aged two to ten years. Children who have had malaria will have antibodies against the Plasmodium parasite in their blood. Children as young as two years old may have already been exposed to the parasite (Kifle, 2013).

3.3.3 Interval Censoring

The event of interest is only known to occur within a certain interval (L, U). Contrary to right and left censoring, we never observe the exact survival time. Interval censoring typically occurs if diagnostic tests are used to assess the event of interest e.g. interval censoring in malaria trial: The exact time to malaria is between the last negative and the first positive test.

3.4 Assumption of uninformative censoring

For the current study, censoring could occur, for example, patients could be lost during follow-up, or the non-occurrence of the outcome event before the study end. Survival data are frequently censored or missing in some way, which is a crucial feature that sets survival analysis apart from other areas of statistics. According to D. W. Hosmer and Lemeshow (1999), uninformative censoring implies that the distribution of time-to-event and the distribution of time-to-censorship do not inform each other. Serious disease progression with reduced survival time such as lung cancer may result in informative censoring that is the censoring provides more information than the fact that survival time exceeded a certain time. Informative censoring may occur for instance, when participants are lost to follow-up for reasons relating to the study, such as when comparing disease-free survival after two cancer treatments, the control arm may be unsuccessful, resulting in more recurrences and patients being too ill to follow-up. If analysis of such data does not take into consideration the information, then this may cause bias in the estimation of the survival time.

The data used for the current research was subject to right-censoring as patients were either lost to follow-up because of moving to other GP practices or being alive at the end of the study period. Furthermore, censoring was assumed to be random, that is, not related to the survival times. We

assumed that the health status and the lifestyle factors of patients were not associated with the transfers to other GP practices. We, therefore, assumed uninformative censoring. Studies have shown that mortality might differ for people moving from deprived to affluent areas, whereby selective migration was associated with lower mortality and better health for younger groups (Maheswaran et al., 2018). Hence, for the current study, it was vital to analyze any distinct pattern of the patients who transferred to another general practice with their health status (diagnosis of TIA and/or stroke) and social deprivation factors. This is addressed in section 5.3.

3.5 Parametric survival models

In parametric survival model, we can completely specify $h(t)$ and $S(t)$ and it is possible to do time-quantile prediction. However, we need to assume an underlying distribution. The following survival distributions are those which are more commonly used in survival analysis. A number of parametric models have successfully served as population models for failure times. Eight most popular parametric models are as follows:

- Exponential
- Gamma
- Weibull
- Extreme Value
- Lognormal
- Log-logistic
- Gompertz distribution
- Gompertz-Makeham distribution

3.5.1 The Exponential distribution

The exponential model, with only one unknown parameter, is the simplest of all life distribution models. The density function is given by:

$$f(t) = \lambda \exp(-\lambda t), \tag{3.12}$$

and other key functions are presented as :

$$F(t) = 1 - \exp(-\lambda t); \tag{3.13}$$

$$S(t) = \exp(-\lambda t); \tag{3.14}$$

$$h(t) = \lambda. \tag{3.15}$$

At any time t, the failure rate reduces to the constant λ. The exponential distribution is the only distribution to have a constant failure rate. An important property of the exponential distribution is the 'loss of memory'(memory-loss) property. The memory-loss property implies that a given probability distribution is independent of its history.

3.5.2 The Gamma distribution

The 2-parameters gamma distribution, which is denoted $G(\alpha, \lambda)$, can be viewed as a generalisation of the exponential distribution. The important functions are:

$$f(t) = \frac{\lambda^{\alpha} t^{\alpha-1} \exp(-\lambda t)}{\Gamma(\alpha)}, \tag{3.16}$$

where parameters $\lambda > 0$ and $\alpha > 0$ and

$$\Gamma(\alpha) = \int_0^{\infty} t^{\alpha-1} \exp(-t) dt, \tag{3.17}$$

$$E(T) = \frac{\alpha}{\lambda}, \tag{3.18}$$

$$Var(T) = \frac{\alpha}{\lambda^2}, \tag{3.19}$$

$$G(1, \lambda) = Exp(\lambda). \tag{3.20}$$

When α is 1, the gamma distribution $G(\alpha, \lambda)$ of Equation (3.16) reduces to $\lambda \exp(-\lambda t)$ which is equivalent to an exponential distribution in Equation (3.20).

3.5.3 The Weibull distribution

The Weibull distribution is also generalisation of the exponential distribution. It is a very flexible life distribution with two parameters. It is related to the extreme-value distribution. It is popular in demographic applications but for mortality studies, it is wise to avoid it for old age mortality (the hazards grows too slow) and mortality in ages $0-15$ (U-shaped hazards, which the Weibull model does not allow), (Broström, 2012). The probability density function is given as :

$$f(t; \alpha, \lambda) = \frac{\alpha}{\lambda}\left(\frac{t}{\lambda}\right)^{\alpha-1} \exp\left(-\left(\frac{t}{\lambda}\right)^{\alpha}\right), \qquad t, \alpha, \lambda > 0 \tag{3.21}$$

The hazard rate of a Weibull distribution is :

$$h(t; \alpha, \lambda) = \frac{\alpha}{\lambda}\left(\frac{t}{\lambda}\right)^{\alpha-1}, \qquad t, \alpha, \lambda > 0 \tag{3.22}$$

where α is a a shape parameter and λ is a scale parameter. When $\alpha = 1$, the Equation (3.21) reduces to $h(t;1,\lambda) = \frac{1}{\lambda}$ which is the exponential distribution with rate $\frac{1}{\lambda}$.

3.5.4 Extreme value Distribution

The Weibull distribution and the extreme value distribution have a useful mathematical relationship. If $t_1, t_2, \ldots, t_n$ are a sample of random failure times from a Weibull distribution, then $\ln t_1, \ln t_2, \ldots, \ln t_n$ are random observations from the extreme value distribution. In other words, the natural log of a Weibull random time is an extreme value random observation, Hanagal (2011). The distribution is often referred to as the Extreme Value Distribution (Type I).

3.5.5 The log-normal distribution

The log-normal life distribution, like the Weibull, is a very flexible model that can empirically fit many types of failure data. The two parameter form has parameters σ = the shape parameter and T_{50} = the median (a scale parameter). If time to failure, t_f, has a log-normal distribution, then the (natural) logarithm of time to failure has a normal distribution with mean $\mu = \ln T_{50}$ and standard deviation σ.

3.5.6 The log-logistic distribution

The log-logistic distribution is very close to the log-normal, has a heavier right tail. Its advantage over the log-normal is that the hazard function has a closed form. It is described as :

$$h(t;(p,\lambda)) = \frac{\frac{p}{\lambda}\left(\frac{t}{\lambda}\right)^{p-1}}{1+\left(\frac{t}{\lambda}\right)^{p}}, \qquad t,p,\lambda > 0. \tag{3.23}$$

The hazard function for a log-logistic distribution with shape $p = 2$ and scale = 1 is shown in figure 3.2.

3.5.7 Other useful survival distributions

- **The Gompertz Distribution**
  The hazard function of the Gompertz distribution Gompertz (1825) is given by :

$$h(t) = \tau \exp(t/\lambda), \qquad t \geq 0; \qquad \tau,\lambda > 0 \tag{3.24}$$

  The hazard function is exponentially increasing. The Gompertz distribution is useful for modelling old-age mortality as the mortality rates grow exponentially with age.

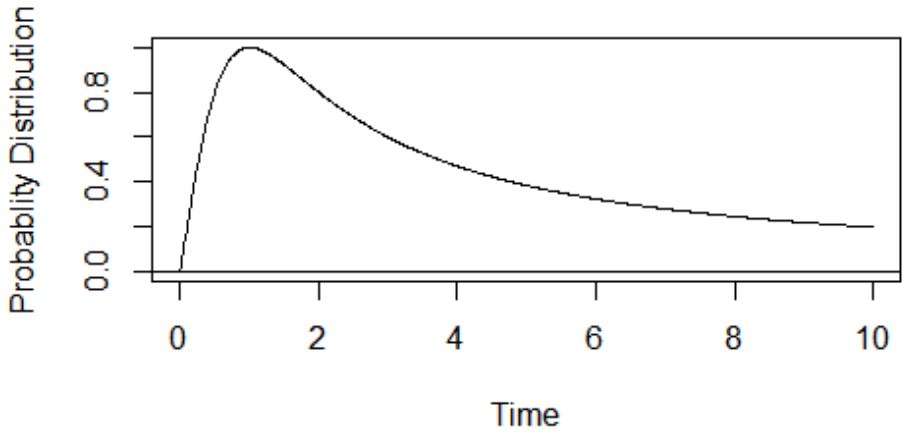


FIGURE 3.2: Log-logistic hazard function with shape 2 and scale 1.

- **The Gompertz-Makeham Distribution**
  The Gompertz distribution was generalised by Makeham (1860). The generalisation consists of adding an extra positive parameter to the Gompertz hazard function to account for mortality that is not related to ageing,

$$h(t) = \alpha + p\exp(t/\lambda), \qquad t, \alpha, p, \lambda > 0 \tag{3.25}$$

  Gompertz-Makeham distribution can be used to model mortality from the age window of 30 to 80 years of age.

3.6 Non-Parametric methods of Survival Analysis

Non-parametric analysis involves the analysis of survival data without parametric assumptions about the form of the distribution (Rodrıguez, 2005).

3.6.1 Kaplan-Meier Estimator

Let

$$t_{(1)} < t_{(2)} < \ldots < t_{(m)}$$

denote the distinct *ordered* times of death (not counting censoring times). Let d_i be the number of deaths at $t_{(i)}$, and let n_i be the number alive *just* before $t_{(i)}$. This is the number exposed to risk at time $t_{(i)}$. Then the Kaplan-Meier or *product limit estimator* of the survivor function is :

$$\hat{S}(t) = \prod_{i:t_{(i)}<t} \left(1 - \frac{d_i}{n_i}\right). \tag{3.26}$$

If we assume that data are not censored, then we can estimate the survival function $S(t)$ non-parametrically by the following empirical survival function

$$\hat{S}(t) = \frac{1}{n} \sum_{i=1}^{n} I\{t_i > t\}, \tag{3.27}$$

where I is the indicator function that takes the value 1 if the condition in braces is true and 0 otherwise. The estimator is simply the proportion alive at t.

Kaplan and Meier (1958) extended the estimate to censored data.

A simple justification is as follows: To survive to time t, a subject first survives to $t_{(1)}$. He or she must then survive from $t_{(1)}$ to $t_{(2)}$ given that he or she has already survived to $t_{(1)}$. If there are no deaths between $t_{(i-1)}$ and $t_{(i)}$, we take the probability of dying between these times to be zero.
The conditional probability of dying at $t_{(i)}$, given that the subject was alive just before, can be estimated by $\frac{d_i}{n_i}$. The conditional probability of surviving time $t_{(i)}$ is the complement $1 - \frac{d_i}{n_i}$. Hence, the overall unconditional probability of surviving to t is obtained by multiplying the conditional probabilities for all relevant times up to t (Rodrıguez, 2005). The Kaplan-Meier estimator is a step function with discontinuities or jumps if death times are observed.
The Kaplan-Meier Estimator is generally used to estimate the observed survival experience of the entire population or it can be stratified by different sub-populations.

3.6.2 Kaplan-Meier Estimator as the Maximum Likelihood Estimator

We can now describe the Kaplan-Meier estimator as the Maximum Likelihood Estimator (MLE). Likelihood provides a natural way to proceed with inference in the presence of censoring. We define c_i as the number of cases censored between $t_{(i)}$ and $t_{(i+1)}$, and d_i as the number of cases that die by $t_{(i)}$.
Thus, the likelihood function takes the form (Rodrıguez, 2005) of

$$L = \prod_{i=1}^{m} \mathbb{P}\left(T_i = t_i\right)^{d_i} \mathbb{P}\left(T > t_i\right)^{c_i}, \tag{3.28}$$

or equivalently

$$L = \prod_{i=1}^{m} \left[S\left(t_{(i-1)}\right) - S\left(t_{(i)}\right)\right]^{d_i} S\left(t_{(i)}\right)^{c_{(i)}}, \tag{3.29}$$

where the product is over the m distinct times of death, and we take $t_{(0)} = 0$ with $S(t_{(0)}) = 1$. If we define $\pi_i = S(t_{(i)})/S(t_{(i-1)})$ for the conditional probability of surviving from $S(t_{(i-1)})$ to $S(t_{(i)})$. Then we can write

$$S(t_{(i)}) = \pi_1 \pi_2 \cdots \pi_i,$$

and the likelihood becomes

$$L = \prod_{i=1}^{m} (1-\pi_i)^{d_i} \pi_i^{c_i} (\pi_1 \pi_2 \cdots \pi_{i-1})^{d_i+c_i}. \tag{3.30}$$

We should note that all cases who die at $t_{(i)}$ or are censored between $t_{(i)}$ and $t_{(i+1)}$ contribute a term π_j to each of the previous times of death from $t_{(1)}$ to $t_{(i-1)}$. In addition, those who die at $t_{(i)}$ contribute $1-\pi_i$, and the censored cases contribute an additional π_i. Let $n_i = \sum_{j \geq i} (d_j + c_j)$ denote the total number exposed to risk at $t_{(i)}$. We can then collect terms on each π_i and write the likelihood as

$$L = \prod_{i=1}^{m} (1-\pi_i)^{d_i} \pi_i^{n_i - d_i}, \tag{3.31}$$

which is a binomial likelihood. The MLE. of π_i is then

$$\hat{\pi}_i = \frac{n_i - d_i}{n_i} = 1 - \frac{d_i}{n_i}. \tag{3.32}$$

Thus the joint likelihood of π_i yields the Kaplan-Meier estimator by multiplying these conditional probabilities as per Equation (3.26), Rodrıguez (2005) .

After obtaining, the Kaplan-Meier estimator, we need to calculate its standard error. The purpose is to determine an approximate 95% confidence intervals for $\hat{S}(t)$. We now describe two ways to estimate the standard errors.

3.7 Greenwood's Formula

The main idea to approximate the variance of the Kaplan-Meier estimator is by using the "delta-method" approximation, Greenwood (1926). From the likelihood function Equation (3.31), it follows that the large sample variance of $\hat{\pi}_i$ conditional on the data n_i and d_i is given by the usual binomial formula :

$$Var(\hat{\pi}_i) = \frac{\pi_i(1-\pi_i)}{n_i}. \tag{3.33}$$

$\mathrm{Cov}(\hat{\pi}_i, \hat{\pi}_j) = 0$ for $i \neq j$, so the covariances of the contributions from different times of death are all zero. To obtain the large sample variance of $\hat{S}(t)$, the Kaplan-Meier estimate of the survival

function, we need to apply the delta method twice. First, we take logs, so that instead of the variance of a product we can find the variance of a sum, working with

$$K_i = \log \hat{S}(t_{(i)}) = \sum_{j=1}^{i} \log \hat{\pi}_j. \tag{3.34}$$

Now we need to find the variance of the log of $\hat{\pi}_i$. The large-sample variance of a function f of a random variable X is

$$\mathrm{Var}(f(X)) \approx f'(X)^2 \mathrm{Var}(X), \tag{3.35}$$

so we just multiply the variance of X by the squared derivative of the transformation. The function is the log and we obtain,

$$\mathrm{Var}(\log \hat{\pi}_i) = (\frac{1}{\hat{\pi}_i})^2 Var(\hat{\pi}_i) = \frac{(1-\pi_i)}{n_i \pi_i}, \tag{3.36}$$

Since K_i is a sum and the covariances of the $\pi_i's$ (and hence of the $\log \pi_i s$) are zero, we find

$$\mathrm{Var}(\log \hat{S}(t_{(i)})) = \sum_{j=1}^{i} \frac{1-\pi_j}{n_i \pi_j}, \tag{3.37}$$

which can be estimated by

$$\sum_{j=1}^{i} \frac{d_j}{n_j(n_j - d_j)}. \tag{3.38}$$

The estimator of the variance of the Kaplan-Meier estimator comes from a second application of the delta method (D. W. Hosmer & Lemeshow, 1999). In this application, the function is

$$f(X) = \exp(X), \tag{3.39}$$

e.g $\hat{S}(t) = \exp\{\log[\hat{S}(t)]\}$. The series expansion is

$$\exp(X) \approx \exp(\mu) + (X - \mu)\exp(\mu) \tag{3.40}$$

and the approximate variance is

$$\hat{\mathrm{Var}}][\exp(X)] \approx \sigma^2[\exp(\mu)]^2. \tag{3.41}$$

Hence, the variance of the survivor function from the variance of its log:

$$\mathrm{Var}(\hat{S}(t_{(i)}) = [\hat{S}(t_{(i)}]^2 \sum_{j=1}^{i} \frac{1-\hat{\pi}_j}{n_i \hat{\pi}_j} \tag{3.42}$$

This is the Greenwood (1926)'s formula. Using the asymptotic normality of its $S(t)$, its confidence interval is

$$\hat{S}(t) \pm z_{\alpha/2}\sqrt{\text{Var}[\hat{S}(t)]}. \tag{3.43}$$

3.8 Nelson-Aalen Estimator

The Nelson-Aalen Estimator is a non-parametric estimator which may be used to estimate the cumulative hazard function (Aalen, 1978).

Consider estimating the cumulative hazard $\hat{H}(t)$. A simple approach is to start from an estimator of $S(t)$ and from Equation 3.10, take $-\log\hat{S}(t)$ as the estimated cumulative hazard. An alternative approach is to estimate the cumulative hazard directly using the Nelson-Aalen estimator:

$$\hat{H}(t) = \sum_{j=1}^{i} \frac{d_j}{n_j}.$$

This expression is estimating the hazard at each distinct time of death $t_{(j)}$ as the ratio of the number of deaths to the number exposed. The cumulative hazard up to time t is simply the sum of the hazards at all death times up to t, and has a nice interpretation as the expected number of deaths in (0,t] per unit at risk. This estimator has a strong justification in terms of the theory of counting processes, (D. W. Hosmer & Lemeshow, 1999).
The variance of $\text{Var}(-\log\hat{S}(t_{(i)}))$ can be approximated by the Greenwood's formula (Rodrıguez, 2005).

3.8.1 Comparison between two survival functions: The Log-rank test and the Wilcoxon test

To compare two survival functions $S_1(t)$ and $S_2(t)$, we need to test the null hypothesis:

$$H_0 : S_1(t) = S_2(t)$$

that the two survival functions are the same (no difference) against the alternative:

$$H_1 : S_1(t) \neq S_2(t).$$

The equation (3.44) is related to the concept of proportional hazards assumption which is explained in later Section (3.9.1). Under H_0, we assume the observations from both distributions come from a single population. Define a dummy variable y taking on the value 0 or 1 according whether an observation comes from first or second distribution.

Let the hazard functions for the two survival functions be $h_1(t) = h_0(t)$ and $h_2(t) = h_0(t)e^{\beta}$, and the two distributions are identical if and only if $\beta = 0$, i.e.,

$$S_2(t) = S_1(t)\exp^{\beta} \tag{3.44}$$

so that by testing $\beta = 0$, we are testing the hypothesis :

$$H_0 : S_2(t) = S_1(t)$$

versus

$$H_1 : S_2(t) = S_1(t)^{\delta}, \delta \neq 1$$

where $\delta = \exp(\beta)$.

Log Rank Test:

Comparison of two survival curves can be perfomed using a statistical hypothesis test called the *log rank* test. It is used to test the null hypothesis that there is no difference between the population survival curves (i.e. the probability of an event occurring at any time point is the same for each population), (Harrington & Fleming, 1982).
Let n_1 and n_2 be the number on individuals in the group 1 and 2, respectively and $n = n_1 + n_2$. Let n_{1i} and n_{2i} be the number of individuals at risk just prior to t_i from the groups 1 and 2 and d_{1i} and d_{2i} be the number of deaths at t_i among the individuals in group 1 and group 2 and

$$d_{1i} + d_{2i} = d_i \; ; \; n_{1i} + n_{2i} = n_i.$$

The number of deaths and survivors are summarised in the contingency table of Table 3.1 for all subjects in the risk set at time t_i :

	Group 1	Group 2	Total
Deaths	d_{1i}	d_{2i}	d_i
Survivors	$n_{1i} - d_{1i}$	$n_i - d_{2i}$	$n_i - d_i$
Total	n_{1i}	n_{2i}	n_i

TABLE 3.1: Number of deaths and survivors in two risk groups at time t_i

The log-rank statistic is given by

$$X_{LR} = \frac{\left[\sum_{i=1}^{r}(d_{1i} - e_{1i})\right]^2}{\sum_{i=1}^{r} V_{1i}}, \tag{3.45}$$

where

$$e_{1i} = n_{1i}\frac{d_i}{n_i}, \tag{3.46}$$

and

$$V_{1i} = \frac{n_{1i}n_{2i}d_i(n_i - d_i)}{n_i{}^2(n_i - 1)}, \qquad i = 1, \ldots, r. \tag{3.47}$$

X_{LR} is the log-rank statistic which approximately has central Chi-Square distribution with one degree of freedom when the null hypothesis is true and the sample size is moderate or large, (Bewick et al., 2004).

Wilcoxon Test:

The *Wilcoxon test*, sometimes also known as the *Breslow test*, is also used to test the null hypothesis that there is no difference in the survival functions for two groups of survival data. The Wilcoxon test statistic is based on the test statistic:

$$X_W = \frac{\left[\sum_{i=1}^{r} n_i(d_{1i} - e_{1i})\right]^2}{\sum_{i=1}^{r} n_i{}^2 V_{1i}}, \tag{3.48}$$

which has central Chi-Square distribution with one degree of freedom when the null hypothesis is true (Custodio Martinez, 2007). The difference between log-rank and Wilcoxon tests is that in the Wilcoxon test, each difference $(d_{1i} - e_{1i})$ is weighted by n_i, the total number of individuals at risk at time $t_{(i)}$. The effect of this is to give less weight to differences between d_{1i} and e_{1i} at those times when the total number of individuals who are still alive is small, that is, at the longest survival times. This statistic is therefore less sensitive than the log-rank statistic to deviations of d_{1i} from e_{1i} in the right tail of the distribution of the survival times and more sensitive at the start.
When the alternative to the null hypothesis of no difference in survival times between two groups is that the risk of death for an individual in one group at any given time is proportional to the risk for a similar individual in the other group at that time, the log-rank test is applicable. This is the assumption of the proportional risk, which is at the heart of a lot of survival data analysis approaches. The survival functions for the two groups of survival data do not cross each other when the hazard functions are proportional. The Wilcoxon test is more appropriate for comparing the two survival functions for other forms of departures from the null hypothesis than the log-rank test.

3.9 The Cox's model

Kaplan-Meier methods and the log-rank test can only study an effect of one factor at a time and cannot be used for multivariate analysis. For this purpose, we use the regression technique developed by Cox (1972). The use of the Cox model has rapidly grown and it became one of the most popular methods of survival analysis used in medical research.

If we have a vector of covariates, x , the Cox's regression models the hazard as :

$$h(t,x) = h_0(t)e^{x'\beta}, \tag{3.49}$$

and the estimation of β can be observed without making any assumptions about the baseline hazard $h_0(t)$.

3.9.1 Cox's Proportional Hazards (CPH) models

Cox (1972) introduced two propositions; his first proposition was a model known today as the *Proportional Hazards Model*. Secondly, he proposed a new estimation method called *partial likelihood*.

Cox's proportional hazards model is the most widely used method in survival analysis and we will use this method to analyse the relationship of survival time distribution to the different potential risk factors. The analysis consists of specifying a linear model for the log-hazard (Fox & Weinsberg, 2011). The Cox's proportional hazards model can be written as

$$h_i(t) = h_0(t)\exp(\beta_1 x_{i1} + \beta_2 x_{i2} + \cdots).$$

The CPH model a multiplicative hazard. No shape is assumed for baseline hazard $\log h_0(t) = \alpha(t)$. That is

$$\log h_i(t) = \alpha(t) + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots$$

The same model can be represented in matrix notation as :

$$h_i(t|x_i) = h_0(t)\exp(X_i^T\beta), \tag{3.50}$$

where $h_0(t)$ is the baseline hazard function and $\exp(X_i^T\beta)$ is the hazard risk (a proportionate increase or reduction) in risk.

The assumption of the Cox model is that the hazard of two populations, the exposed (cases) and the unexposed (controls) are proportional over time (D. W. Hosmer & Lemeshow, 1999); (Kalbfleisch

& Prentice, 2011). This entails that the ratios of their estimated parameters are constant with respect to follow-up time. This equation states that the hazard for individual i at time t is the product of two factors:

1. a baseline hazard function $h_0(t)$ that is unspecified,
2. a linear function of a set of fixed p covariates $(x_1,\ldots,x_p)$, which is exponentiated.

If we further specify $\alpha(t) = \alpha$, which is the simplest function that says the hazard is constant over time, we will get the exponential model. If we specify $\alpha(t) = \alpha t$, we will get the Gompertz model. And if we specify $\alpha(t) = \alpha \log(t)$, we will have a Weibull model. Thus, the proportional hazards model is generalization of the Exponential, Weibull and Gompertz model (Allison, 2010).

The hazard for any individual is a fixed ratio of the hazard for any other individual, hence the name proportional hazards model. For illustration, we can take the ratio of the hazards for two individuals i and j and use Equation (3.50) :

$$\frac{h_i(t)}{h_j(t)} = \exp \beta_1(x_{i1} - x_{j1}) + \ldots. + \beta_p(x_{ip} - x_{jp}) \tag{3.51}$$

The $h_0(t)$ cancels out of the numerator and denominator. As a result, the ratio of the hazards is constant over time.

3.10 Fitting the Proportional Hazards Model

To fit the proportional hazards model, the unknown coefficients $\beta_1, \beta_2, \beta_3, \ldots, \beta_p$ have to be estimates. Let $\beta^T = (\beta_1, \beta_2, \beta_3, \ldots, \beta_p)$ denote the vector of coefficients corresponding to the p covariates. Sometimes, the baseline hazard $h_0(t)$ needs to be estimates. the most common method to estimate β^T is the method of maximum likelihood.

Let t_i is the observed survival time, $\mathbf{x}_i$ be the covariate and c_i is the censoring indicator variable for each individual i out of a group of n individuals.

Suppose that there are r ordered distinct death times such that $t_{(1)} < t_{(2)} < \ldots < t_{(r)}$. This means that there are $n - r$ right-censored survival times. The ties are not considered here but will be explained later. Suppose that the set of individuals that are at risk at time $t_{(j)}$ are denoted by $R\left(t_{(j)}\right)$, the risk set.

If we consider the result from Equation (3.8), we can formulate :

$$f(t, \mathbf{x}, \beta) = h(t, \mathbf{x}, \beta) \times S(t, \mathbf{x}, \beta) \tag{3.52}$$

Considering next the likelihood function from regression models,

$$L(\beta) = \prod_{i=1}^{n} \left\{ [f(t_i, \mathbf{x}_i, \beta)]^{c_i} \times [S(t_i, \mathbf{x}_i, \beta)]^{1-c_i} \right\} \tag{3.53}$$

Substituting Equation 3.52 into Equation 3.53 gives :

$$L(\beta) = \prod_{i=1}^{n} \left\{ [h(t_i, \mathbf{x}_i, \beta) \times S(t_i, \mathbf{x}_i, \beta)]^{c_i} \times [S(t_i, \mathbf{x}_i, \beta)]^{1-c_i} \right\} \tag{3.54}$$

$$= \prod_{i=1}^{n} \left\{ [h(t_i, \mathbf{x}_i, \beta)]^{c_i} \times [S(t_i, \mathbf{x}_i, \beta)] \right\} \tag{3.55}$$

Substituting $h(t_i, \mathbf{x}_i, \beta) = h_0(t_i) e^{\mathbf{x}_i^T \beta}$ and $S(t_i, \mathbf{x}_i, \beta) = [S_0(t)]^{\exp\left(\mathbf{x}_i^T \beta\right)}$ into the above equation results in

$$L(\beta) = \prod_{i=1}^{n} \left\{ \left[h_0(t_i) e^{\mathbf{x}_i^T \beta} \right]^{c_i} \times [S_0(t_i)]^{\exp\left(\mathbf{x}_i^T \beta\right)} \right\} \tag{3.56}$$

$$\ln[L(\beta)] = \ell(\beta) = \sum_{i=1}^{n} \left\{ c_i \ln[h_0(t_i)] + c_i \mathbf{x}_i^T \beta + e^{\mathbf{x}_i^T \beta} \ln[S_0(t_i)] \right\} \tag{3.57}$$

The maximum likelihood estimator method in Equation needs to be maximised with respect to β and a parametric model for the baseline hazard. To avoid defining the baseline hazard, the proportional hazard is considered.

Cox (1972) proposed the 'partial likelihood' function implying that this likelihood depends only on the parameters of interest in order to find the estimators. In his paper, he showed that the distributional properties of the partial likelihood were the same as the full maximum likelihood estimators. According to Fleming and Harrington (2011), the partial likelihood can be estimated in the counting process approach by :

$$L(\beta) = \prod_{j=1}^{r} \frac{\exp\left(\mathbf{x}_{(j)}^T \beta\right)}{\sum_{l \in R\left(t_{(j)}\right)} \exp\left(\mathbf{x}_l^T \beta\right)} \tag{3.58}$$

The conditional probability in Equation (3.58) relates to individuals who experience the event. Let c_i be the censoring which is 0 when the individual is censored and 1 otherwise, the likelihood can hence be modified to

$$L(\beta) = \prod_{i=1}^{n} \left[\frac{\exp\left(\mathbf{x}_i^T \beta\right)}{\sum_{l \in R\left(t_{(j)}\right)} \exp\left(\mathbf{x}_l^T \beta\right)} \right]^{c_i}. \tag{3.59}$$

Taking the natural log of the likelihood yields

$$\ln L(\beta) = \sum_{i=1}^{n} c_i \left[\mathbf{x}_i^T \beta - \ln \sum_{l \in R\left(t_{(j)}\right)} \exp\left(\mathbf{x}_l^T \beta\right) \right] \tag{3.60}$$

If the log likelihood is then maximised with respect to β, the following is obtained,

$$\begin{aligned} \frac{\partial \ell(\beta)}{\partial \beta_k} &= \sum_{j=1}^{r} \left[x_{(jk)} - \frac{\sum_{l \in R\left(t_{(j)}\right)} x_{lk} \exp\left(\mathbf{x}_l^T \beta\right)}{\sum_{l \in R\left(t_{(j)}\right)} \exp\left(\mathbf{x}_l^T \beta\right)} \right] \\ &= \sum_{j=1}^{r} \left\{ x_{(jk)} - \bar{x}_{wjk} \right\} \end{aligned}$$

where

$$\bar{x}_{wjk} = \sum_{l \in R\left(t_{(j)}\right)} w_{jl} x_{lk}$$

and

$$w_{jl} = \frac{\exp\left(\mathbf{x}_l^T \beta\right)}{\sum_{l \in R\left(t_{(j)}\right)} \exp\left(\mathbf{x}_l^T \beta\right)}$$

$x_{(jk)}$ is the value of the covariate x_k for a subject with observed ordered survival time $t_{(j)}$. The derivative is then set to zero and then solved to obtain the estimator β. Generally, iterative method is used to solve for the unknown parameters.
The Information matrix is the negative value of the second derivative of $\ell(\beta)$ as :

$$I(\beta) = -\frac{\partial^2 \ell(\beta)}{\partial \beta \partial \beta^T}$$

Hence, the variance for the estimator of β is

$$\text{var}(\hat{\beta}) = I(\hat{\beta})^{-1}$$

3.10.1 Confidence Interval for estimated coefficients

The likelihood ratio test is used to test the statistical significance of covariates in the Cox's regression model (Grambsch & Therneau, 1994).

If $\beta^{(0)}$ is the true theoretical value of the coefficients and $\hat{\beta}$ be the estimated coefficients. To test the global null hypothesis $H_0 : \hat{\beta} = \beta^{(0)}$, the likelihood ratio test is given by :

$$\text{LRT} = 2\left(\ell(\hat{\beta}) - \ell\left(\beta^{(0)}\right)\right) \tag{3.61}$$

It is equivalent to twice the difference between the likelihood of a null model(model without covariates) and the CPH model with p parameters. LRT follows a χ^2 distribution with p degrees of freedom, where p is the difference of the number of estimated coefficients.

To check the statistical significance of a specific covariate x, i.e., to test whether x is significantly associated with survival time, the Wald test is used (Grambsch & Therneau, 1994). The Wald test assesses the hypothesis that the estimated coefficient for covariate x, denoted by $\hat{\beta}_x$, from the Cox's model, is significantly different from zero. The test statistic, Z, is given by

$$Z = \frac{\hat{\beta}_x}{\widehat{\text{se}}\left(\hat{\beta}_x\right)}$$

where $\widehat{\text{se}}\left(\hat{\beta}_x\right)$ is the standard error associated with $\hat{\beta}_x$. The test statistic Z follows a standard normal distribution and can also be used to formulate confidence interval for $\hat{\beta}_x$. The $(1-\alpha)\%$ confidence interval (CI) for $\hat{\beta}_x$ is given by

$$\text{CI}\left(\hat{\beta}_x\right) = \hat{\beta}_x \pm Z_{1-\alpha/2}\widehat{\text{se}}\left(\hat{\beta}_x\right)$$

where Z_α represents the critical value at $(1-\alpha)\%$ level from the standard normal distribution. When the $(1-\alpha)\%$ CI of an estimated coefficient contains zero, it means that the covariate associated with that estimated coefficient does not significantly affect the hazard of failure.

3.10.2 Frailty in Survival model

Univariate survival models assume homogeneity i.e., all individuals are subject to the same risks $h(t)$ or the survivor function $S(t)$. Unobserved sources of heterogeneity can be modelled in the study population by the method of univariate frailty models (Vaupel et al., 1979). The univariate survival model is an extension of the Cox model Equation (3.50) where the hazard rate is directly proportional to a parameter Z.

$$h(t|X,Z) = Zh_0(t)\exp X^T\beta \tag{3.62}$$

The frailty component Z is assumed to have an expected value of 1 and to be multiplicative in the hazard rate. The Survival function of an individual conditional on the frailty Z, is defined as :

$$\begin{aligned} S(t|X,Z) &= \exp[-H(t|Z,X)] && (3.63)\\ &= \exp[-Z\Lambda_0(t)\exp(X^T\beta)] && (3.64) \end{aligned}$$

where $\Lambda_0(t)$ is the integrated or cummulative baseline hazard function at time t.

Suppose there is I groups in which n individuals are assigned to, such that $\sum_{i=1}^{I} n_i = n$.

Let $D_i = \sum_{j=1}^{n_i} \delta_{ij}$ denote the number of events experienced by the i^{th} group, where δ_{ij} is the censoring indicator which takes on 1 on event occurrence or 0 if not.

The hazard for the j^{th} individual from the i^{th} group is given by

$$h_{ij}(t) = h_0(t)\exp\left(\mathbf{x}_{ij}^T\beta + w_i\right)$$

where $\mathbf{x}_{ij}$ is a vector of p covariates for individual j in group $i, h_0(t)$ is the baseline hazard, and w_i is the random effect for the i^{th} group. The w_i 's are an independently and identically distributed random sample from a density $f_W(\cdot)$.

The function can also be expressed as :

$$\begin{aligned} h_{ij}(t) &= h_0(t)\exp(w_i)\exp\left(\mathbf{x}_{ij}^T\beta\right) && (3.65)\\ &= z_i h_0(t)\exp\left(\mathbf{x}_{ij}^T\beta\right) && (3.66) \end{aligned}$$

where $z_i = \exp(w_i)$ is the frailty term. The z_i 's are independent, and are assumed to have common density $f_Z(\cdot)$.

The z_is are usually assumed to follow gamma density as :

$$f_Z(z) = \frac{\alpha^\alpha z^{\alpha-1}e^{-\alpha z}}{\Gamma(\alpha)}$$

which corresponds to a log-gamma density for W. The mean and variance for Z are given by

$$\begin{aligned} E[Z] &= 1\\ \mathrm{Var}[Z] &= \frac{1}{\alpha} \end{aligned}$$

The baseline hazard in Equation (3.66) can be left unspecified or can be explicitly specified. Under parametric estimation for the baseline hazard, MLE procedures can be used but in the case when the baseline hazard is unspecified, the unknown parameters in the shared frailty model can be estimated by methods namely (Phipson, 2006):

1. Expectation-Maximization (EM) algorithm
2. Penalized Partial Likelihood (PPL) approach
3. Markov Chain Monte Carlo (MCMC) methods
4. Monte Carlo EM (MCEM) approach (Ripatti et al., 2002)
5. Different methods using Laplace approximation.

3.10.3 Ties or grouped observations

Survival times may have ties in survival models, arising either due to imprecision of incomplete data or the rounding of continuous times. It also may arise in the case of discrete. Ties may affect the partial likelihood of the Cox model. The presence of ties in the data causes imprecision in the estimation of the partial log-likelihood function. The next set of equations illustrate the problem of tied data. Consider k individuals in time order and two individuals experiencing failures at the same time. If ties are ignored, then the log-likelihood function can either be of the two next forms (considering only the first two forms):

$$\left(\frac{r_1}{r_1+r_2+\ldots+r_k}\right)\left(\frac{r_2}{r_2+r_3+\ldots+r_k}\right) \tag{3.67}$$

or

$$\left(\frac{r_2}{r_1+r_2+\ldots+r_k}\right)\left(\frac{r_1}{r_1+r_3+\ldots+r_k}\right) \tag{3.68}$$

but it is difficult to predict which of Equation (3.67) and (3.68) reflects the true likelihood function. Grambsch and Therneau (1994) proposed three methods to tackle the problem of tied observations, namely :

1. Breslow approximation,
2. Efron approximation and
3. Discrete method.

Breslow approximation

The Breslow approximation is one of the simplest methods to estimate the likelihood function for tied survival data. It has been proposed independently by Breslow and Peto (1972). It uses the

complete sum of $r_1 + r_2 + \ldots + r_k$ as the denominator of the partial log-likelihood function :

$$\left(\frac{r_1}{r_1 + r_2 + \ldots + r_k}\right)\left(\frac{r_2}{r_1 + r_2 + \ldots + r_k}\right) \tag{3.69}$$

Efron approximation

The Efron approximation, developed by Efron (1977) is considered as a more accurate partial likelihood using an average denominator.

$$\left(\frac{r_1}{r_1 + r_2 + r_3}\right)\left(\frac{r_2}{0.5r_1 + 0.5r_2 + r_3}\right) \tag{3.70}$$

Discrete approximation

The discrete method does not assume any ordering of the tied events and treats time as a discrete variable. Under this scenario, the log-likelihood function yields to

$$\frac{r_1 r_2}{r_1 r_2 + r_1 r_3 + r_2 r_3} \tag{3.71}$$

In conclusion, the Efron approximation method is the most reliable approximation among the three methods, does not produce bias, and is computationally more efficient.

3.11 Proportional Hazard Regression Diagnostics

Hypothesis tests can be employed to check model assumptions (proportional hazards assumption), verify whether covariates need to be transformed and to detect the presence of outliers and unduly strong influential observations. In addition, we can further check the assumptions graphically with residual plots,using:

- Martingale residuals
- Deviance residuals
- Delta-beta residuals
- Schoenfeld residuals

3.11.1 Martingale Residuals

If we consider data for each subject to be (y_i, δ_i, x_i) where δ_i is the count of the number of events for the i^{th} subject (0 or 1) and $H_0(t)$ is an estimate of the baseline cumulative hazard function,then the cumulative hazard for subject i is therefore,

$$H_i(t_i|X) = H_0(t)exp(\beta^T X)$$

Martingale residuals compare the observed values of δ_i to expected:

$$r_{Mi} = \delta_i - \hat{H}_i(t_i|X) \tag{3.72}$$

They are motivated by the fact that, for large samples, the quantity r_{Mi} would be a martingale evaluated at the time t_i. Under correct model specification, martingale residuals have a mean zero and are uncorrelated with one another across subjects. The martingale residual r_{Mi} is a measure of the degree to which the i^{th} subject is an outlier, after adjusting for the effect of X.

This represents the discrepancy between the observed value of a subject's failure indicator and its expected value, integrated over the time for which the patient was at risk. Patients who had positive values died sooner than expected, whereas those who had negative values lived longer than expected (or censored).

While martingale residuals are uncorrelated and have a mean of zero, their disadvantage is that:

1. Their maximum is +1, but their minimum is $-\infty$.
2. Their distribution is quite skewed to the left.

The heavily skewed distribution of martingale residuals makes them hard to use to identify outliers. For this, deviance residuals are a better option.

3.11.2 Deviance residuals

Constructing a "deviance" residual is a technique for obtaining symmetric, normalised residuals that is commonly used in generalised linear modelling. The idea behind the deviance residuals is to examine the difference between the log-likelihood for subject i under a given model and the maximum possible log-likelihood for that subject.

Deviance residuals are defined as (Gillen, 2016) :

$$r_{Di} = \text{sign}(r_{Mi})\big[-2r_{Mi} + \delta_i \log(\delta_i - r_{Mi})\big]^{\frac{1}{2}}. \tag{3.73}$$

By definition, r_{Di} has the same sign as r_{Mi}. The quantity inside the square bracket is positive.

Compared to r_{Mi} , r_{Di} has a shorter left and a longer right tail. r_{Di} is more symmetrical around zero. The distribution of the deviance residual is better approximated by a Gaussian distribution than is the distribution of the martingale residuals.

3.11.3 Delta-beta residuals

The idea behind delta-beta residuals is very simple. Let $\hat{\beta}_j^{(i)}$ denote the estimate of the coefficients $\hat{\beta}_j$ obtained if we leave subject i out of the model.

The delta-beta residual for coefficient j, subject i is therefore defined as

$$\hat{\Delta}_{ij} = \hat{\beta}_j - \hat{\beta}_j^{(i)} \tag{3.74}$$

Delta-beta plots are always fascinating to look at and provide a wealth of information about the inner workings of complex models. Furthermore, they can show whether a few individuals dominate the estimation of a coefficient, which would be cause for concern.

3.11.4 Schoenfeld residuals

Schoenfeld (1982) proposed that instead of a single residual for each individual, we need to assign a separate residual for each individual and each covariate. Schoenfeld residuals are not defined for censored individuals.

For the i^{th} subject and k^{th} covariate, the estimated Schoenfeld residual, r_{ik}, is given by (D. W. Hosmer & Lemeshow, 1999) :

$$\hat{r}_{ik} = x_{ik} - \hat{\bar{x}}_{w_i k}$$

Where x_{ik} is the value of the k^{th} covariate for individual i, and $\hat{\bar{x}}_{w_i k}$ is a weighted mean of covariate values for those in the risk set at the given event time. A positive value of r_{ik} shows an X value that is higher than expected at that death time.

We normally plot the Schoenfeld residuals against time to evaluate the proportional hazard assumption. The residuals represent the difference between the observed covariate values and the expected given the risk set at that time and are calculated for each covariate. Schoenfeld (1982) showed that the r_i's are asymptotically uncorrelated and have expectation zero under the Cox model. Thus a plot of r_{ik} versus X_i should be centered about zero if the PH is satisfied. On the other hand, non-PH in the effect of X_i could be revealed in such a plot.

3.12 Assessing overall goodness of fit

In comparison to traditional regression models, assessing the goodness-of-fit of survival models involves unique issues. The dependent variable or variables being modelled are often observed directly in a traditional statistical study. For example, the heights of a group of people are measured directly and forecasted using a model that takes into consideration confounders like age and gender.

In survival analysis, on the other hand, the timing of the event is observed (e.g., age at death), but the hazard is predicted. (Price & Jones, 2017).

There are different tests available for assessing and comparing different models. When the number of occurrences increases, Machin et al. (2006) recommends using the likelihood ratio test for stability and consistency.

3.12.1 Wald Test

When fitting a Cox's model of a single variable x, which may be a binary variable or a continuous variable, we obtain an estimate, b, of the associated regression coefficient, β, together with the standard error $SE(b)$. The statistical significance of this variable, which is equivalent to a test of the null hypothesis, $\beta = 0$, is established by use of the $z = \dfrac{b}{SE(b)}$ test. Equivalently, the Wald test can be used, where $W = z^2$.

This test can also be used whether the addition of another variable, to a Cox model already containing v variables, improves the model. However, the test is only strictly valid if the estimated regression coefficients for the variables are not unduly influenced by the presence of the additional variable (Machin et al., 2006).

3.12.2 Likelihood Ratio Test

The likelihood ratio (LR) test is a more general test than the Wald test which can cope with categorical variables of more than two levels and with adding several variables simultaneously to a Cox's model.
To illustrate this test, we first define the null model which, in the terminology of the Cox's model, specifies that no variable influences survival. When the variable is treatment, this is equivalent to setting $\beta = 0$ in Equation (3.49) to imply $\lambda_N(t) = \lambda_0(t)$, or that, irrespective of treatment, all patients have the same hazard.
The likelihood for this model is denoted as ℓ_0 where ℓ denotes likelihood and the zero, 0, represents the fact that all the regression coefficients are set to zero. In contrast, the likelihood of the model which contains v different regression coefficients is written as ℓ_v and the regression coefficients are estimated by the method of maximum likelihood. The larger ℓ_v is relative to ℓ_0, the better the model explains, or fits, the observed data. The fit of each model can be tested using the Likelihood Ratio (LR) as defined by

$$LR = -2\log(\frac{\ell_0}{\ell_v}) = -2(L_0 - L_v) \tag{3.75}$$

where $L_v = \log \ell_v$ and $L_0 = \log \ell_0$. Under the hypothesis of no difference in the two models, that is, where including the variables in the Cox model does not help to explain the survival data any more

satisfactorily than the null model with no variables, the LR of Equation (3.75) has a χ^2 distribution with df = v, (Machin et al., 2006).

To assess the relative fit of two models, one with $(v+k)$ regression coefficients and the other with the fewer v regression coefficients, we can use the statistic

$$LR_k = -2(L_{v+k} - L_v) \tag{3.76}$$

This statistic also has a χ^2 distribution, but with df = $(v+k) - v = k$. The null hypothesis here is that there is no improvement in the fit of the model, by including the extra k coefficients.

3.12.3 Harell concordance

Concordance, also known as the C- statistic, is a frequent and useful measure of model discrimination that may be applied to any ordered result. In general, choose two patients at random and note if the model accurately forecasts order for each pair, e.g., a higher model score for a better result (Kremers, 2007). The fraction of pairs for which the model is true is called concordance. The concordance of a perfectly random forecast is 0.5, whereas the concordance of a perfect rule is 1.

According to Kremers (2007), concordance is most familiar from logistic regression, where it is also known as the area under the receiver operating curve. It can also be described for survival data while allowing for censoring and no distributional assumption need be made for motivation or calculation.

The C-statistic is used in survival modelling to determine the likelihood that a randomly selected patient who had an event (such as a disease or condition) had a higher risk score than a patient who did not. It runs from 0.5 to 1 and is equivalent to the area under the Receiver Operating Characteristic (ROC) curve (D. Hosmer, 2000).

- A value below 0.5 indicates a very poor model.
- A value of 0.5 means that the model is no better than predicting an outcome than random chance.
- Values over 0.7 indicate a good model.
- Values over 0.8 indicate a strong model.
- A value of 1 means that the model perfectly predicts those group members who will experience a certain outcome and those who will not.

3.12.4 AIC and BIC

The two most common information criteria used for model selection are the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). AIC and BIC are respectively given

by

$$AIC = -2\ell - 2k \tag{3.77}$$

$$BIC = -2\ell + k\ln(n) \tag{3.78}$$

where ℓ is the partial log-likelihood function of the regression model, k is the number of estimated parameters in the regression model and n is the number of events in the survival analysis. BIC penalizes larger models more heavily as it depends on the number of events, n and tends to perform better for smaller sample size model fitting in comparison to AIC.

3.12.5 Model Selection

Selecting a model involves deciding which terms should be included in a response variable model. The model can include every tested factor as well as interactions and covariates as terms (e.g., potential nuisance variables that were recorded for statistical control). The goal of model selection is to choose a sparse statistical model that adequately explains the data. Parsimony (model simplicity), Goodness-of-Fit Test (model well fits the data), and generalisability (model fits the data well) are the three main qualities of a good model (model can be used to describe or predict new data).

A decent model also only includes the elements and covariates that are absolutely necessary in order to avoid being "underfit" (too simple), "overfit" (unnecessarily complex), and thirdly, to take into consideration potential confounding. Regression modelling offers a variety of model construction techniques. Exhaustive search, forward selection, backward selection, and step-wise selection are examples of traditional methods; regularisation methods, however, are receiving more and more attention (ridge regression, LASSO, elastic net).

In the forward selection approach, the model starts with just an intercept term in the initial model. The addition of each variable is then evaluated using a predetermined criterion, and the variable (if any) that enhances the model the most is included. The process stops when no more variable considerably enhances the model. Similar reasoning underlies backward selection, however in the opposite direction. Backward selection involves starting with a model that contains all relevant variables. The variable with the lowest criterion is then eliminated from the model. Up until every last predictor reaches a predetermined degree of significance, this procedure iteratively continues.

Last but not least, step-wise selection might be seen as a fusion of forward and backward selection. This means that step-wise selection could start as a backward selection process, eliminating variables that don't enhance the model. If certain variables have changed in significance as measured by the chosen criterion, they may be reintroduced into the model. Some statistical software automates the forward, backward, and step-wise techniques.

Any technique has the potential to either include or exclude covariates that are crucial (Harrell et al., 2001; Hosmer Jr et al., 2013). In comparison to the other stepwise procedures, backward elimination is favoured because it is least likely to leave out significant factors that are only significant in respect to other covariates ((Harrell et al., 2001; Hosmer Jr et al., 2013).

In order to choose a final mode, some standard criterion is used to judge whether or not the model "improved" during the model selection process. There are many selection criteria available to try to describe how well the model fits the data. One model selection criteria is the significance of the factors and covariates based on the p-value. Using forward, backward, or step-wise selection techniques, the p-value of an individual variable (a factor or a covariate) may be used as an evaluation criterion.

The likelihood ratio test (Section 3.12.2), another similar criterion, makes use of the idea of likelihood. Given a chosen model, likelihood assesses the "probability" of the observed data; the higher the likelihood, the better the model's goodness of fit to the data. The likelihood ratio test statistic compares the fit of two nested models and has a chi-squared distribution so that a test of significance can be performed. Depending on the selection strategy chosen, terms are continued to be added or eliminated until there is no discernible difference between the new model and the old one.

The Akaike's Information Criterion (AIC) and the Bayesian Information Criterion (BIC) (Section 3.12.4) compare various possible subsets of factors based on a trade-off between complexity and lack of fit (measured by model likelihood) (measured by the number of parameters included in the model). Model selection based on AIC has a non-zero likelihood of overfitting the data, that is, choosing insignificant covariates that make the model overly complex, while model selection based on BIC has a zero probability of overfitting the data as the sample size grows to infinity (Hastie et al., 2009).

The model selection started with a model with main effects and second-order interactions. Second-order interaction effects were backward eliminated for this study. Since main effects' importance in influencing survival prospects had already been demonstrated in earlier studies, it was believed that only interaction effects would be dropped from the model and not main effects. A manual elimination approach was used. The decomposition of Cox's model was performed using backward elimination by the LRT approach (ANOVA command) till all interaction effects were significant at the 1% level. Interactions between factors with small frequencies in their categories were not considered. A survival model with three to 4 interaction effects for TIA and IS cohorts was produced by this manual backward elimination procedure.

3.12.6 Verifying the proportionality assumption

Graphical representation or numerical methods can be used to verify the proportionality assumption. In the previous section, we introduced the methods of the Schoenfeld residual scaled plot (Section

3.11.4).

cox.zph test

The cox.zph() function from the survival package allows to test this assumption for each variable by creating interactions with time. A small p-value would indicate a violation of the proportionality assumption. cox.zph function can be used to generate a plot for each of the individual predictors in the model. The plot of scaled Schoenfeld residuals should be a horizontal line. The includes a hypothesis test identifying whether the gradient differs from zero for each variable. The function outputs a table in which each row represents one variable, with the last row serving as Schoenfeld's global test for the proportional assumption's violation. The column provides the correlation coefficient between transformed survival time and the scaled Schoenfeld residuals (rho), a chi-square statistic (chisq), and the two-sided p-value (p).

Kolmogorov-Smirnov and Cramer von Mises tests

In a situation of time-varying effects, Scheike and Martinussen (2004) proposed that it is preferable to test each component separately, starting with the model where all effects are permitted to change over time and gradually simplifying the model as necessary. The test statistics proposed by them are based on the asymptotic analysis of the cumulative regression functions of the proposed model.

For example, to test whether the effect of x_j(covariate) is constant, where $a_j(t)$ is the effect of a regression coefficient, the null hypothesis is considered as $H_0 : a_j(t) = \beta_j$, for all $t \in [0, \tau]$, where τ is study endpoint.

To test H_0 there are many possibilities, a simple test that relies only on $\hat{a}_j(t)$ is to look at

$$T_j(t, \widehat{\alpha}_j) = \widehat{\alpha}_j(t) - \frac{1}{\tau} \int_0^\tau \widehat{\alpha}_j(s) ds,$$

for $j = 1, \ldots, p$. Scheike and Zhang (2008) derived the asymptotic distribution of this test process and proposed to compute the p- value of the test based on a Kolmogorov-Smirnov type test-statistic $\sup_{t \in [0,\tau]} \left| T_j(t, \hat{a}_j) \right|$ or by a Cramer von Mises type test-statistic $\int_0^\tau \left\{ T_j(s, \widehat{\alpha}_j) \right\}^2 ds$. The Cramer von Mises test is an alternative to the Kolmogorov-Smirnov test. Both tests can be performed and plots of the simulated test processes under the null hypothesis can help to visually examine whether a specific covariate has a time-varying effect. All large sample properties and resampling techniques used for the test statistics are explained in Scheike and Martinussen (2004).

Estimation routines for the model can be found in the *timereg* R package (Scheike & Zhang, 2011). The resampling method is used to test the time-varying effect. The Cox regression model is fitted similarly to how it is in the survival package, with the exception that the statistical inference is made using resampling techniques, necessitating the specification of the number of simulations (n.sim=500). The end of the observation period when estimates are computed is specified by the max.time argument (Scheike & Zhang, 2008).

The results typically produce two sets of summaries, one for the test for non-significant effects and the next set for the tests of time-invariant effects with outputs for the Kolmogorov-Smirnov test and the Cramer von Mises test. The significance of the first table essentially demonstrates whether given variables have significant effects on the survival outcome relative to the null hypothesis that the coefficients under test are not significantly different from 0. The test also produced plots for the estimated cumulative coefficients with 95% pointwise confidence intervals where the effects of the variables over time can be observed. From the output of the time-invariant effects and plots, we can proceed by finalizing which variables will ultimately have time-fixed effects by using the const() function. Both tests are flexible to deal with time-varying dynamics of covariate effects in the sense that they investigate how the effect of the covariates acts on an absolute scale rather than on a relative scale. While cox.zph test provides the first overview to identify potential time-varying variables, the Kolmogorov-Smirnov and Cramer von Mises tests can be used as a next step to fully confirm the variables having a time-varying effect as they show exactly where non-proportionality is present in terms of the covariate effects. One advantage of this procedure is that the model reduction in each step is kept to a relatively low degrees of freedom (Scheike & Zhang, 2008).

3.12.7 Tackling time-dependent effect

The proportional hazards assumption may not hold in most of the survival studies. In the previous section, it was demonstrated how to verify the proportional hazards assumption following the fitting of a Cox regression model, and in this section, it is demonstrated how the model should be changed to best reflect the data in the event that the assumption is violated. Time-varying covariates, a frequent occurrence in clinical research, happens when a particular covariate changes over time during the follow-up period, for example, age being a time-varying factor. The Cox model has the advantage of being able to account for factors that change over time. A time-dependent effect could be included such as for the continuous variable as a function of time or the splitting follow-up time methods.

- **Using step function**
  A step function, such as

$$g(t) = I(t \geq t_o)$$

  , where t_o is a predetermined value, can be used to describe time-varying coefficients. This technique stratifies the Cox proportional model for the various time periods after dividing the analysis time into discrete chunks. By stratifying data by time, we can investigate whether the effect of constant baseline factors grows or weakens over time. After observing the plot produced by the the cox.zph() function, we can decide on the time intervals. The survSplit() function enables to divide the data set into cut time points. It is then important to check if the proportional hazards assumption holds for the stratified Cox regression model used in the different time windows. This method may prove to be problematic if different

covariates have different time cut-off periods, resulting in lots of time windows and when the proportional hazards assumption violation persists with the stratified analysis (Therneau & Grambsch, 2000).

- **Stratification**
  Stratification, a practical and straightforward alternative for non-proportional risk analysis, enables the consideration of various baseline hazard functions in various strata. A separate partial likelihood is created for each stratum. If a factor fails to produce proportional hazards between categories, we may stratify on the category. However, we would not be able to observe the overall effect of the variable of interest.

- **Parametric approach**
  Parametric modelling offers the benefit of specificity where the underlying hazard function is known or specified and the accommodation to a variety of different shapes for the hazard function ranging from the simple constant hazard to complex "bathtub" which can be more relevant in modelling human life (D. Hosmer, 2000). To circumvent the problem of non-proportionality, the current study used the double-Cox parametric model developed by Begun et al. (2019) which was used in time-to-failure of hip replacement, and simulation studies by Begun and Kulinskaya (2022). The model basically employs an additional, distinct Cox-regression term that is dependent on the vector of covariates to parameterize the baseline hazard function's shape parameter. The R programs for fitting the double-Cox model were made available on Github by Dr A. Begun on https://rdrr.io/github/AB5103/doubleCoxr/f. The next section presents the Double-Cox model in more detail.

3.13 Shape and scale: Weibull model

The Weibull model (developed by Waloddi Weibull in 1939) is a significant extension of the exponential model with two positive parameters. The model's second parameter gives it a lot of versatility and lets the hazard function take multiple shapes. The versatility of the Weibull model, on the one hand, and the simplicity of the hazard and survival functions, on the other, make it useful for empirical research.

Survival time T has a Weibull distribution of $W(t|a,b)$ with pdf :

$$f(t) = \frac{b}{a}\left(\frac{t}{a}\right)^{b-1} e^{-(t/a)^b}$$

and hazard function defined as:

$$h(t) = \frac{b}{a}\left(\frac{t}{a}\right)^{b-1}$$

where a is defined as the scale parameter and b is defined as the shape parameter. The hazard function, which can take many different forms, is influenced by these parameters. The Weibull,

Gompertz, and gamma distributions all support decreasing and increasing hazards in a monotonic manner. A non-monotonic hazard rate shape does not exist in the Weibull distribution.

The hazard is decreasing for shape parameter $b < 1$ and increasing for $b > 1$. For $b = 1$, the Weibull distribution is equivalent to an exponential distribution with rate parameter 1/a.

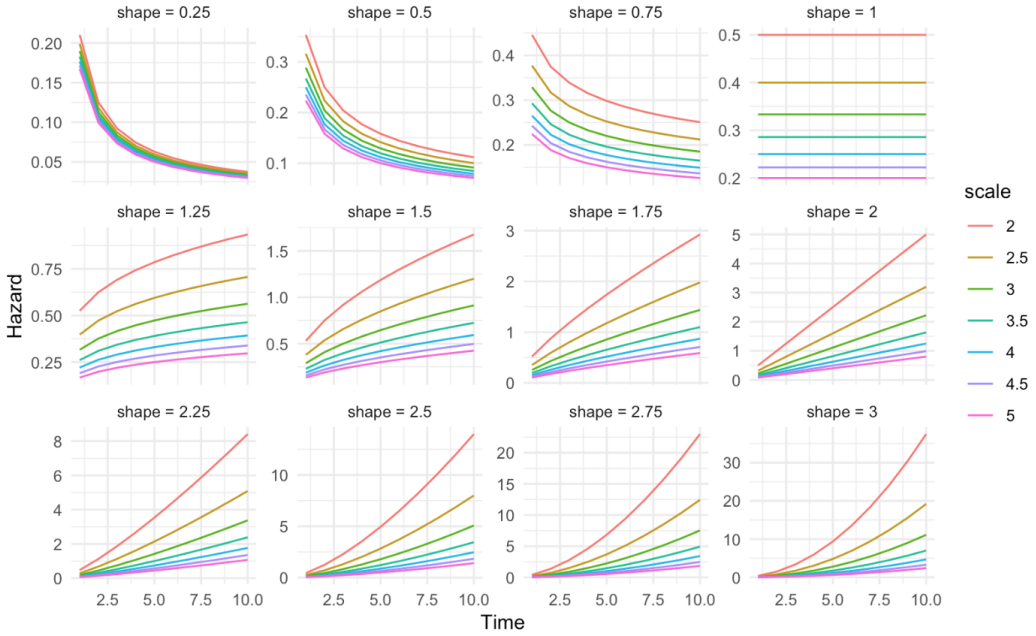


FIGURE 3.3: Illustration of the effects of the different shapes and scales on the Weibull hazards

Source : Devin (2019)

3.14 The Weibull Double-Cox model with shared frailty

We explained the different ways to circumvent the problem of non-proportional hazards in the earlier section. Another efficient method is to use the double-Cox type model. This method not only entails a parametric approach with regards to the modelling of the baseline hazard function but also allows different shapes for violating covariates. This method has been introduced by Begun et al. (2019) based on frailty survival models with application to hip replacement outcomes.

In the model, the shape parameter of the baseline hazard function can be specified using the additional, separate Cox-regression term, i.e., the shape parameter $b(\mathbf{u})$ of the hazard function is written as $b\exp(\beta_{\mathbf{u}})$ and depends on the vector of the covariates $\mathbf{u}$ (Begun & Kulinskaya, 2022).

The double-Cox model with shared frailty is a re-parametrized model in the sense that the non-proportional hazard regression model with frailty Z has a Weibull two-parameter (scale and shape) baseline hazard function, where both parameters log-linearly depend on the observed covariates.

The conditional hazard function is defined as

$$\tilde{h}(t \mid \mathbf{u}, Z) = Zh(t \mid \mathbf{u}) = Z\frac{be^{\beta_{\text{shape}}\,\mathbf{u}}e^{\beta_{\text{scale}}\,\mathbf{u}}}{a}\left(\frac{t}{a}\right)^{b\exp(\beta_{\text{shape}}\,\mathbf{u})-1} \tag{3.79}$$

The cumulative conditional hazard function is defined as

$$\tilde{H}(t \mid \mathbf{u}, Z) = ZH(t \mid \mathbf{u}) = Ze^{\beta_{\text{scale}}\,\mathbf{u}}\left(\frac{t}{a}\right)^{b\exp(\beta_{\text{shape}}\,\mathbf{u})} \tag{3.80}$$

$a > 0$ and $b > 0$ are the scale and the shape parameters of the baseline survival distribution, respectively, $\mathbf{u}$ is the vector-column of the covariates, β_{scale} and β_{shape} are the vector-rows of the Cox-regression parameters.

A gamma-distributed frailty Z with mean 1, variance σ^2 is considered for which the probability density function is $f\left(z \mid \sigma^2\right)$ and the "random effect" is defined by $\omega = \ln Z$. The population is divided into N_{cl} clusters and all individuals from the cluster i, $i = 1,\ldots,N_{cl}$, share the same frailty Z_i. The conditional and the marginal survival functions are given by

$$\begin{aligned} S(t \mid \mathbf{u}, Z) &= \exp(-\tilde{H}(t \mid Z, \mathbf{u})), \\ S(t \mid \mathbf{u}) &= \mathbb{E}S(t \mid \mathbf{u}, Z) = \left(1 + \sigma^2 H(t \mid \mathbf{u})\right)^{-1/\sigma^2}. \end{aligned} \tag{3.81}$$

The conditional likelihood function is defined by

$$\mathscr{L}_c\left(\text{ Data } \mid \zeta, Z_1, \ldots, Z_{N_{cl}}\right) = \prod_{i,j}\left(-\frac{\partial}{\partial t_{ij}}\right)^{\delta_{ij}} \exp\left(-\sum_j \tilde{H}\left(t_{ij} \mid \mathbf{u}_{ij}, Z_i\right)\right) \tag{3.82}$$

where $\zeta = (a, b, \beta_{\text{scale}}, \beta_{\text{shape}})$ is the vector of parameters, δ_{ij} stands for censoring (1 if censored and 0 , otherwise) and indices i and j correspond to a cluster and a subject in that cluster, respectively.

3.14.1 Point estimation of the parameters

The marginal likelihood is calculated by marginalizing out the frailty Z in the conditional likelihood (Equation 3.82). From a random effect, the marginal likelihood inherits only parameter σ^2. The k-dimensional vector of the unknown parameters $\zeta_\sigma = \left(\zeta, \sigma^2\right)$ is estimated by maximizing the marginal likelihood function (or its logarithm)

$$\begin{aligned} &\mathscr{L}_m\left(\text{ Data } \mid \zeta_\sigma\right) = \mathbb{E}\mathscr{L}_c\left(\text{ Data } \mid \zeta, Z_1, \ldots, Z_{N_{cl}}\right) \\ &= \prod_{i,j}\left(-\frac{\partial}{\partial t_{ij}}\right)^{\delta_{ij}}\left(1 + \sigma^2 \sum_j H\left(t_{ij} \mid \mathbf{u}_{ij}\right)\right)^{-1/\sigma^2} \end{aligned} \tag{3.83}$$

The marginal likelihood function is a Laplace transform of the frailty distribution calculated at the point $H(t \mid \mathbf{u})$. The Gamma frailty results in a closed-form marginal likelihood.

3.14.2 Confidence intervals for the parameters

Standard-error based confidence intervals

The standard errors of the parameter estimates can be obtained from the inverse of the Hessian matrix (Begun & Kulinskaya, 2022). Let μ be the k-vector parameter and $\hat{\mu}$ its maximum likelihood estimate (MLE). Assume that $\left[\mu_i^l, \mu_i^u\right]$ is the $1-\alpha$ confidence interval for the j^{th} component $\hat{\mu}_j, j = 1,\ldots,k$. The coverage probability for component j is defined by

$$P_{cov}(\mu_j) = \mathbb{P}\left(\mu_j \in \left[\mu_j^l, \mu_j^u\right]\right).$$

This probability can be estimated from simulations as a proportion of cases when the true value of μ_j lies in interval $\left[\mu_j^l, \mu_j^u\right]$.

The scale and the shape parameters a and b of the baseline hazard distribution and the Cox-regression parameters β_{scale} and β_{shape} do not lie on the boundary. But the frailty variance σ^2 can be equal to zero corresponding to the model without frailty. If $\sigma^2 > 0$ and σ^2 is large compared with the standard error for $\hat{\sigma}^2$, the confidence intervals can be specified using the asymptotic normality of MLE:

$$\sqrt{n}(\hat{\mu} - \mu) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \Sigma),$$

where n is the number of observations, $\mathcal{N}$ is the k-variate normal distribution with zero mean and the $k \times k$ asymptotical variance-covariance matrix Σ with elements $\Sigma_{l,m}, l,m = 1,\ldots,k$. In general, for any values of σ^2, the confidence intervals can be calculated using a mixture of the truncated normal distribution and a point mass at zero for $\hat{\sigma}^2$:

$$\sqrt{n}(\hat{\mu}_j - \mu_j) \xrightarrow{d} \Phi\left(\frac{\nu}{\kappa}\right) \mathcal{TN}_1(\mu_\nu, \Sigma_j, s, t) + \Phi\left(-\frac{\nu}{\kappa}\right) \mathcal{N}\left(-\frac{\Sigma_{j,k}}{\Sigma_{k,k}}\nu, \Sigma_{k,k} - \frac{\Sigma_{j,k}^2}{\Sigma_{k,k}}\right)$$

$$\sqrt{n}\hat{\sigma}^2 \xrightarrow{d} \Phi\left(\frac{\nu}{\kappa}\right) \mathcal{TN}\left(\nu, \kappa^2, 0, \infty\right) + \Phi\left(-\frac{\nu}{\kappa}\right) \chi_0^2$$

where χ_0 is a point mass at zero, $\kappa = \sqrt{\Sigma_{k,k}}, \nu = \sqrt{n}\sigma^2, \mu_\nu = (0,\nu)^T, \Sigma_j = \begin{pmatrix} \Sigma_{j,j}\Sigma_{j,k} \\ \Sigma_{k,j}\Sigma_{k,k} \end{pmatrix}$, $j = 1,\ldots,k-1, \mathcal{TN}_1(x,V,s,t)$ is the marginal distribution of the first component of a truncated bivariate normal distribution with mean vector x, covariance matrix V, and the lower and upper truncation limits $s = (-\infty, 0)^T$ and $t = (\infty,\infty)^T$, respectively. Begun and Kulinskaya (2022) refer these intervals as standard-error-based.

Calculating confidence intervals of hazards and survival functions using bootstrapping method

The hazards and survival probability at time t of a specific profile are calculated by fitting $\zeta = (a, b, \beta_{\text{scale}}, \beta_{\text{shape}}, \sigma^2)$ parameters and covariates $\mathbf{u}$ to Equation 3.79 and Equation 3.81, respectively.

To calculate their confidence interval at a specific time point t, the parameters ζ and their corresponding standard errors (derived from the Hessian Matrix) are used. N (say 100,000) bootstrap samples of parameters are then produced from the original sample of parameters. They are eventually fitted to simulate N hazard or survival estimate at that time point. The 95% bootstrap confidence interval for a point estimate is then obtained by finding the 2.5^{th} percentile and the 97.5^{th} percentile.

3.15 Actuarial Translation of the survival models

With regards to the actuarial application, the survival models after IS and TIA, could not be easily translated into life expectancy due to the time scale being "time since study entry".

The original survival models after TIA and IS were useful for medical application and interpretability. The models included predictors like birth cohort and age categories at diagnosis of stroke and TIA.

For the actuarial implementation of the model, using age as continuous was more relevant. Changing the time scale to age (continuous) at entry resulted in illogical findings and this was accounted for by the dependency of survival time and age at diagnosis. The model was then changed by setting the time scale as the time of entry but considering the age at entry as a continuous predictor in the model. The model provided hence provided important insights into finding benefits or harms associated with certain medications, lifestyle interventions, and medical conditions.

The models for actuarial translations for the case of IS and TIA are presented in Tables C.7 and B.9. In TIA model, significant shape parameters were obtained for birth year category, diagnosis of heart failure and antiplatelets prescriptions and significant scale interactions between cases/controls status and antiplatelets. In the IS model, birth cohort and prescription of antiplatelet had significant shape effects. Significant scale interaction of cases/controls status with hypertension was present. The model building for life expectancy is explained in Chapters 5 and 6. The formula below shows how to calculate the expected life expectancy. Using relevant parameters (from the aforementioned models), covariate values, and time points, the survival probabilities and the integral of the survival probabilities can be calculated which are then used as an expectation as :

The remaining life expectancy at age z is

$$e(z) = \frac{\int_z^\infty S(t)dt}{S(z)} \tag{3.84}$$

where $\int_z^\infty S(t)dt$ is the survival time from age z till limiting age ω adjusted by number of survivors, $S(z)$. The functions $\int_z^\infty S(t)dt$ and $S(z)$ can be found using the Weibull double Cox model using equation (following from Equation 3.81):

$$S(t|a,b,\sigma^2,\beta_{\text{scale}},\beta_{\text{shape}}) = \quad (1+\sigma^2 H(t|a,b,\sigma^2,\beta_{\text{scale}},\beta_{\text{shape}})^{-1/\sigma^2}$$

where $H(t|a,b,\sigma^2,\beta_{\text{scale}},\beta_{\text{shape}})$ is the cumulative hazard which requires input parameters such as : time is the time past from diagnosis t, u is the covariate vector, a and b are slope and shape parameters defining the hazard functions (Weibull), β_{shape} shape and β_{scale} scale are the Cox-regression parameters for shape and scale, respectively, and σ^2 is the variance of frailty.

3.16 Missingness

In longitudinal studies, where subjects are followed over some extended period of time, follow-up makes missing data fairly inevitable. It subsequently presents a major challenge for analysis (Yeatts & Martin, 2015). For nearly a century, scientists have been deleting missing cases or arbitrarily filling in missing cases, which may have caused bias (Dziura et al., 2013). Great advances have been made in terms of analytical techniques to deal with missing data. The following section presents the different types of missing data mechanisms and gives an overview of handling missing data.

In the current study, we faced missing data problems when patients' medical measurements such as BMI, smoking, or alcohol status, were unavailable. This could be attributed to patients missing visits to the medical practices for numerous reasons or non-response. Andrinopoulou (2014) explains that missing values in longitudinal studies occur in fundamentally two different ways. The first type is when patients' data are missing at random times, meaning that other measurements are observed following missing values. The second type of missing data occurs when data is not available for a subject after some time point, and the patient is said to have dropped out of the study or censored (in our case, when a patient is transferred out of medical practice).
The main concern in longitudinal analysis with missing data arises when there is an association between the longitudinal profile and the missing process. In order to decide on the appropriate methods to analyse the incomplete longitudinal data, we need to know the missing data mechanism. There are three types of mechanisms, namely (R. Little et al., 2002);

- **Missing Completely at Random (MCAR)**: When the probability that the responses are missing, is unrelated to any variables. For example, when a patient forgets to attend an appointment or is transferred out of practice (e.g. subject moved to another GP practice).
- **Missing at Random (MAR) data** is caused when the probability of missingness depends on recorded side-effects, on known baseline characteristics or their conditions are not controlled sufficiently according to protocol criteria (Dziura et al., 2013).
- **Missing Not at Random (MNAR)**: When the probability that the longitudinal responses are missing depends on observed and unobserved data.

So in the case of MCAR, the incomplete observed data could be considered as a random sample of the complete data, Andrinopoulou (2014). Therefore, under this assumption, we could proceed with the analysis by using only the observed data. Conversely, in the case of MAR mechanism, the missing observations are no longer random, considered as a random sample of the complete data. Hence, the missing process must be modelled together with the longitudinal process. Data under the MAR assumption could be ignorable while deriving valid inferences (Andrinopoulou, 2014). MNAR is more general and represents the most complex missing data scenario. In this case, a joint distribution of the longitudinal and the missing processes is required in order to obtain valid inferences. Different types of modelling mechanisms for handling missing data in the longitudinal setting have been considered in the literature. Some of the examples include selection models, pattern mixture models and shared-parameter models (R. J. Little & Rubin, 1987; Molenberghs & Kenward, 2007). Selection and pattern mixture models are applied for discrete times.

There is no discernible difference between observed and unobserved data using MCAR data (Graham et al., 2009). For instance, if the equipment fails, the BP values or blood glucose readings may not be recorded. When using MAR data, the observed data completely accounts for the systematic difference between observed and unobserved data (Graham et al., 2009). As older patients are more likely to have a record of their blood glucose readings or cholesterol readings, recorded cholesterol readings or blood glucose readings may, in this case, be higher than missing readings given that the data include both age and medical readings. There is a systematic difference between observed and unobserved data in MNAR data, and unobserved data can at least partially account for this disparity (Graham et al., 2009). Since the data include both medical readings and drug therapy, for instance, individuals with high cholesterol or blood glucose readings who do not adhere to drug therapy may be less likely to attend doctor appointments. This could imply that patients who don't adhere to medication therapy have higher recorded cholesterol or blood glucose values than those who do. It is known that there is a systematic discrepancy between observed and unobserved data in primary care data Marston et al. (2010).

People who are sick will see their general practitioner more frequently, who then are more likely to collect background information on these patients compared to healthier individuals. Because of the

correlation between lifestyle and medical disorders, information on lifestyle is primarily recorded Marston et al. (2010). Recording has significantly improved in primary care since the Quality and Outcomes Framework (QOF) was introduced in 2004 as a compensation system to enhance the quality of healthcare provided by general practitioners. Regarding the history of medical conditions, lifestyle, and sociodemographic factors, patients with and without complete medical records were compared to one another for this study. It was considered that patients who had not received a diagnosis or treatment for a medical ailment did not have the condition or did not get treatment for it. This indicates that the only lifestyle factors with missing information were those related to body mass index, blood pressure, alcohol usage, and smoking. Chi-square tests were used to determine whether there was a significant difference in the proportion of missing data in a lifestyle covariate by the medical conditions, therapies, socio-demographic characteristics, and other lifestyle covariates. The Chi-square of independence results shows that individuals with and without complete medical data consistently vary from one another.

Multiple imputation entails a number of model construction steps and associated missing data assumptions, where any poor choices could result in skewed estimations. The most crucial assumption is that the data are missing (totally) at random (Allison, 2001). Although the type of missing data cannot be directly verified, knowing why the data might be missing could help in deciding whther to use or not to use multiple imputation. Multiple imputation has been demonstrated to be a useful technique for producing accurate estimates when there are missing data in primary care data (Marston et al., 2010). For the purposes of this study, it was therefore considered that the observed data on medical conditions, treatments, soci-demographic characteristics, and lifestyle factors could completely account for the cause for missing records and that the missing records were therefore absent at random. Multiple imputation was utilised to address bias and imprecision brought on by the presence of missing data because there were missing data that ranged from more than 20% to less than 50% in the TIA and IS cohorts.

3.16.1 Approaches to missing data

Incorrectly explaining the missing data mechanism during analysis, could lead to flawed conclusions. The following is an overview of analytic approaches that are employed to handle missing data. Four general strategies are devised to cope with missingness (Carpenter et al., 2002).

1. Use only data from participants completing the study with no missing data (complete-cases analysis).

2. Impute (either single or multiple) values for missing data and analyse with complete case models.

3. Develop a model for the data that includes a model for the missing data process.

Complete-case analysis

According to Yeatts and Martin (2015), complete-case analysis results in a biased estimate of the treatment effect in the absence of data rising from an MCAR mechanism; there is a loss of precision in the estimation of the treatment difference (larger standard errors and wider confidence limits, a loss of power in the test of significance).

Maximum likelihood

Allison (2001) explains that maximum likelihood is an excellent method for handling missing data in a wide variety of situations. Maximum likelihood have the desirable properties : *consistency*, *asymptotic efficiency* and *asymptotic normality*. It works with data that are MAR but not that are MCAR. The likelihood can be maximized using either the expectation-maximization (EM) algorithm or direct maximum likelihood.

Imputation

Imputation has been described as " the practice of filling in missing data with plausible values" (J. L. Schafer, 1999). Single imputation approaches assign a single value to the missing result; multiple imputation (MI) methods repeat the process of imputing a response to account for uncertainty in the outcome assignment. (Yeatts & Martin, 2015).

Last Observation Carried Forward is a popular type of single imputation that can be effective in longitudinal trials since it replaces the missing result with the last observed outcome assessment. The approach assumes that the participant's responses (e.g., outcome measures) would have been stable from the point of dropout to trial completion, rather than declining or improving further. The assumption's validity could be influenced by the time of the last accessible assessment in relation to the missing outcome, as well as the response observed at the last available assessment (Yeatts & Martin, 2015).

Multiple imputation is a Monte Carlo technique in which the missing values are replaced by $m > 1$ simulated versions, where m is typically small (say, 5 to 10). In R. J. Little and Rubin (1987)'s method for 'repeated imputation' inference, each of the simulated complete datasets is analysed by standard methods, and the results are later combined to produce estimates and confidence intervals that incorporate missing-data uncertainty.

MI is generally preferred over single imputation because the inference correctly reflects the uncertainty associated with the procedure. Multiple imputed data sets are generated and analysed, and the inference from each is statistically combined to estimate the treatment effect and its corresponding SE and p-value. For instance, outcomes could be imputed via logistic regression, where favourable outcome is predicted according to baseline characteristics (e.g, stroke severity, age, and sex).

3.17 Multiple Imputation

Although the maximum probability is a great strategy for dealing with missing data, it does have some drawbacks. The main drawback is that all of the variables must be assigned to a joint probability distribution, which is not always straightforward to find. MI is a fantastic alternative with statistical features that are essentially identical to maximum Likelihood.

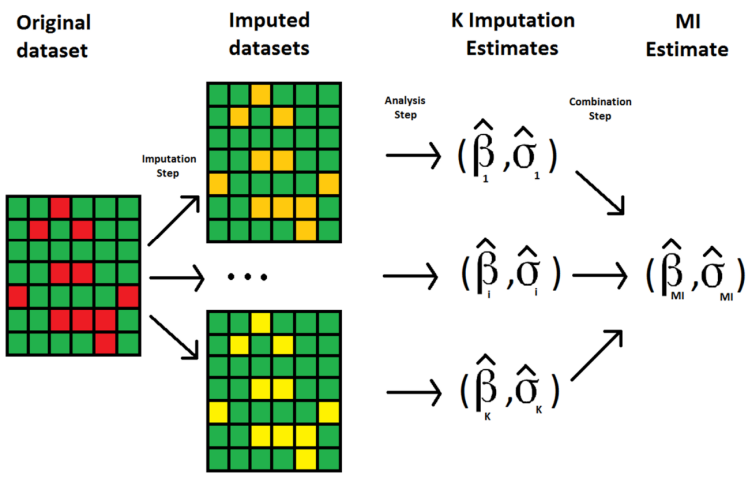
Multiple imputation works effectively when the data is MAR rather than MCAR, which is a more stringent condition. All covariates from the analysis model and covariates linked to missing data should be included in the imputation model (Van Buuren, 2018). The ideal number of covariates to include is up to 30 (Van Buuren, 2018). If the imputation model is flawed, the analysis model's results may be biased when it incorporates more covariates than the analysis model. The predicted values of a covariate provided by a regression model that regresses that covariate on the other covariates would be a more appropriate replacement for missing observations of that covariate. Since the factors related to the cause of the missing data are taken into account in the regression model, imputation by a regression model yields estimates that are unbiased. The imputation, on the other hand, understates the standard errors, giving the results deceptive precision. The measurement scale of each covariate with missing data determines the imputation model. Linear regression is used to impute continuous variables. The imputation model's goal is to provide a range of reasonable values, not to impute the values that would have been seen if the data were not missing (Van Buuren, 2018), hence the imputed covariate does not need to be normally distributed. A logistic regression is used to impute binary factors. A multinomial regression is used to impute covariates with more than two categories.To produce unbiased estimates, covariates should be imputed on their original measurement scale and transformed after imputation. Any covariates generated from covariates with missing records should also be imputed rather than being calculated using the passive imputed values (Van Buuren, 2018). This is done to guarantee that the imputed values match the analysis model. Covariates were measured on several scales for both TIA and IS studies, with body mass index, being continuous measures while smoking status and alcohol consumption were categorical. Due to the clustering of patient medical records by general practice, the data had a hierarchical structure. For mixed continuous-categorical data with a hierarchical structure and a generic missing pattern, JOMO imputation was used to describe such imputation model . This software uses joint modelling to impute missing data.

JOMO is a **R** package to extend this to allow for a mix of multilevel (clustered) continuous and categorical data. Our data is multi-level given the patients have been assumed to be nested in the GP practice clusters.

The Markov Chain Monte Carlo (MCMC) algorithm, which is based on linear regression, is the most extensively used approach for multiple imputation. After producing predicted values based on linear regressions, random draws are done for each regression equation from the (simulated) error

distribution. The imputed values are created by adding these random 'errors' to the projected values for each individual (Allison, 2001).

FIGURE 3.4: Illustration of the Multiple Imputation process.



The first stage involve producing multiple imputed datasets based on the regression model. The next stage is running the model on the imputed dataset and finally, pooling the estimates using the Rubin's Rules.

Source : Quartagno et al. (2019)

To introduce the general ideas of joint modelling multiple imputations, we consider a data set made up of N observations on three continuous variables; we further assume that one of these variables, X, is fully observed, while the other two, Y_1 and Y_2, have missing values. The main concept underlying joint modelling imputation is to create a joint multivariate model for all of the variables in the data set that will be utilised for imputation. The multivariate normal model is the easiest joint model to understand:

$$
\begin{aligned}
Y_{1,i} &= \beta_{0,1} + \beta_{1,1} X_i + \varepsilon_{1,i} \\
Y_{2,i} &= \beta_{0,2} + \beta_{1,2} X_i + \varepsilon_{2,i} \\
\begin{pmatrix} \varepsilon_{1,i} \\ \varepsilon_{2,i} \end{pmatrix} &\sim \mathcal{N}(\mathbf{0}, \mathbf{\Omega})
\end{aligned}
$$

After choosing the imputation model, Bayesian methods are used to impute the missing data with the use of Gibbs sampling which deals with missing data using a data augmentation algorithm (Quartagno et al., 2019). This entails taking new values from the relevant conditional distribution for each of the model's parameters one at a time. The parameters of the model are the fixed effects β and the covariance matrix Ω, and the missing data. The process stops when the stationary distribution is reached. The first sample of missing values is mixed with the observed data to make the first imputed

data set. The sampling is run for an adequate number to ensure guarantee stochastic independence between consecutive imputations.

Given the Bayesian method, prior distributions are important and the package uses flat priors for all the parameters, except for covariance matrices. If observations in the data set are nested in J clusters, the model can be changed to accommodate random intercepts as follows:

$$
\begin{aligned}
Y_{1,i,j} &= \beta_{0,1} + u_{1,j} + \beta_{1,1} X_{i,j} + \varepsilon_{1,i,j} \\
Y_{2,i,j} &= \beta_{0,2} + u_{2,j} + \beta_{1,2} X_{i,j} + \varepsilon_{2,i,j} \\
\begin{pmatrix} \varepsilon_{1,i,j} \\ \varepsilon_{2,i,j} \end{pmatrix} &\sim \mathcal{N}(\mathbf{0}, \mathbf{\Omega}) \\
\begin{pmatrix} u_{1,j} \\ u_{2,j} \end{pmatrix} &\sim \mathcal{N}(\mathbf{0}, \mathbf{\Omega}_u)
\end{aligned}
$$

When running the imputation, the number of burn-in iterations for the MCMC sampler, the number of iterations between imputations, and the number of imputations need to be provided.

By using latent normal variables, for instance, binary and category data can be incorporated into the imputation model. If Y_1 is binary in this model, a latent continuous variable Y_1^* is added to it, so that $Y_{1,i}^* > 0$ for individuals i for whom $Y_{1,i} = 0$ and $Y_{1,i}^* \leq 0$ for individuals for whom $Y_{1,i} = 1$ (H. Goldstein et al., 2009). Similar to this, $K-1$ latent normal variables that indicate the distinctions between categories can be used to represent a categorical variable with K categories. As opposed to models without categorical data, this approach necessitates limitations on the variance-covariance matrix to provide model identifiability. These three functions, *'jomo1rancon'*, *'jomo1rancat'*, and *'jomo1ranmix'*, respectively, impute clustered continuous, categorical, and mixed data. *'jomo1rancat'* and *'jomo1ranmix'* use the latent normal model for the categorical variables. In the imputation model, each of these functions has a constant, common covariance matrix that spans all of the clusters (level-2 units).

The running of the imputation model can be time-consuming and computationally intensive if a large dataset is involved, so it is good practice to run the MCMC chain function with a small number of burns, say 2, and observe the resulting output. Then, the function can be run for more iterations, and trace plots can be useful in deciding the correct number of burn-in and between-imputation iterations (Quartagno et al., 2019). It is important to check if MCMC chains converge during imputations. The 'jomo.MCMCchain function' allows to check it.The substantive model is fitted to the imputed data sets and pooled using Rubin's rules.
In particular, five to ten imputed datasets—which is the standard recommendation for multiple imputation (Van Buuren, 2018; Zhou & Reiter, 2010). The inferences drawn from five imputed datasets

will probably remain the same even if we impute more than five times(Van Buuren, 2018). The default 10 value was used for this study. Using 10 imputed datasets resulted in a similar selection of prognostic indicators (See Chapters 5 and 6). Bar and density plots were used to verify the imputed values. The imputed values should have a similar distribution to the recorded observations if the missing data were lost at random (Van Buuren, 2018). This supports Rubin (1987)'s claim that 5 to 10 imputed datasets are sufficient to attain high efficiency. Each of the imputed datasets underwent individual analysis, including covariate selection.

3.17.1 Rubin's Rules

Rubin (1987)'s Rules are used to combine the results. For parameter estimates (e.g., regression coefficients), the combined estimate $\bar{\theta}$ is the average of the estimates from the m imputed data analyses:

$$\bar{\theta} = \frac{\sum_{j=1}^{m} \theta_j}{m}$$

$$\text{Var}(\bar{\theta})_{\text{within}} = \frac{\sum_{j=1}^{m} \text{Var}(\theta_j)}{m}$$

The between imputation variance $\text{Var}(\bar{\theta})_{\text{between}}$ is the sum of the squared deviation of the parameter estimate of each imputed data analysis from the pooled parameter estimate weighted by $m-1$:

$$\text{Var}(\bar{\theta})_{\text{between}} = \frac{\sum_{j=1}^{m} (\theta_j - \bar{\theta})^2}{m-1}$$

The variance of the parameter estimates is then calculated by combining the within and between variance:

$$\text{Var}(\bar{\theta}) = \text{Var}(\bar{\theta})_{\text{ within}} + \left(1 + \frac{1}{m}\right) \text{Var}(\bar{\theta})_{\text{ between}}$$

For significance testing of the pooled parameter, a univariate Wald test (Van Buuren, 2018) is used as follows :

$$Wald_{\text{Pooled}} = \frac{(\bar{\theta} - \theta_0)^2}{V_T}$$

where $\bar{\theta}$ is the pooled mean difference and V_T is the total variance and is equal to the SE_{Pooled} (pooled standard error) and θ_0 is the parameter value under the null hypothesis (which is 0) and which follows a t-distribution. The p-value can hence be derived. The value for t depends on the degrees of freedom, according to:

$$t_{df,1-\alpha/2}$$

Where df is the degrees of freedom and α is the reference level of significance. The derivation of the degrees of freedom for the t-test is complex. Because t^2 is equal to F at the same number of

degrees of freedom, we can also test for significance using an F-distribution, according to:

$$F_{1,df} = t^2_{df,1-\alpha/2}$$

3.17.2 How many imputations are required?

Multiple imputation can yield accurate estimates and confidence ranges from a small number of imputed datasets (m), even as little as two (Van Buuren, 2018). Multiple imputation is effective when m is low because it increases the variance $\mathrm{Var}(\bar{\theta})_{\text{between}}$ between imputations by a factor of $1/m$ before computing the overall variance in $\mathrm{Var}(\bar{\theta})$, that is

$$\mathrm{Var}(\bar{\theta}) = \mathrm{Var}(\bar{\theta})_{\text{within}} + \left(1 + \frac{1}{m}\right) \mathrm{Var}(\bar{\theta})_{\text{between}}$$

For moderate amounts of missing data, the conventional suggestion is to use a low number of imputations—between 3 and 5. To achieve certain efficiency, it is then advisable to set m high such as >20. The impact of m on various parts of the outcomes was examined by several authors.
The following justification underpins the recommendation for low m. Since multiple imputation is a simulation approach, simulation error can affect both the difference between the expected value of the estimate and the true value ($\bar{Q}$) and its variance estimate $T = \mathrm{Var}(\bar{\theta})$. Theoretically, all simulation errors are eliminated when m is set to ∞, achieving T_∞. Rubin (1987) demonstrated the two variances T_∞ and T_m are connected by according to

$$T_m = \left(1 + \frac{\gamma_0}{m}\right) T_\infty$$

where γ_0 is the amount of missing information. So Van Buuren (2018) explains that setting m to be 5, would make the confidence interval longer by 3% than the ideal confidence interval, and increasing m to 10 to 20, the factor would decrease by 1.5% and 0.7%, respectively. According to J. Schafer et al. (1996), creating and analysing more than a few imputed datasets rarely serves any useful purpose in some cases and J. L. Schafer and Olsen (1998) maintains that more than a few imputations would require more resources to build and maintain, which would not be an efficient use of resources.

According to Royston (2004), m must be "at least 20 and maybe more". He advocates that the length of the confidence interval also depends on degrees of freedom and thus on m. The impact of m on a test's statistical power for detecting a tiny effect size (0.1) was examined by Graham et al. (2009). In instances where strong statistical power is required, they advise setting m high and proposed to use $m = 20$ to 100 for different percentages of missing information.

Bodner (2008) suggested using a linear rule with m = (3, 6, 12, 24, 59, 114, 258) for missing percentage, γ_0 = (0.05, 0.1, 0.2, 0.3, 0.5, 0.7, 0.9) after investigating the fluctuation of the length of the 95% confidence interval, the p-value, and γ_0 under varied m. By using a quadratic rule, Von

Hippel (2020) demonstrated that a relationship between m and γ_0 is more understandable. Using a two-step process, he proposed that missing information of less than 0.5, to use a lower m and higher m, otherwise. The first phase of the two-step process uses the rule estimates γ_0 and its 95% confidence interval.

Higher m should always be used, according to theory, but doing so requires more computation and storage. Setting m high might not be worthwhile if the point estimates—rather than standard errors, p-values, and other statistics—are what matter most. Under moderate missingness, employing $m = 5 - 10$ will be sufficient (Van Buuren, 2018).

3.17.3 Variable selection after Multiple Imputation.

Three stages make up the typical multiple imputation method:

- Imputation of m times the missing data;
- Analysis of the datasets that were "m imputed"
- Combining the parameters from m different analyses.

This scheme is difficult to apply if step-wise model selection is part of the statistical analysis in phase 2 (Van Buuren, 2018). Step-wise variable selection techniques may produce sets of variables that vary between the m datasets. When the imputation model has more covariates than the analysis model, the analysis model's findings may be biased if the imputation model is wrong (J. L. Schafer & Graham, 2002). When the analysis model has more variables than the imputation model, the analysis model's conclusions are still valid but may understate the impact of the covariates that were left out of the imputation model if the analysis model is accurate. For the study, the covariate selection was carried out on each of the imputed datasets. The imputed datasets were analysed independently. Extra covariate selection did not happen during the process but if was the case, the Brand (1999)'s method would have been used.
Brand (1999)'s method is a two-step process for choosing covariates following MI. On each imputed dataset independently, step-wise model selection is first performed. Next, a new supermodel is built, containing all the variables that appeared in at least half of the first models. The purpose behind this criterion is to exclude variables that were unintentionally chosen. Backward elimination is then applied. Another approaches could also be envisaged like using a Bayesian framework (Yang et al., 2005) and scenario-based framework (Van Buuren, 2018; Wood et al., 2008) like : 'stack'(Combine the imputed datasets into a single dataset, give each record a fixed weight, and use the standard variable selection techniques.), 'majority'(a process that chooses final variables from the pool that are present in at least half of the models) and 'Wald'(The Wald statistic derived from the multiply imputed data serves as the foundation for step-wise model selection.)

3.18 Chapter summary

In this chapter, the statistical methods used in the analysis of survival analysis are presented. Non-parametric models are used for the analysis of data when the distributional assumption is not required. Parametric models are used for survival data assuming a defined distribution. Additionally, the Cox's proportional hazards model presented in this chapter can be extended to accommodate time-dependent, stratified variables and frailties. The novel double- Cox Weibull model is presented. The Multiple Imputation method is explained.

4 Data Selection and Description

This section describes the THIN database, the inclusion-exclusion criteria and the extracted variables.

4.1 The Health Improvement Network (THIN) database

This thesis makes use of a retrospective cohort study. Electronic primary care medical records were extracted from The Health Improvement Network (THIN). This section of the chapter provides a general overview of the methods used. Healthcare databases are an excellent source of data on the incidence and prevalence of many diseases and can provide much useful information on stroke morbidity. While most studies have endeavoured to estimate the incidence of stroke using hospital discharge data, only a few have been able to validate the accuracy of the data. It has been shown that discharge data can overestimate the incidence of stroke (Ellekjær et al., 1999). The THIN data has widely been used for epidemiological studies (Denburg et al., 2011).

THIN is a database that contains anonymised electronic patient records from more than 580 general practices from the UK Blak et al. (2011). The THIN database has data on 12 million patients, of whom 3.6 million are current and the rest are former (left the practice) or deceased patients, and represents around 6% of the UK population. It also includes complete prescription records, which are particularly relevant for pharmaco-epidemiological research, as well as data on hospitalisations and referrals, making it a valuable source of stroke morbidity data. Since 2002, THIN has been collecting data from general practices. It all started with a partnership between the makers of Vision software (InPractice Systems), which is used by some general practices, and the Epidemiology and Pharmacology Information Core (EPIC), which provides primary care data for medical research. (Taggar et al., 2012).

Every GP consultation from participating clinics is included in the data. Read codes are a hierarchical coding system that makes it simple to record patient events such as diagnosis, symptoms, procedures, and test findings. Data on pharmacological therapy administered to patients are also kept track of. To safeguard patient privacy, the combined data is anonymised.

4.2 Structure of the THIN database

The structure of the THIN database is summarised in Figure 4.1. Data is organised by general practice then patient and linked by practice ID and patient ID. There are practice files, four main files

(patient, therapy, medical and additional health data (AHD) files) and three linked files (postcode variable indicators (PVI), staff and consult files), which are described below.

4.2.1 Practice file

The practice file is a separate file for each general practice, which includes the date of computerisation, acceptable mortality reporting (AMR) date (when the practice mortality rate is similar to the UK mortality rate), date of last data collection and country. This file is linked to the four main files below.

4.2.2 Main files

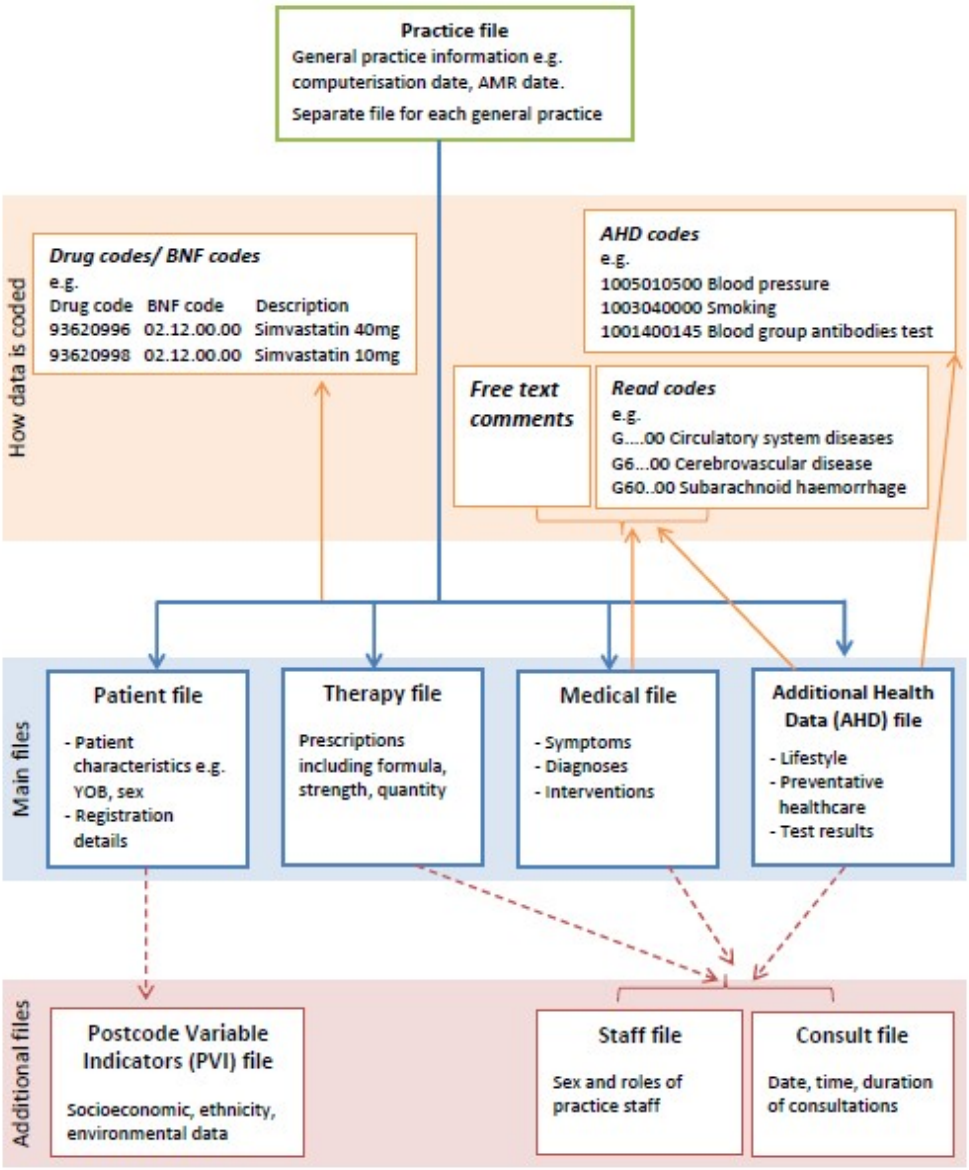
1. **Patient file**: Demographic information (including sex and year of birth) and registration information (such as registration date and date patient left the practice).
2. **Therapy file**: Data on prescriptions (including strength and formulation) which is automatically recorded when a GP or nurse issues a prescription.
3. **Medical file**: Diagnoses and symptoms, which are recorded at consultations and information, provided from secondary care discharge notes.
4. **AHD file**: Other information includes lifestyle factors (such as smoking and alcohol), information for preventive healthcare (such as height, weight and cholesterol) and lab test results.

4.2.3 Linked files

1. **PVI file**: Socioeconomic and environmental data at an area level to maintain anonymity (linked to the patient file).
2. **Staff file**: Data on the sex and roles of general practice staff (linked to the therapy, medical and AHD files).
3. **Consult file**: Information on consultations including date, time and duration (linked to the therapy, medical and AHD files).

With the exception of free text, all data in THIN is displayed as coded information. Read codes are used to code clinical data including diagnoses, procedures, investigations, and signs and symptoms. The read codes are organised in chapters and categories and are made up of seven characters (e.g. G64z200: left-sided cerebral infarction). Diagnoses are also coded according to the International Classification of Diseases-9 (ICD-9) system. AHD codes are used to code additional clinical information, such as clinical measures whereby some measures can be numeric for example weight, BMI, etc. The therapy file includes drug codes for specific drug formulations as well as British National Formulary (BNF) codes depending on BNF chapter.

FIGURE 4.1: Structure of The Health Improvement Network (THIN) database



4.3 Data Extraction

The study design was defined as follows:

- Study entry period: 1986-2017
- Study follow-up period: 1986-2017

4.3.1 Selection criteria

- Medical records selected from 1986 onwards.
- Medical records selected after the acceptable mortality reporting (AMR) date of the practice. The Acceptable Mortality Reporting (AMR) date is determined by QuintilesIMS (provider and licensor of THIN) internally and is determined by comparing trends in mortality reporting for each particular THIN practice to the forecasted death rates derived from national statistics given the practice's demographics. The AMR date is the year that the practice is assumed to have started proportionally reporting all-cause mortality based on these statistics. The use of this variable in data filtering before the selection of cases minimizes the possibility of biases in disease occurrence and ensures that there are no "immortal" periods present in the data.
- Medical records selected of practices with an active status. A patient is considered to be "currently active" if they are listed with an active THIN practice and have not passed away or changed practices before the practice's latest collection date.
- Data pertains to adults born in or before the year 1960.

Data flags used in THIN database

The patient Flag indicates the integrity of the data for the patients and the Registration Flag provides an indicator of any data issues related to the patient's registration to practices. The flags that are used in the THIN database are presented in Table 4.1 and Table 4.2.

TABLE 4.1: Patient Flag code used in THIN database

Patient Flag	Description
A	Acceptable record
C	Acceptable: transferred out dead without additional death information found in the data
D	Not permanently registered
E	Out of sequence Year of Birth (YOB). YOB greater than registration date.
F	Out of sequence registration date. i.e. greater than transferdate.
G	Regstat 5 and missing or invalid transfer out date
H	Missing or invalid registration date
I	Year of birth missing, invalid or over 115 years of age
J	Not male or female
K	Invalid transfer out date
N	Family number invalid
P	Invalid Regrea
Q	Out of sequence deathdate i.e before YOB or greater than last collection
R	No registration time i.e registration date = last collection or transfer out
S	Acceptable but no medical, therapy or AHD events recorded
M	Multiple problems. More than one of the above errors
X	Re-allocation of patient ID : 2 different patients with same patient ID

Practice Inclusion Criteria

In an endeavour to obtain an established cohort of participating practices for the purpose of the analysis, a contributing practice was chosen as follows from greater of the following: One–year post Vision installation date and acceptable mortality reporting date.
The justification for choosing practice with one year after installation of the Vision practice software is to ensure that the software was sufficiently implemented. Practices were only included if practices had an end date later than 09^{th}January 2017; those who stopped contributing to THIN before 09^{th} January 2017 were excluded. Subsequently, we identified a stable cohort of 392 practices that had been continuously active in THIN from 1^{st} January 1986 to 09^{th} January 2017 for the stroke research.

Patients Inclusion Criteria

Patients were included if they had been registered at the practice for 12 months or more and had met the following criteria:

- Patient flag was either ‘A’ or ‘C’ indicating that the patient record had passed the internal THIN validation process and was an ‘acceptable record’ or was an ‘acceptable record and transferred out of data due to the patient being dead’.

TABLE 4.2: Registration Status Flag used in THIN database

Registration Status (Regstat)	Description
01	Applied
02	Permanent
03	Temporary resident <16 days
04	Temporary resident 16 days to 3 months
05	Transferred out
07	Immediately necessary treatment
08	Emergency treatment
09	Child Health Surveillance
10	Contraception
11	Maternity
12	Minor surgery
13	Private
14	Referred
15	Walk in centre
16	GP with special Interest
17	Minor Injury Clinic
18	HMP inmate
19	Visitor (EC111)
99	Death
00	Null record

- Registration flag was either '01', '02', '05' or '99' indicating that a patient has 'Applied', or 'Permanent' residence or patient has 'transferred out' of the practice or patient has died'.
- Patient start date was earlier than 09^{th} January 2017 and the patient start date was earlier than the patient end date.

Selection of cases

Stroke patients were selected if they have had the ischaemic or transient ischaemic attack-type of stroke for the first time in 1986 or later. A search was done for the earliest reference of an occurrence of stroke according to the defined Read Codes (See Appendix 1). Patients with a Read code indicating TIA/IS stroke, but missing diagnosis dates were excluded. Patients who had practice status W (Withdrawn) with diagnosis date before AMR date were excluded.

The Read Codes followed the NICE clinical pathway as per Figure 2.5. Read codes suggesting TIA/ stroke mimics were filtered out. Read codes of TIA used for the study were based on the transience of the neurological symptoms after mimics had been ruled out at initial assessment (such as blood sugar test to exclude hypoglycaemic mimics). Read codes for IS were based on the persistence of neurological symptoms and diagnosis of confirmed infarction through brain imaging results according to NICE (2020b). The TIA cohort study considered TIA patients after filtering out mimic cases, ischaemic and hemorrhagic stroke cases.

Medically qualified colleagues reviewed the Read codes for the inclusion of diagnoses of stroke.

Further Exclusion criteria: Some medical conditions

Medical records of patients with major cancer (melanoma, pancreatic and lung cancer, metastatic cancer), dementia, chronic kidney disease stage 3+ and other types of stroke (haemorrhagic) before the corresponding event date for each patient in cases. The rationale for excluding haemorrhagic stroke was to allow a stable cohort of patients as haemorrhagic stroke tends to be more fatal than IS/TIA types of stroke.

Selection of controls

Select medical records of patients who have not had any type of stroke ever matched by age, sex and practice by that date. Matching of $1:3$ was performed.

4.3.2 Variables of interest

Drugs:
Anti-hypertensive drugs, Anticoagulant drugs, Lipid regulating drugs and anti-diabetic drugs.
Medical conditions:
Asthma, Atrial Fibrillation, Chronic Kidney Disease, Coronary heart disease, Peripheral Arterial

Disease (PAD), Hypothyroidism, Chronic Obstructive Pulmonary Disease, Diabetes, Hypercholesterolemia, Hypertension, and Myocardial Infarction.
Other (Demographical and lifestyle conditions):
Latest Blood-Pressure, HDL Cholesterol, Serum Cholesterol, BMI, gender, date of birth, age at a point in time when diagnosed stroke, smoking status, and alcohol status.

4.4 Dataset

From 3,515,292 patients, we found 171,326 patients with ischaemic stroke diagnosis and 151,818 patients with a transient ischaemic stroke diagnosis.
From 171,326 patients with ischaemic stroke diagnosis and 151,818 patients with a transient ischaemic stroke diagnosis, we selected patients:

- with dates of diagnosis after acceptable mortality reporting
- with dates of diagnosis after at least 12 months of registration date
- if registration dates of patients are after computerization date, then data flag (flag to indicate data issue) of practice does not matter, however, if the registration dates of patients are before computerization date, then select patients with null data flag (flag to indicate data issue)
- Only select patients with patflag= 'A'. A patflag is a flag to indicate the integrity of the data for that patient and patflag= 'A' means acceptable record.

After matching with controls, the total number of patients in the datasets changed. More details of the number are found in the flowcharts in Chapter 4 and Chapter 5.

4.5 Strength and Limitations of THIN Database

The THIN database's main advantage is that it is demographically representative of typical primary care. Non-interventional since data is obtained on a regular basis (Turner, 2016). Clinical and prescription data are abundant in the THIN database. As the data is validated using vision software, the chances of inexplicable figures being submitted are quite slim. As cases and controls may be collected from the same population source, the THIN data is very apt to be used in epidemiological study designs.

They have been widely validated to be used for producing population-level estimates (Denburg et al., 2011; Public Health England, 2018). Data collected in THIN includes both primary and secondary care data, meaning out-of-hospital health events are more likely to be captured (Public Health England, 2018). THIN contains data from 6% of the population of England; this is a large database, and it has been shown to be representative of the general population (Blak et al., 2011; Denburg et al., 2011; Hippisley-Cox et al., 2007).

Data are continuously updated, allowing for robust time-series analysis of events in both the cohort and individual patients. The data can be used for longitudinal research allowing for long-term follow-up of patients (Public Health England, 2018). THIN has been used extensively in other epidemiological studies, for example in illustrating the prescribing of statins for the primary prevention of CVD, frequency of antibiotic prescribing as well as underuse of prevention drugs in the prevention of strokes and TIA (Turner, 2016).

However, when evaluating the data, there are a few caveats to keep in mind. The strategy implies that the epidemiology of stroke in the small variety of practices that participate in THIN accurately represents both behavioural and fixed risk factors for stroke in England's general population. Only the first-ever incidence of stroke can be reliably quantified because it is impossible to know whether future coding of stroke data in the GP record is related to the continued care of the first stroke or are new incident instances. Extrapolated estimates from THIN assume that case ascertainment is good and GP practices will receive all health information for patients from all relevant medical sources. Because it is impossible to examine missing data from THIN, any missing data is believed to be absent at random. The THIN database has an under-representation of practices from disadvantaged communities (Public Health England, 2018).

4.6 Data description

TABLE 4.3: Description of variables used in the study

Variable	**Category**	**Coding**
Demography	Sex	Male/Female
	Year of Birth Category	1900 to 1920,1921 to 1930
		1931 to 1940,1941 to 1950,
		1951 to 1960
	Socio-Economic status	
	measured by MOSAIC(2009)	15 groups
Lifestyle	Smoking Status	Current/Ex/Non
	Alcohol Status	Current/Ex/Non
	Body Mass Index (weight in kg)/(height in m)^2	Normal/Underweight/Overweight/Obese
District	Index of Multiple Deprivation IMD Deciles	Deciles (1 to 10)
	Townsend score	Quintiles (0 to 5)
Treatments	Angiotensin-Converting-Enzyme (ACE) Inhibitor	Yes/No (Date of first prescription at stroke date)
	Alpha I Receptor Blockers	Yes/No (Date of first prescription at stroke date)
	Angotensin Receptor Blockers (ARBs)	Yes/No (Date of first prescription at stroke date)
	Calcium- Channel-Blockers (CCBs)	Yes/No (Date of first prescription at stroke date)
	Centrally Acting Anti-hypertensive	Yes/No (Date of first prescription at stroke date)
	Direct Vasodilators	Yes/No (Date of first prescription at stroke date)
	Drugs Affecting the renin angiotensin	Yes/No (Date of first prescription at stroke date)
	Loop Diuretics	Yes/No (Date of first prescription at stroke date)
	Potassium Sparing diuretics	Yes/No (Date of first prescription at stroke date)
	Thiazide Type diuretics	Yes/No (Date of first prescription at stroke date)

TABLE 4.3: Description of variables used in the study

Variable	Category	Coding
	Oral Anticoagulants	Yes/No (Date of first prescription at stroke date)
	Parenteral Anticogulant	Yes/No (Date of first prescription at stroke date)
	Statins	Yes/No (Date of first prescription at stroke date)
	Beta Blockers	Yes/No (Date of first prescription at stroke date)
Medical Conditions	Stroke diagnosis	IS/TIA/Both IS and TIA, Date of first diagnosis
	Asthma	Yes/No, Date of first diagnosis
	Atrial Fibrillation	Yes/No, Date of first diagnosis
	Coronary Heart Disease	Yes/No, Date of first diagnosis
	Chronic Obstructive Pulmonary Disease[1]	Yes/No, Date of first diagnosis
	Minor Cancer (*excluding metastatic cancer*)	Yes/No, Date of first diagnosis,
	Chronic Kidney Disease (*CKD*) (stage of CKD determined subsequently)	last *eGFR* value with date
	Cardio-Vascular Disease(*CVD*)	Yes/No, Date of first diagnosis
	Heart Failure	Yes/No , Date of first diagnosis
	Diabetes (*Type I and Type II*)	Yes/No, Last HbA1c in mmol/L reading with date
	PAD	Yes/No , Date of first diagnosis
	Diabetes	Yes/No , Date of first diagnosis
	Hypothyroidism	Yes/No, Last BP recording before stroke with date
	Hypercholesterolaemia	Yes/No, Last *LDL*, *HDL* and *triglycerides* readings

[1]CKD excludes stages 4 and 5

4.6.1 Description of variables

Demography, Socio-economic and Social deprivation factors

Based on the review of literature on stroke incidence, mortality, and morbidity, the following lifestyle and socio-economic variables were extracted. The selected demographic variables were: Sex and Year of birth. The Mosaic (2014) was selected as a measure of socioeconomic status and the Index of Multiple Deprivation, IMD (2014) and the Townsend (2017) as social deprivation.

One of the risk factors most strongly associated with stroke is gender. Stroke incidence is approximately 25% higher in men than in women (Townsend et al., 2012). The next well-established non-modifiable risk factor is age. The Year of birth was used to calculate the age of the person at stroke.

Despite the declining trends in stroke mortality during the last decades, socioeconomic inequalities in stroke have persisted and have widened during the last few years. There are high rates of mortality and incidence of stroke in low socioeconomic groups and this evidence has been consistent with previous findings (Hippisley-Cox et al., 2007).
The Mosaic (2014) is used as a measure of socioeconomic status. Mosaic is a tool that helps businesses understand the demographic and lifestyle features of their customers so they can target their products and services to the right people in the right places. The tool is an area-based classification system that allocates individuals to one of the 15 Mosaic groups or 67 types based on the nature of the people living within the same postcode area. All individuals living in these households will be classified to the same Mosaic group based on their 'average' features because the classification is done at the level of the entire UK postcode, which is similar to around 15 households (Mosaic, 2014; Sharma et al., 2010).

Mosaic (2014) UK classification is produced by Experian each year. Mosaic data were provided for each individual in THIN according to their postcode, categorised into 67 'types', and their aggregated 15 broader 'groups' based on the 2009 version.

TABLE 4.4: Description of 15 Mosaic groups based on the Experian Mosaic (2014)

Group	Description
A	Alpha Territory
B	Professional Awardfs
C	Rural Solitude
D	Small town Diveristy
E	Active Retirement
F	Suburban Mindsets
G	Careers and kids
H	New Homemakers
I	Ex- Council Community
J	Claimant Cultures
K	Upper Floor living
L	Elderly needs
M	Industrial heritage
N	Terraced Melting Pot
O	Liberal Opinions

Social Deprivation

The Index of Multiple Deprivation, IMD (2014), and Townsend scores (quantiles) were employed to assess deprivation. The official measure of relative deprivation in England is the Index of Multiple Deprivation, or (IMD, 2014). It ranks every neighbourhood in England from 1 (the poorest) to 32,844 (the wealthiest) (least deprived area). It is usual to categorise an area's relative deprivation by stating whether it is in the top 10% or 20% of deprived districts in England (IMD, 2014). The Indices of Deprivation aim to quantify a broad concept of multifaceted deprivation, which is comprised of numerous separate dimensions, or domains, of deprivation:

- Income
- Employment
- Health and Disability
- Education Skills and Training
- Barriers to Housing and Services
- Crime
- Living Environment

These combine to make an overall measure of deprivation, which is a relative ranking of the 32,844 neighbourhoods (Lower Layer Super Output Areas - LSOAs) in England. LSOAs typically have between 1000 and 3000 people living in them with an average population of 1500. In most cases,

these are relatively small areas, thus allowing the identification of pockets of deprivation (IMD, 2014).

Deprivation measures are not the same as income measures; they refer to how individuals live. Deprivation is the result of a lack of income and other resources, which can be perceived as living in poverty when taken together. The relative deprivation approach to poverty investigates deprivation indices, which are then linked to income and resources (Mack, 2020). Townsend (1979) created a list of sixty indicators of the population's "style of living" for a study on living standards in the United Kingdom conducted in 1968/69 to extend this relative deprivation method. Diet, clothing, fuel and light, home amenities, housing, and health were all included in the indicators.

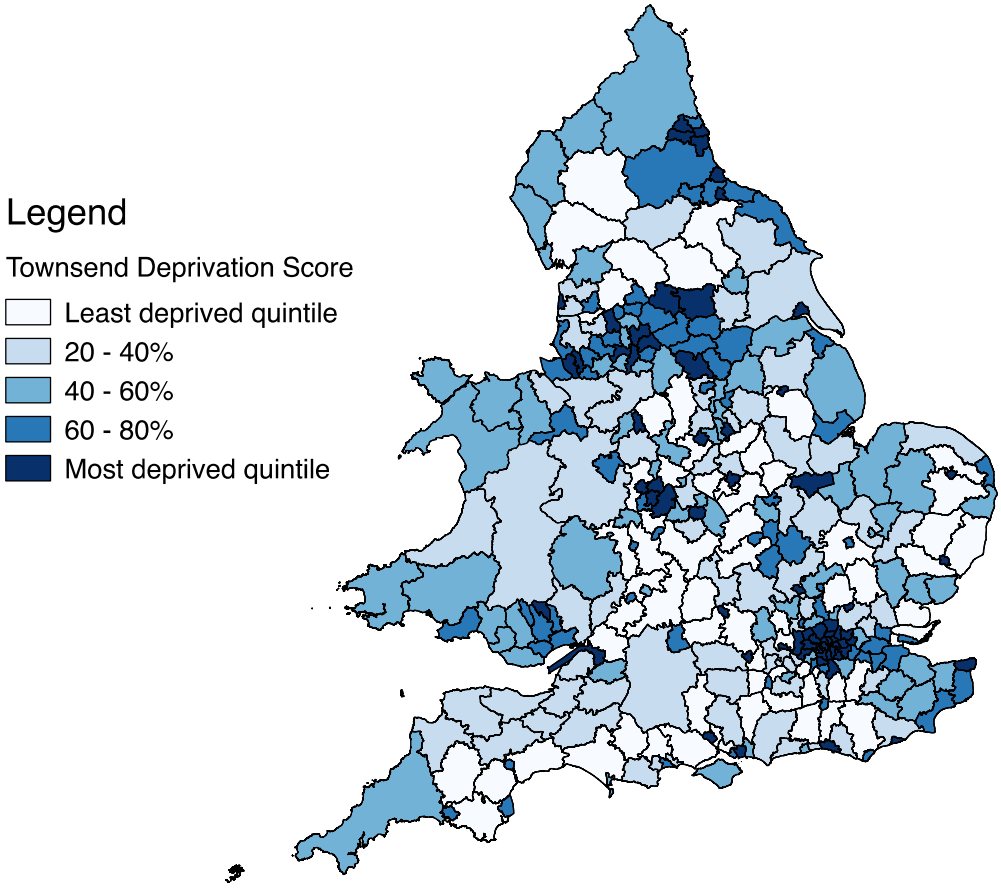


FIGURE 4.2: Townsend Deprivation Score for England and Wales from Office for National Statistics, 2011 Census.

Source : Extracted from Jones (2015)

The Townsend deprivation score can be used to show which areas are more or less deprived. The chart above, for example, demonstrates that the North East and North West have higher levels of

relative deprivation. The Townsend score in quintiles is calculated using data from THIN. The results for each variable were converted from precise scores to quintiles: five groups of equal size, numbered 1 to 5, to represent the amount of deprivation, ethnicity, and pollution in the area. The output area Townsend deprivation quintile was then matched to UK postcodes. The end result was a table with a deprivation quintile for each of the UK postcodes. This process was then performed for all ethnicity and pollution factors.

Lifestyle factors

The lifestyle factors included in the study were the smoking status, the alcohol consumption status and the Body Mass Index (BMI). The smoking and alcohol status was categorised as Ex (smoker/consumer), Current (smoker/consumer) and Non (smoker/consumer).
Body Mass Index (BMI) is a simple index of weight-for-height that is commonly used to classify underweight, normal weight, overweight and obesity in adults. It is calculated by dividing the weight in kilogrammes by the square of the height in metres (kg/m^2). The BMI is categorised by World health Organisation (2020) into Underweight (BMI$<$ 18.5) , Normal (18.5 $\leq$ BMI $\leq$ 24.9) , Overweight (25 $\leq$ BMI $\leq$ 29.9) and Obese (BMI $>$ 30).

Treatments

Drug management (both pre-stroke and long-term) used for TIA and IS stroke were selected from the NICE (2020b) recommended list and included the following classes:

1. Anti-Hypertensive drugs : Angiotensin-Converting Enzyme inhibitor (ACE), the Alpha1 receptor Blockers, the Angiotensin-Receptor Blockers (ARBs), Calcium Channel Blockers (CCBS), Centrally Acting Anti-hypertensive, Direct Vasodilators, Drugs affecting the renin-angiotensin, loop diuretics, potassium-sparing diuretics and Thiazide type diuretics.
2. Anti-coagulants drugs: Oral anticoagulants and parenteral anticoagulants.
3. Lipid-Lowering drugs: Beta-blockers and Statins.

Medical Conditions

The following medical illnesses were included in the study because they either exacerbate mortality after a stroke, or they may pose a stroke risk or are stroke side effects. They were made up of the following: Asthma, Atrial Fibrillation, Hypercholesterolemia, Coronary Heart Disease, Chronic Obstructive Pulmonary Disease (COPD), Cancer (excluding metastatic cancer), Chronic Kidney Disease (CKD) stage 1 to 3, Cardio-Vascular Disease (CVD), Heart Failure, Diabetes Type I and Type II, Hypertension. CKD stages 4 and 5 were excluded at baseline.

According to Lloyd-Jones et al. (2010), atrial fibrillation accounts for 35 to 50% of all strokes. Blood clots can occur during atrial fibrillation and can then travel on to form an embolic stroke.

A high cholesterol concentration in the blood is known as hypercholesterolaemia. It might be present from birth due to an inherited genetic abnormality, and it can lead to atherosclerosis and coronary heart disease (CHD) at a young age (NICE, 2020a). If an adult's total cholesterol concentration is greater than 7.5 mmol/L, or if there is a family history of hypercholesterolaemia, hypercholesterolaemia is suspected.

Hypertension is defined as a clinic blood pressure of 140/90 mm Hg or higher, with a subsequent ambulatory blood pressure monitoring (ABPM) daytime average or home blood pressure monitoring (HBPM) average blood pressure of 135/85 mm Hg or higher, according to NICE guidelines (NICE, 2020a).

Medical Tests and their use for diagnoses

The following medical recordings (markers) were included in the study and used for diagnoses:

1. Hypertension: clinic blood pressure is 140/90 mm-Hg or higher.
2. Hypercholesterolaemia: total cholesterol concentration is greater than 7.5 mmol/L or LDL-C $>$ 190 mg/dL (4.9 mmol/L) in adults..
3. Chronic Kidney Disease (CKD) [2] : The *estimated- glomerular filtration rate* (eGFR) result provides staging of CKD from one of the five:
    - stage 1 (G1) - a normal eGFR (above 90 ml/min), but other tests have detected signs of kidney damage
    - stage 2 (G2) - a slightly reduced eGFR (60-89 ml/min), with other signs of kidney damage
    - stage 3a (G3a) - an eGFR of 45-59 ml/min
    - stage 3b (G3b) - an eGFR of 30-44 ml/min
    - stage 4 (G4) - an eGFR of 15-29 ml/min
    - stage 5 (G5) - an eGFR below 15 ml/min, meaning the kidneys have lost almost all of their function.

    Alternatively , the *albumin:creatinine ratio* (ACR) result provides the staging from one to three :
    - A1 - an ACR of less than 3 mg/mmol
    - A2 - an ACR of 3-30 mg/mmol

[2] CKD stages 4 and 5 are excluded from the current study)

- A3 - an ACR of more than 30 mg/mmol

For both eGFR and ACR, a higher stage indicates more severe kidney disease.

4. Diabetes: HbA1c measurement of 48 mmol/mol or higher

4.6.2 Asthma

According to the Asthma and Foundation (2019), asthma affects approximately 12 percent of the population (8 million individuals). The Health Improvement Network (THIN) database records for 2004-13 were used in the aforementioned study. Approximately, 5.4 million people (8.2%) receive therapy for the disease (Asthma & Foundation, 2019). Their findings also revealed that asthma may be over-diagnosed. According to them, asthma affected 12% of the population in 2012. From the years 2004 to 2012, ladies were somewhat more likely than males to get asthma. According to recent statistics, young people (ages 21 to 30) and young adults (ages 16 to 20) are the age groups most likely to be diagnosed with asthma.

Our crude rate (13.8%) slightly matches the latter. It also concurs with the fact asthma is diagnosed more in females than males.

4.6.3 Atrial Fibrillation

According to Davis et al. (2012), atrial fibrillation was shown to be slightly more common in men (2.4 percent) than in women (1.6 percent). Using prospectively selected groups, the research team came up with the following conclusion: 3960 people aged 45 and above were chosen at random from the UK population. With increasing age, the risk of atrial fibrillation increased. Furthermore, utilising the Renfrew-Paisley cohort, National Collaborating Centre for Chronic Conditions et al. (2006) discovered that of an original cohort of men and women aged 45 – 64 years ($N = 15,406$), there were 100 (0.65 percent; 95 percent CI 0.53 to 0.79 percent) reported occurrences of AF. Men were found to have more incidences (53 out of 7,052) than women (47 of 8,354).

Our results (6.9%) diverge from the two documented percentages because it is assumed that more stroke patients visit practices after a stroke event and it is a known fact that AF is a clinical risk factor for stroke thereby our crude rate being higher than published research.

4.6.4 Hypercholesterolemia (Dyslipidemia)

Over half of all persons in England have high cholesterol (>5 mmol/L), according to Heart UK (2020). The World Health Organisation's key findings demonstrate that the WHO Region of Europe has the highest prevalence of elevated total cholesterol (54 % for both sexes). Our results differ significantly from the recorded population estimates, with a difference of about 7.5%.

4.6.5 Chronic obstructive pulmonary disease (COPD)

An estimated 1.2 million people in the United Kingdom have COPD, according to the British Lung Foundation (2020). COPD affects approximately 2% of the population, with 4.5% of those over the age of 40 suffering from the disease. Between 2008 and 2012, there was a 9% increase in prevalence. During the period $2004-13$, Scotland and the northern areas of England experienced a higher proportion of persons diagnosed with COPD for the first time than the rest of the UK. Males suffer more than females. According to the British Lung Foundation (2020), the majority of those diagnosed with COPD are above the age of 40. Our results reveal that an estimate of 6.7% of the THIN patients was diagnosed with COPD. This high percentage is accounted for by the fact that our database consists of more older patients who are more likely to be diagnosed with lung-related problems due to old age. It can be further justified by the fact that a large proportion of them were smokers.

4.6.6 Hypertension

In England, 25% of adults are diagnosed with hypertension, according to NICE (2022). Unprecedented numbers of people go misdiagnosed; up to 7 million people in the UK are unaware of their risk, according to (British Heart, 2019). Because Due to the sample being made up of more elderly patients, and hypertension being predominantly driven by age, our estimate of 45.9% is significantly higher than national estimates.

4.6.7 Diabetes

According to Diabetes (2018), about 3.7 million people in the United Kingdom have been diagnosed with diabetes. This equates to roughly 7% of the population. Men have a higher prevalence of diabetes than women, with 9.6% versus 7.6% in England, according to Public Health England (2016). The increased incidence is linked to the higher average age of the population of interest and the fact that a significant percentage was socially deprived, according to our estimates of 27.8%. Males tend to be diagnosed more than females.

4.6.8 Heart Failure

Around 2.3 million people in the United Kingdom have coronary heart disease, with 500,000 of them suffering from heart failure. In 2013, the estimated frequency in men of all ages in the United Kingdom was 1.22 percent. For the UK (Nichols et al., 2014), the prevalence was 0.76 percent in women. There was a definite link between getting older and having a higher risk of heart failure. Our crude estimate from the THIN extracted data was 10.2% and again the high number of elderly patients in the file can be accountable for this high incidence. Our results also showed that females were diagnosed more than males, which differs from the recent estimates of heart failure in the UK.

This could be a result of female patients visiting the practice more regularly than their counterparts and hence being diagnosed more often with the given medical condition.

4.6.9 Peripheral artery disease (PAD)

A cohort study in THIN of individuals aged 50 to 89 years identified annually between 2000 and 2014 produced the most latest estimates of PAD incidence in the UK to be published. It was also the largest study of temporal patterns to date. In 2014, the incidence of symptomatic PAD per 10,000 person-years was predicted to be 17.3 (men: 23.1; women: 12.4). The incidence and prevalence of PAD are decreasing in the United Kingdom, according to the findings (Cea-Soriano et al., 2018).
Our cohort's estimated percentage was 9.7%. The discrepancy in the crude rate is due to a disparity in the population's age distribution.

4.6.10 Hypothyroidism

According to NICE UK (2018), spontaneous hypothyroidism affects 1–2% of the population. It is up to ten times more common in women than in men (Vanderpump, 2011).
Our findings reveal that female cases outnumber male cases, which is consistent with national figures. The overall crude incidence rate in our THIN database is 8.2 percent of all individuals (1.5 percent for males; 6.6 percent for females). The values are greater due to the study cohort's advanced age.

4.6.11 Chronic Kidney Disease, CKD

According to Public Health England (2019), 2.6 million people aged 16 and over are estimated to have CKD stage 3 or higher (diagnosed and undiagnosed). This equates to 6.1 percent of the population in this age group (95 % CI: 5.3-7.0 %). Women have a higher frequency of CKD stage 3-5 than males, with 7.4 percent versus 4.7 percent. With 1.9 percent of persons aged 64 and under having CKD stage 3-5, 13.5 percent of people aged 65-74 having CKD stage 3-5, and 32.7 percent of people aged 75 and beyond having CKD stage 3-5, there is a definite link between rising age and higher CKD prevalence. With increasing age, the disparity in CKD stage 3-5 prevalence between males and females widens. Our overall crude rate is 9.4%, which is higher than previously published estimates.

4.6.12 Cardio-Vascular Diseases (CVD)

According to Townsend et al. (2012), male adults aged 16 and above had a prevalence of 11.1 percent in 2011 while female adults had a prevalence of 9.1 percent. Our crude rate is 16.4%. Male cases (9.6% of all cases) outnumber female cases (6.7 percent). The gender distribution is in accordance with the national statistics.

4.6.13 Minor Cancer

Annually, more than 360,000 new cancer cases are diagnosed in the United Kingdom (2013-2015). In 2015, over 183,000 new cancer cases were diagnosed in males in the United Kingdom, whereas around 177,000 new cancer cases were diagnosed in females (Roth et al., 2020). Over a third (36%) of all cancer cases in the UK are detected in adults aged 75 and above each year (2013-2015). In the United Kingdom, the rates of all cancers combined are highest in those aged 85 to 89. (2013-2015). Breast, prostate, lung, and colon cancer account for more than half of all new cancer cases, according to UK Cancer Research UK (2020) in 2015. Our crude incidence rate for small cancer is 16.2 %, with male cases outnumbering female cases.

4.7 Exploratory data analysis

The time trends of incidence of some selected medical conditions by gender were studied using the original dataset. The selected medical diagnoses of atrial fibrillation, hypertension, diabetes Type II, myocardial infarction, and hypercholesterolaemia experienced a rise from 1980 to 2016 (See the left panel Figure 4.3 and Figure 4.4). Similarly, the right panels of the figures show the trends in the proportion of prescriptions for treatments of these cardiovascular diseases such as anticoagulants, anti-hypertensives, anti-diabetic, antiplatelets, and statins by calendar year. No significant gender differences were detected. While the selected prescriptions were on rising over time, the unusual decline in the prescriptions of antiplatelets is observed post after 2011 in Figure 4.5a.

Figure 4.5b describes the trends in mortality within a year after first ever ischaemic stroke. In IS patients, the one year all-cause mortality increased between 1996 and 2001 (annual absolute change + 1.5%) then decreased from 2001 to 2008 (- 0.3%) and continued to decrease further for the next decade 2008 to 2015 (- 0.5%) . A similar trend is observed for the 1 month all-cause mortality with an absolute change of + 0.5%, - 0.4% and - 0.3% for the time periods $1996-2001$, $2001-2008$ and $2008-2016$, respectively. These results indicate that stroke outcomes have improved over time. A considerable decline in mortality is observed post calendar year 2008. In 2008, the NICE guidelines were amended to include mandatory administration of alteplase, surgery if required, acute care, neuro-imaging, better management of post-stroke complications and secondary prevention and rehabilitation (NICE, 2019). The observed decline may be due to these changes in the 2008 NICE guidelines.

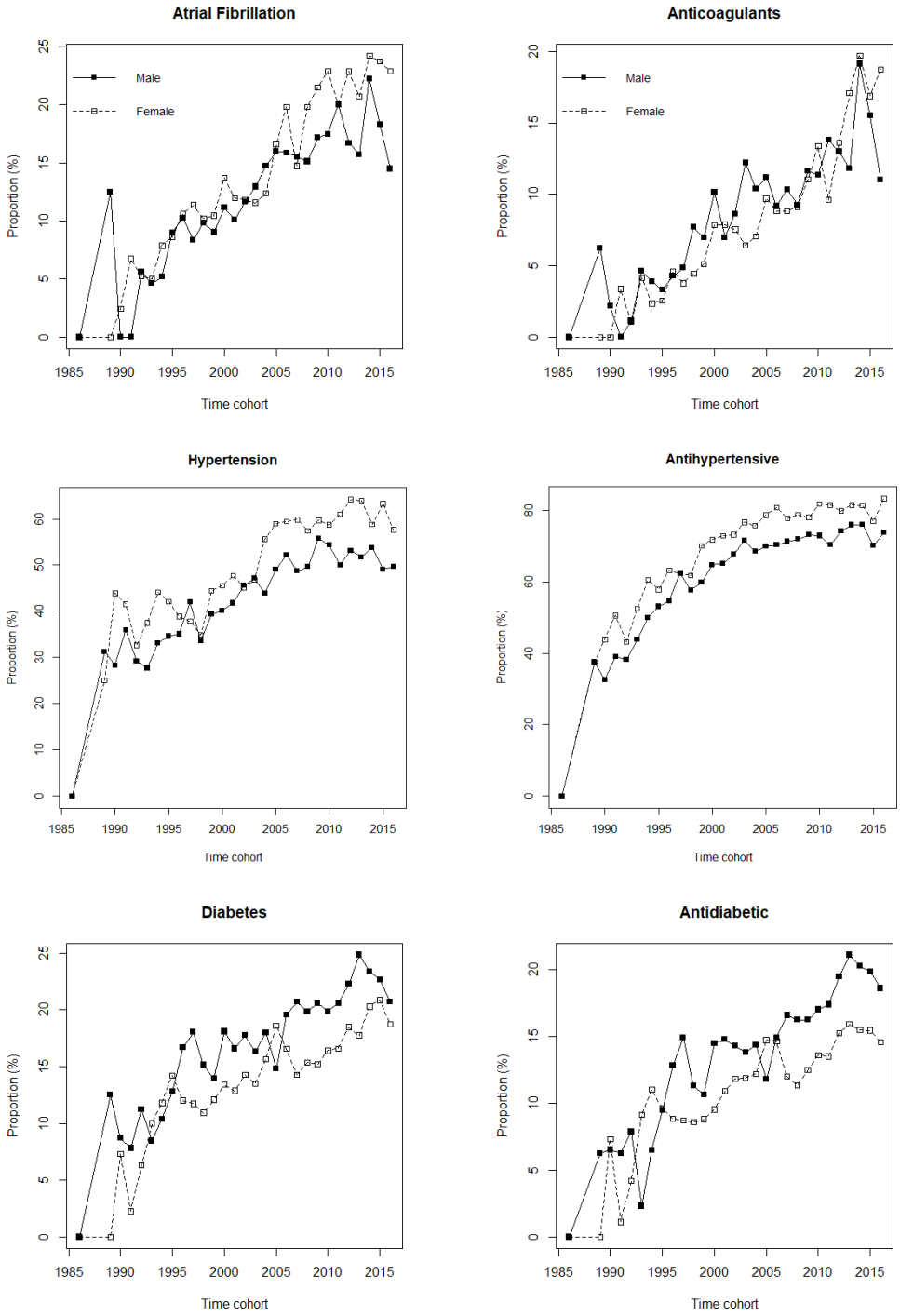


FIGURE 4.3: Incidence of prior risk factors and medication use over time in patients with first-ever IS by gender.

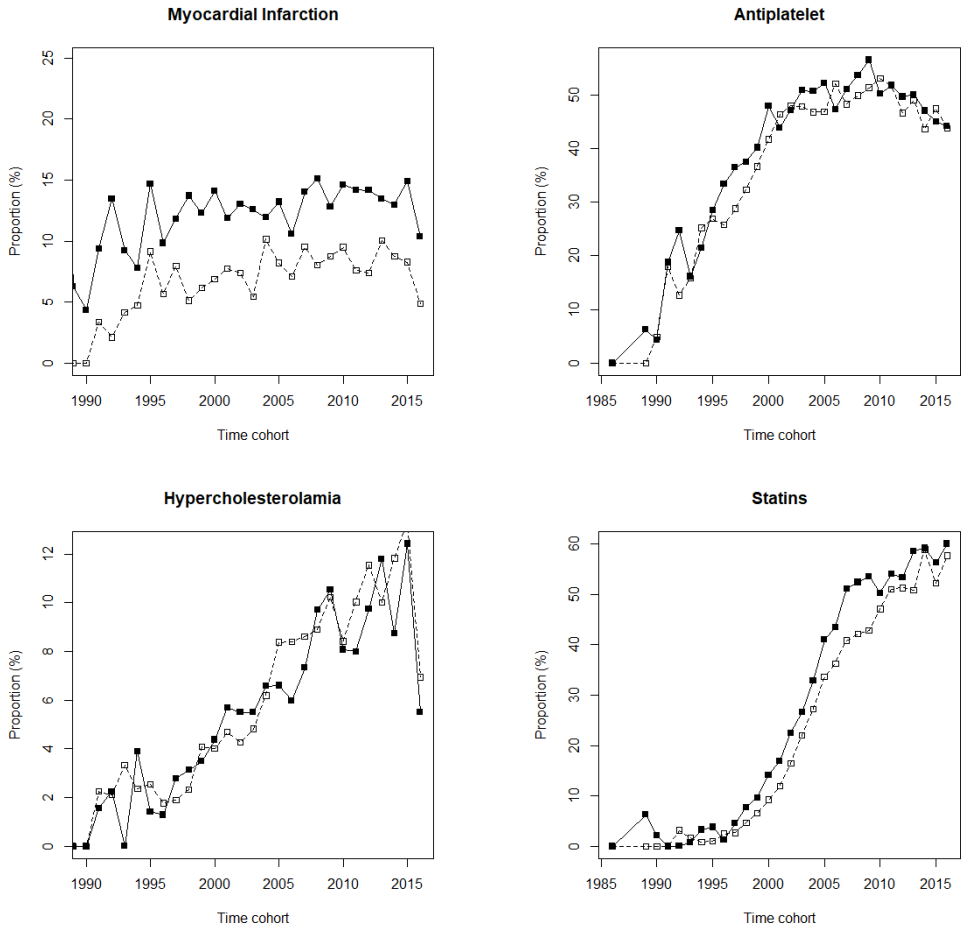
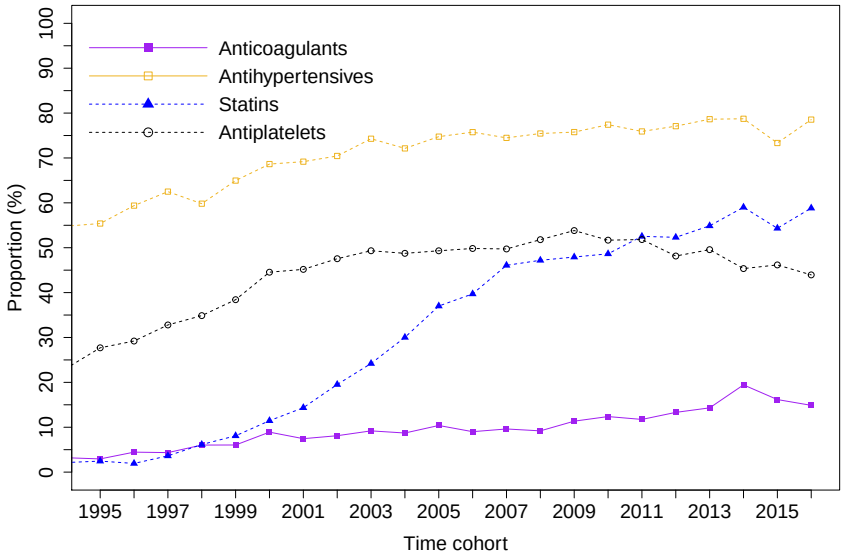
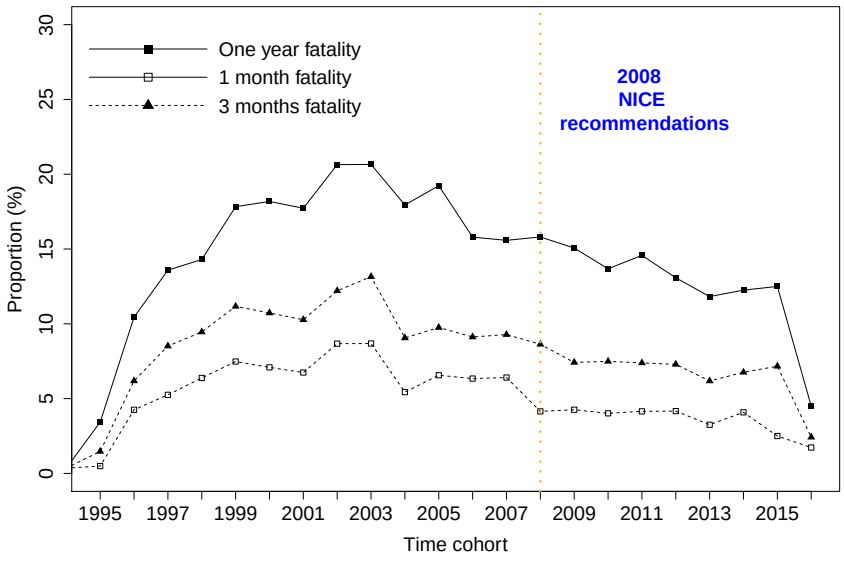


FIGURE 4.4: Incidence of prior risk factors and medication use over time in patients with first-ever IS by gender.



(A)



(B)

FIGURE 4.5: (a) Proportion of pre-morbid prescription of selected drugs to IS patients by calendar year. (b) 1 month, 3 months and one-year all-cause mortality for first-ever Ischaemic stroke patients 1990-2016.

4.8 Chapter Summary

The THIN database is presented as the source of data in this chapter. The inclusion-exclusion criteria and the selection process are explained. The prevalence of co-morbidities described is compared with published studies.

5 Survival analysis after Transient Ischaemic attack

This chapter presents the development of the survival model for the analysis of the survival of transient-ischaemic attack patients. Firstly, a description of the model development is provided. Secondly, the fitted full-case model is described and assessed. Thirdly, multiple-imputation procedures are explained and a comparison of the models of the full-case and multiple imputed data is provided. The findings of the final survival model are then presented using Forest plots, hazard curves and hazard ratio plots. The important findings are then discussed.

5.1 Study Design

5.1.1 Selection criteria

The validity and structure of the THIN database has been described earlier in chapter 3. A retrospective case-control design was used to investigate how the diagnosis of TIA affects the survival of patients. 129,668 records of patients were extracted in total which consisted of 33,909 patients with first-ever TIA diagnosis between the years 1986 to 2017 matched to 95,759 controls in Figure 5.1.

TABLE 5.1: The distribution of variables with missing values in the TIA dataset

Variables	Levels	Number (%)[1]
Smoking status	Non-smoker	32140(41%)
	Ex-Smoker	19316(24%)
	Current Smoker	8403(11%)
	Missing	19408(24%)
BMI category	Underweight	1362(2%)
	Normal	17678(22%)
	Overweight	19680(25%)
	Obese	10760(14%)
	Missing	29787(38%)
Alcohol consumption	Abstainer	8968(11%)
	Ex-consumer	1616(2%)
	Current consumer	31144(39%)
	Missing	37539(47%)

[1]The proportions are determined out of a total of 79,267 records.

One of the entry requirements was a valid IMD deprivation value, i.e, only residents of England. Table 5.1 displays the proportions of missing data for BMI category, smoking status and alcohol consumption. BMI includes 37% of missing data, smoking status 24% and alcohol consumption status 46%. The flowchart (Figure 5.1) illustrates the data selection process to attain the full case dataset without missing values which included full records of 9,377 TIA survivors and 27,535 controls. To ensure a balanced dataset, the 9,377 TIA patients were matched again to their controls. Only 15,420 out of 27,535 controls were retained (not all 27,535 controls with full records could be matched with 9377 TIA survivors' full records). All 9377 TIA survivors were matched to at least one control. The final full cases dataset hence consisted of 9377 TIA patients and 15,420 matched controls.

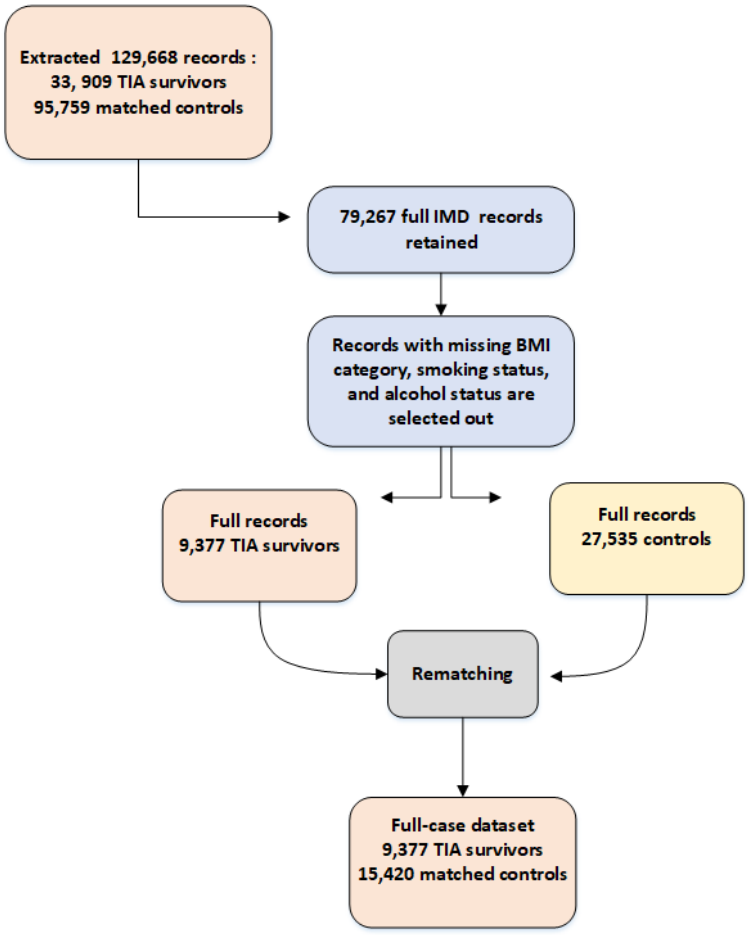


FIGURE 5.1: Data extraction in the TIA dataset

Statistical analysis was then performed in two stages. First, the full-case analysis was performed to get a parsimonious model. The model was then used to inform the imputation regression model where the same predictors were used. In the second stage, new analyses were performed on the 10 imputed datasets and the estimates were then pooled using Rubin's rules.

5.2 Description of the full-case dataset

TABLE 5.2: Baseline characteristics of TIA dataset

Variable	Type of patients	
	Cases ($n = 9,377$)	**Controls** ($n = 15,420$)
Demographical variables		
Age (mean (std dev.) in years)	69.4(10.0)	67.8(9.4)
39-60	1846(19.7%)	3453(22.4%)
61-70	3013(32.1%)	5572(36.1%)
71-76	2049(21.9%)	3448(22.4%)
77+	2469(26.3%)	2947(19.1%)
Male	4723 (50.4%)	7931 (51.4%)
Year of TIA diagnosis		
1986-1992	554(5.9%)	1027(6.6%)
1993-1999	2328(24.8%)	3896(25.3%)
2000-2006	4080(43.5%)	6794(44.0%)
2007-2016	2415(25.8%)	3703(24.0%)
IMD Quintile		
1	1333(14.2%)	2061(13.4%)
2	1747(18.6%)	2807(18.2%)
3	1935(20.6%)	3205(20.8%)
4	2179(23.2%)	3712(24.1%)
5	2183(23.2%)	3635(23.6%)
Pre-morbid conditions		
Asthma	981(10.5%)	1447(9.6%)
Atrial Fibrillation	754(8.0%)	597(3.9%)
Diabetes II	1323(14.1%)	1534(9.9%)
CHD	2248(24.0%)	2249(14.5%)
CKD stages 1-2	406(4.3%)	591(3.8%)
COPD	462(4.5%)	551(3.5%)
Heart Failure	468(5.0%)	422(2.7%)
Hypertension	4317(46.0%)	5396(35.0%)
Hypercholesterolaemia	767(8.2%)	981(6.4%)
Hypothyroidism	700(7.5%)	808(5.2%)
Myocardial Infarction	850(9.1%)	903(5.9%)
PVD	2194(23.4%)	2758(17.8%)
Lifestyle factors		
BMI category		

Table 5.2 *Continued from previous page*

Variable	Type of patients	
	Cases ($n = 9,377$)	**Controls** ($n = 15,420$)
Normal	3179(33.9%)	5279(34.2%)
Underweight	196(2.1%)	285(2.3%)
Overweight	3825(40.8%)	6430(42.1%)
Obese	2177(23.2%)	3426(22.3%)
Smoking status		
Non smoker	4661(549.7%)	8359(54.2%)
Ex-smoker	3297(35.1%)	4954(32.1%)
Current smoker	1419(15.1%)	2107(13.7%)
Alcohol status		
Abstainer	2112(22.5%)	3032(19.7%)
Ex-Consumer	435(4.6%)	496(3.2%)
Current consumer	6830(72.8%)	11892(77.1%)
Premorbid prescriptions		
Anticoagulant agents	685(7.3%)	647(4.2%)
Antiplatelet agents	4877(52.0%)	3314(21.5%)
Lipid lowering agents	2995(31.9%)	3265(21.17%)
Antidiabetic agents	608(6.5%)	1019(6.6%)
Antihypertensive agents		
ACE inhibitors	1815(19.3%)	2061(13.4%)
Alpha1 receptor blockers	370(3.9%)	401(2.6%)
ARBS	616(6.6%)	749(4.9%)
Beta blockers	1974(21.0%)	2231(14.4%)
CCBs	1738(18.5%)	2005(13.0%)
Centrally acting antihypertensive drugs	54(0.58%)	39(0.25%)
Direct Vasodilators	10(0.11%)	13(0.08%)
Drugs affecting the renin angiotensin system	4(0.04%)	6(0.04%)
Loop diuretics	700(7.5%)	666(4.3%)
Potassium sparing diuretics	385(4.1%)	367(2.4%)
Thiazide type diuretics	1311(14.0%)	1730(11.2%)

The general characteristics of the study population at baseline are presented in Table 5.2. The stroke survivors were generally somewhat older than their matched controls with a mean age of 69.4 years compared to 67.8 years. The distribution of gender and social deprivation were roughly the same across cases and controls. Based on the medical conditions prior to TIA events, hypertension was the leading ailment (46% in cases and 35% in controls), followed by peripheral-arterial disease

occurring in 23% of cases and 18% of controls. Coronary heart disease and Diabetes type II were the next two most common co-morbidities. In terms of lifestyle factors, the majority of both cases and controls were overweight, non-smokers, and current alcohol consumers. In terms of medical therapies, cases were generally more medicated than controls.

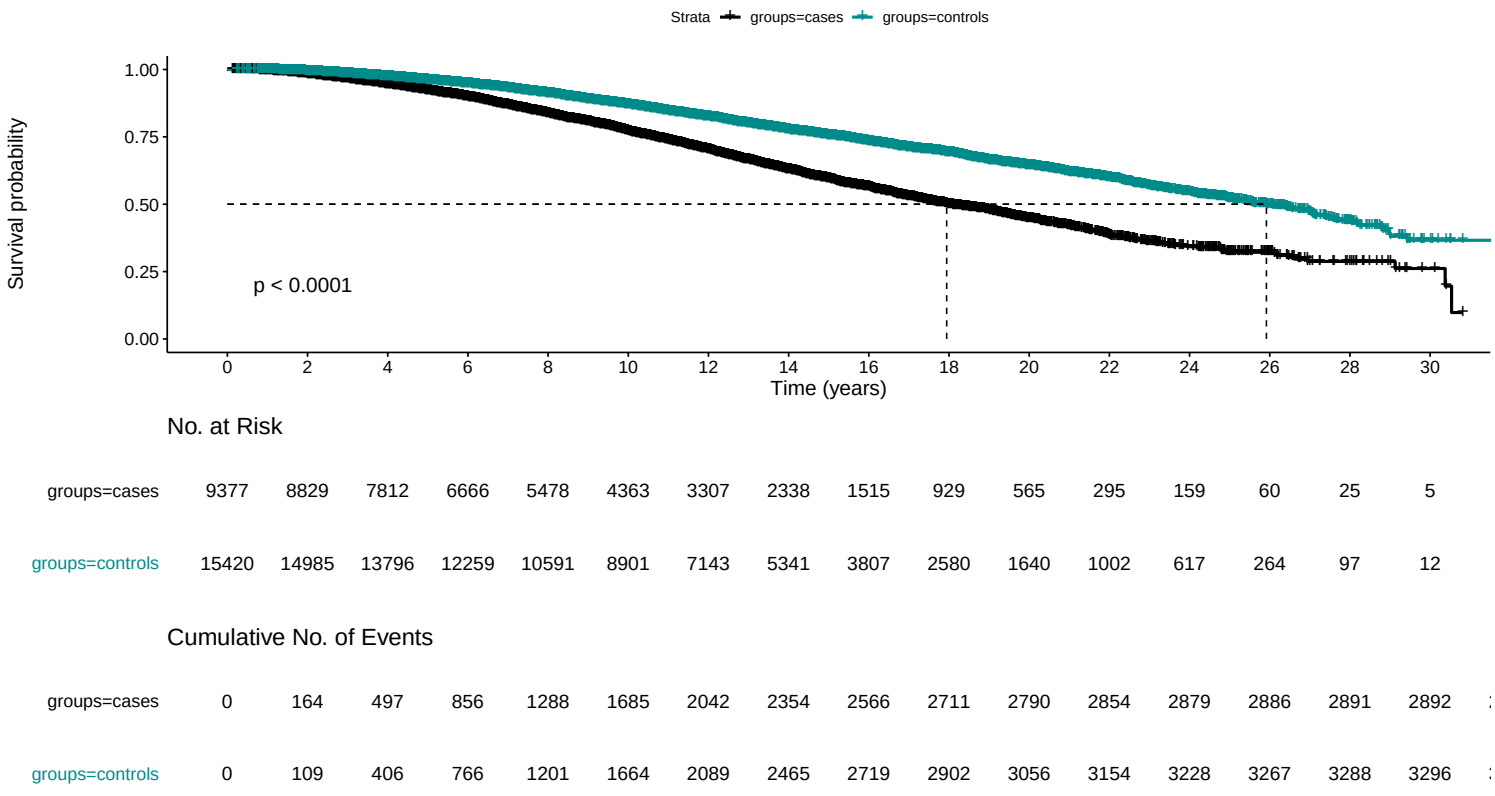


FIGURE 5.2: Kaplan-Meier plot after first-ever TIA event.
Black line indicates the cases; dark cyan,controls.

The median survival for TIA patients was 18 years whereas that of the controls was 26 years (Figure 5.2). Patients surviving a TIA have a shortened life expectancy compared to the matched controls. After a decade, nearly 18 % of TIA survivors passed away (1685 out of 9377) compared to 10% controls. After another decade, 29 % of cases died compared to 20 % controls. There is a significant survival difference between cases and controls (the p-value for the difference (log-rank) test, was $p < 0.001$). The survival disparity is also apparent in different age groups at TIA events (Figure 5.3).

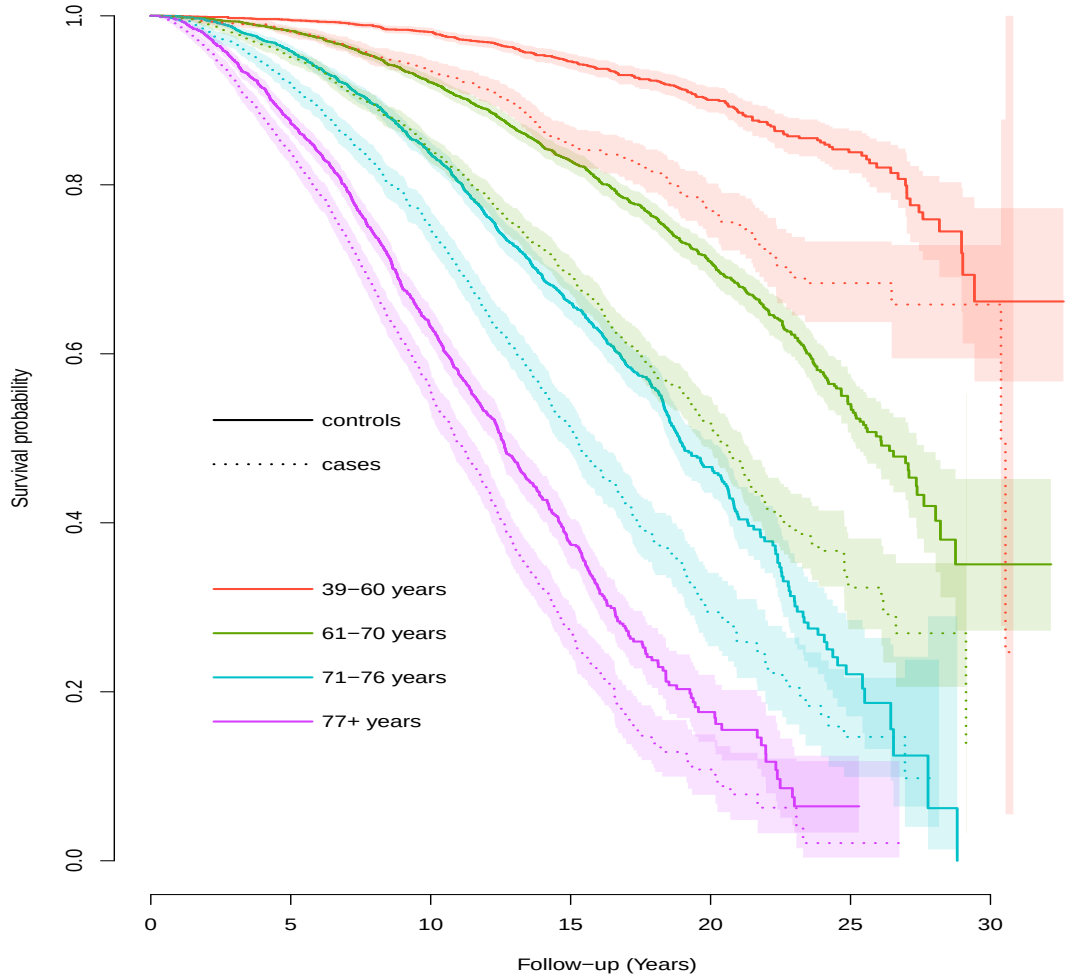


FIGURE 5.3: Unadjusted Kaplan-Meier plot by age groups at TIA event and case-control status.
Solid lines indicate the controls; dashed lines for cases.

5.2.1 Univariate analysis

TABLE 5.3: Univariate analysis of predictors for the TIA cohort.

Study variable	Levels	Coefficient (Beta)	Exp(Beta)	SE	Z statistics	P Value
Birth cohort	1900-1960					
	1921-1930	-0.68	0.51	0.03	-21.58	0.00
	1931-1940	-1.5	0.22	0.04	-40.78	0.00
	1941-1960	-2.51	0.08	0.06	-40.61	0.00
Gender	Female					
	male	-0.23	0.8	0.03	-8.88	0.00
IMD category	1					
	2	-0.13	0.87	0.04	-3.14	0.00
	3	-0.13	0.88	0.04	-3.07	0.00
	4	-0.27	0.76	0.04	-6.64	0.00
	5	-0.36	0.7	0.04	-8.36	0.00
Age category	39-60					
	61-70	1.1	3	0.05	21.01	0.00
	71-76	1.8	6.06	0.05	34.04	0.00
	77+	2.63	13.85	0.05	50.01	0.00
Groups	Controls					
	Cases	0.61	1.83	0.03	23.7	0.00
BMI category	Normal					
	Obese+Overweight	-0.32	0.72	0.03	-12.57	0.00
Asthma	Present	0.32	1.38	0.04	7.8	0.00
COPD	Present	1.16	3.19	0.05	22.26	0.00
CKD	Present	0.87	2.4	0.08	10.38	0.00
Myocardial Infarction	Present	0.82	2.26	0.04	20.3	0.00
PAD	Present	0.5	1.65	0.03	16.34	0.00
Smoking	Non-smoker					
	Current Smoker	0.33	1.39	0.04	9.31	0.00
	Ex-smoker	0.37	1.44	0.03	12.91	0.00
Alcohol	Abstainer					
	Consumer	-0.25	0.78	0.03	-8.65	0.00
Atrial Fibrillation		1.04	2.84	0.05	22.68	0.00
Hypertension	Yes	0.73	2.08	0.03	27.42	0.00
Antiplatelet	None					
	Aspirin only	0.66	1.93	0.03	23.66	0.00
	DUAL Therapy	0.81	2.24	0.06	13.28	0.00
	Other APL agents	0.69	1.98	0.07	9.89	0.00
Diabetes	No					
	Yes and Treated	0.8	2.23	0.04	19.66	0.00
	Yes and Untreated	0.62	1.86	0.07	8.39	0.00
Ancoagulant agent	Yes	0.96	2.6	0.05	20.01	0.00

Table 5.3 shows the model-based estimates univariate analysis of the important predictors. All the predictors were significant ($p < 0.00$). The selected prognostic factors were then used for the multiple Cox regression.

5.3 Model development and diagnostic tests

One of the assumptions underlying a good model is uninformative censoring, i.e Censoring must be independent of survival. Informative censoring can lead to biased results in survival analysis, the true survival probability at any time t can be overestimated. Our data was non-informative. Formally, non-informative censoring is defined as the "Probability of being censored at time t does not depend on prognosis for failure at time t ". In the above context, non-informative censoring implies that the probability of being censored (transferred out) for any subject in the risk set at time t does not depend on that subject's prognosis for failure at time t. Any distinctive pattern with regards to the patients who were transferred out given their social deprivation, TIA diagnosis and age was checked. Figure 5.4 shows no distinct transfer pattern for cases < 70 years. It is noteworthy though that at older ages (> 70 years), more cases were lost to follow-up. This may be due to moving to nursing homes and/or old-age accommodation. No trends were observed for the major part of the data and hence uninformative censoring was assumed. The survival model also assumed no time-tied deaths. Time-tied deaths were handled by Efron's approximation.

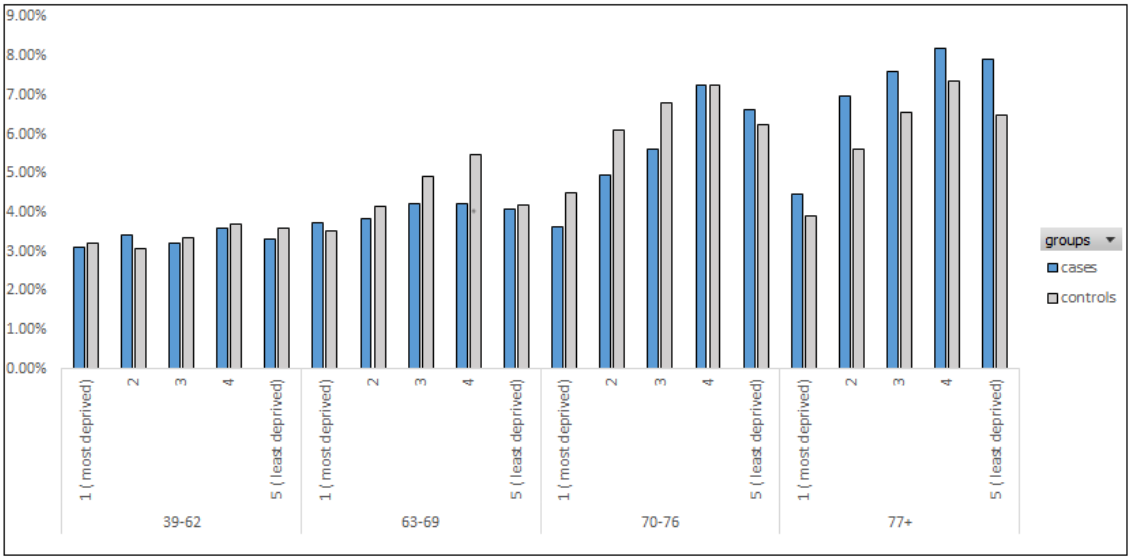


FIGURE 5.4: Percentage of cases and controls lost due to follow-up by social deprivation, age category at entry and TIA diagnosis.

A Cox proportional hazards regression model was fitted to estimate the effect of the first-ever TIA and other factors on the hazard of all-cause mortality. The variables used to build the model are presented in Table B.3. The initial model included all the main effects and two-way interactions of all variables. A Gamma frailty (random effect of general practice) was also included. Starting with a full model, the model was gradually reduced using a backward-elimination process. The covariates and factors were removed based on the removal significance level of 0.05 for main effects and 0.01 for interactions in the ANOVA.

Our penultimate model consisted of the following covariates: random frailty effect of GP practice, birth cohort, age category, sex, case/controls status, IMD Quintile, BMI category, antiplatelet drugs, asthma, CKD, COPD, heart failure, myocardial infarction, PVD, smoking status, alcohol consumption, hypertension, diabetes, atrial fibrillation as main effects and significant interaction terms between age category, antiplatelet drugs and hypertension with cases/controls status. However, the proportionality assumption was not satisfied for several covariates; the hazard ratios were not constant over time, which is an underlying premise for use of the Cox proportional hazards (CPH) model (See Table B.4).

Grambsch and Therneau (1994)'s cox.zph test was performed to evaluate the adequacy of the proportional hazards assumptions. The global test was highly significant (global p-value = 4.5e-15), providing evidence of non-proportionality. The scaled Schoenfeld residuals are very useful for investigating the dependence of an effect on time. Violation to proportional hazards assumption for birth cohort, age category, case/controls status, and antiplatelet drugs among others were detected. Figures 5.5 show the plots of Schoenfeld residuals of the violating variables versus time for some selected factors. Systematic departures from a horizontal line and clear downward or upward trends were indicative of non-proportional hazards.

Next, Martinussen and Scheike (2007)'s method for extended multiplicative hazards models was used, the *timereg* package in R. The following variables were found to have time-varying effects based on the Kolmogorov-Smirnov and the Cramer von Mises tests: birth cohort, age categories, antiplatelets and heart failure (See Tables B.7 and B.8). The shape effects were then added for the covariates which were violating the PH assumptions in the Weibull double-Cox model. The Weibull Double-Cox model is explained next.

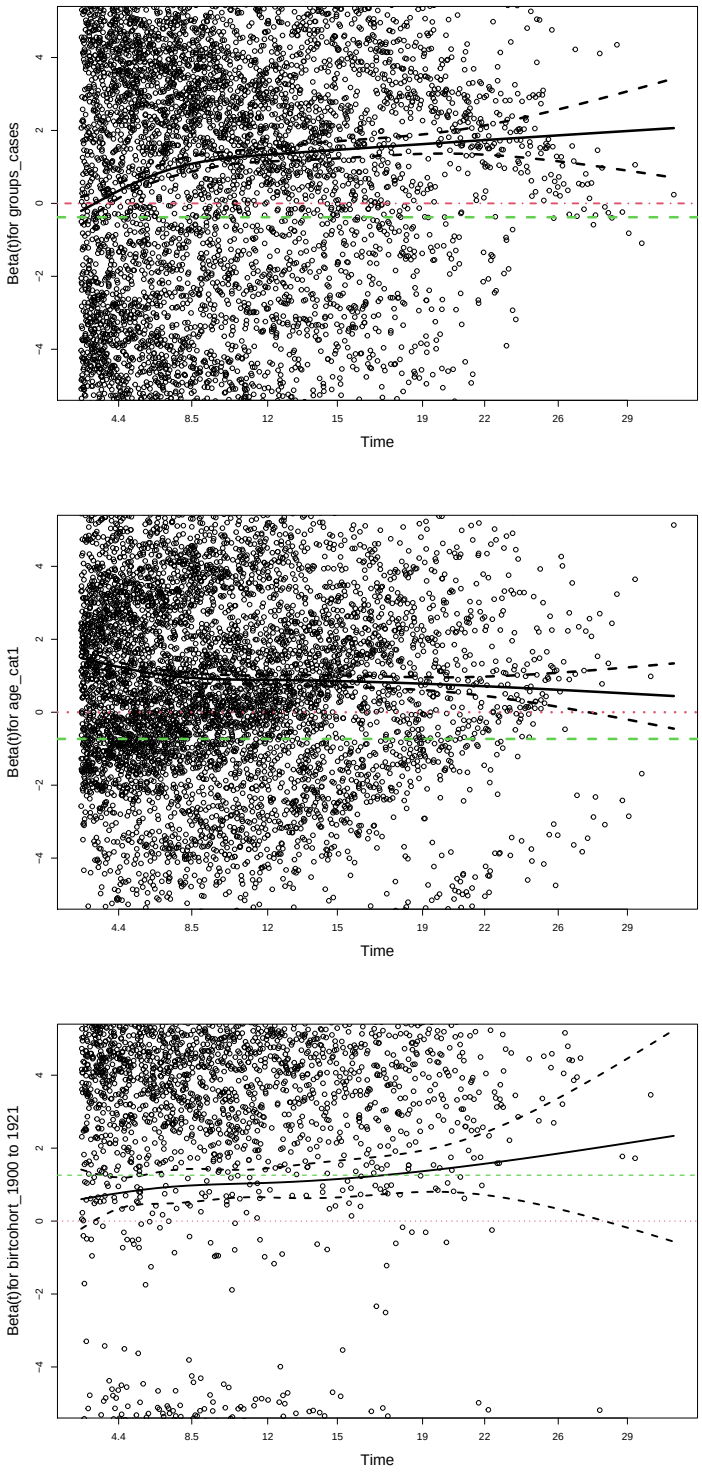


FIGURE 5.5: Schoenfeld residuals against the transformed time.

for variables of cases,age category($39-60$ years), birth cohort($1900-1921$) violating the PH assumption

5.3.1 Weibull Double-Cox model.

A number of approaches to tackle the non-proportionality issue have been described in Chapter 3. A Weibull double-Cox model was constructed which allowed non-proportionality of hazards by modelling both shape and scale parameters, adapted from Begun et al. (2019). To allow covariates to be time-varying, the chosen methodology includes also separate regressions for scale and shape parameters to circumvent the problem of non-proportionality. First, five types of parametric baseline survival distributions were fitted- Weibull, exponential, logistic, log-logistic, Gompertz and log-normal to a subgroups of data (cases and controls) with baseline parameters. Kindly note that the results were based on a model estimated with all individuals but only the baseline hazard form given. Figure 5.6 shows the plots of the cummulative hazards of these parametric distributions compared to baseline hazards estimated by the Kaplan-Meier method. The Weibull distribution fitted the baseline hazard reasonably well whereas the rest of the parametric distributions fail to match the convexity of the Kaplan-Meier hazard function. The goodness of fit of the parametric distributions was evaluated using the Akaike Information Criterion (AIC). The relative values of AIC are presented in the plots.

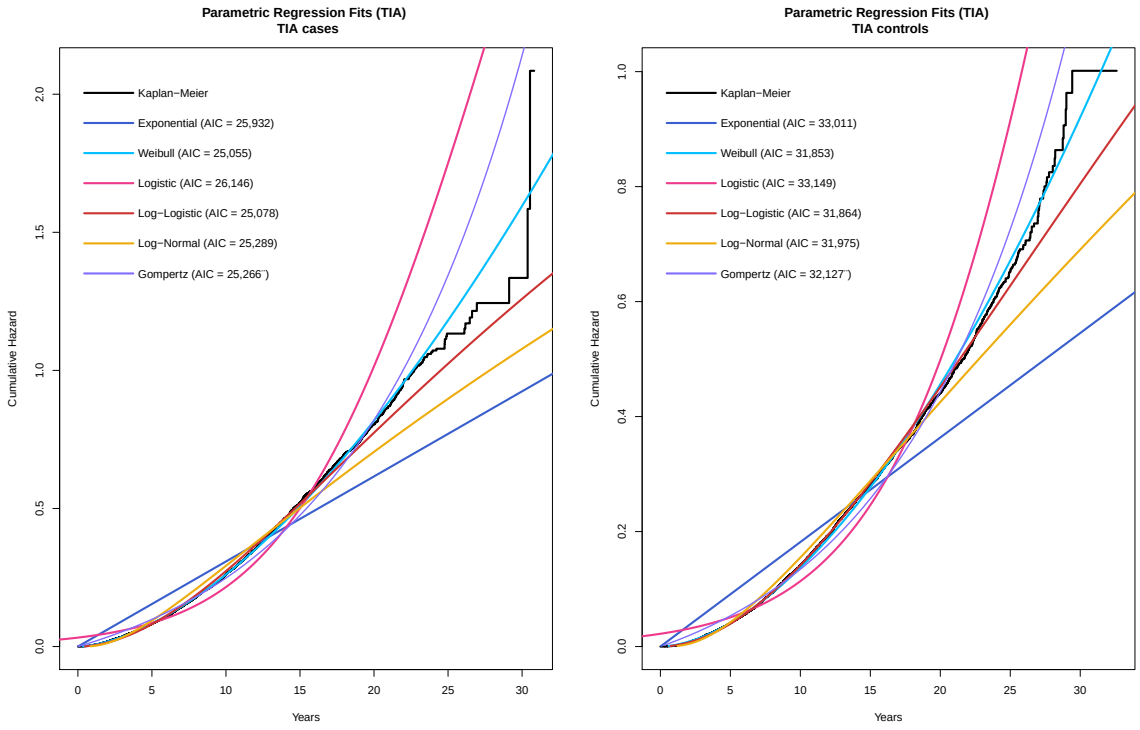


FIGURE 5.6: Graphical comparison of the fit of several distributions:

Exponential, Weibull, Logistic,log-logistic, Log-Normal and Gompertz cummulative hazards compared to KM cummulative hazards function.

The Weibull model had the lowest AIC (AIC =25,055 for cases and AIC= 31,853 for controls), indicating a better overall fit than the other models.

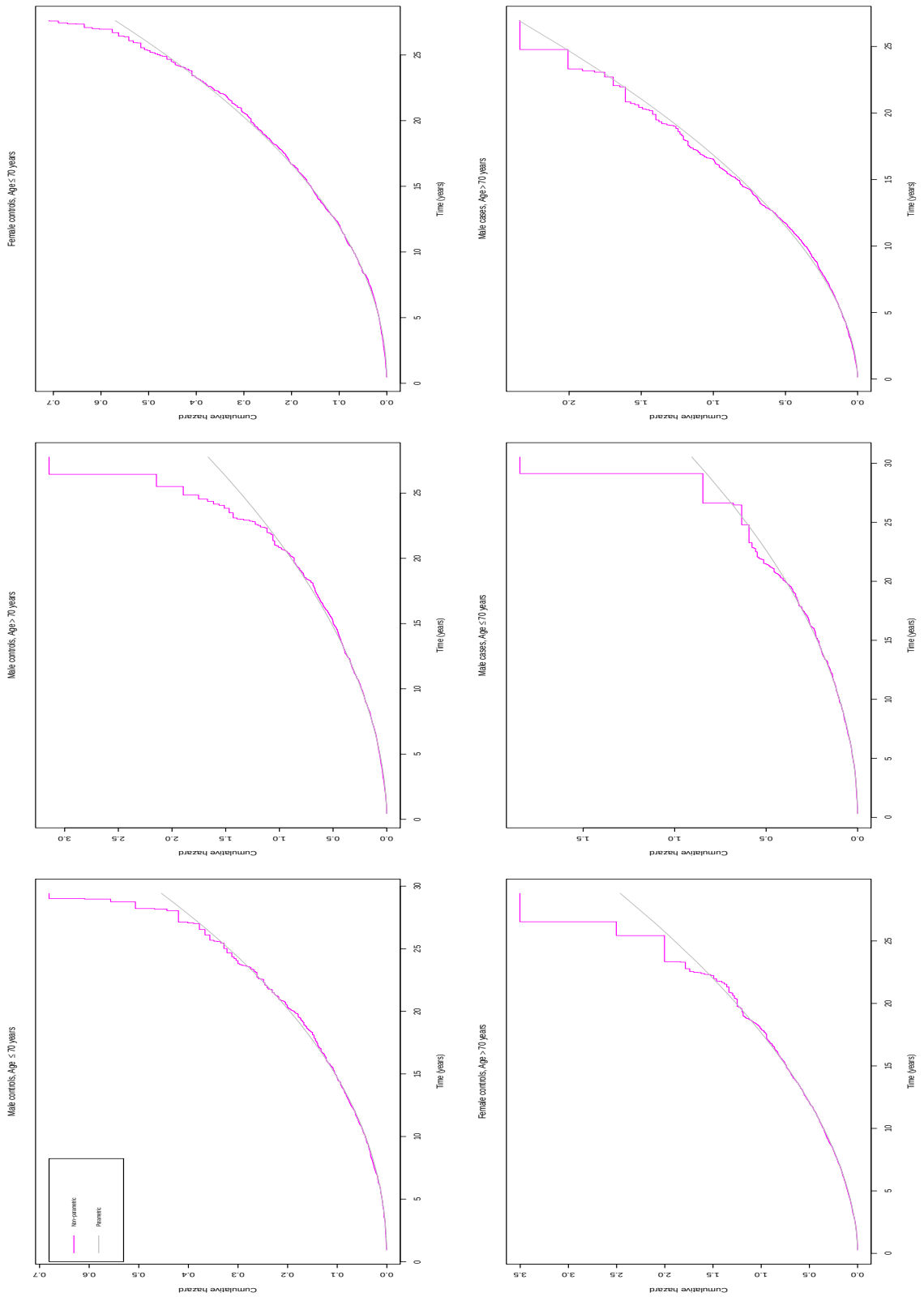


FIGURE 5.7: Comparison of the baseline cumulative hazard functions estimated using semi-parametric (magenta) and parametric (Weibull model, grey) methods. Age, sex and case-controls groups.

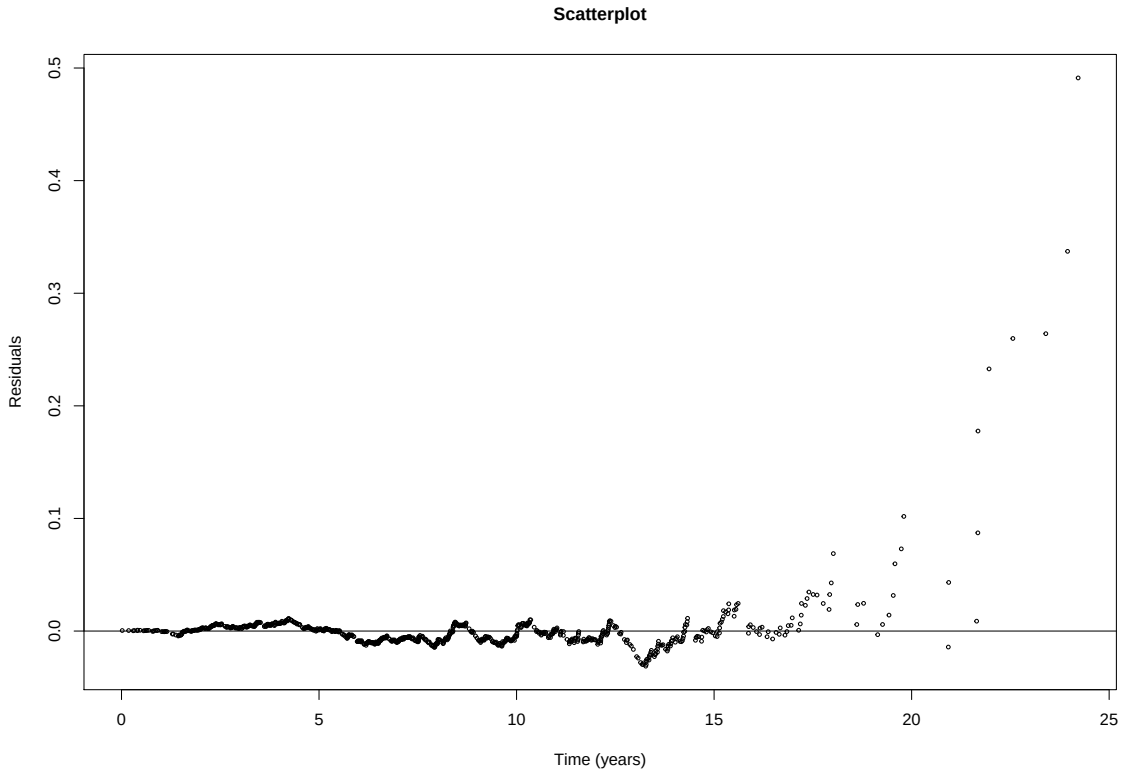


FIGURE 5.8: Plot of the residuals between the parametric and semi-parametric estimates of the baseline hazards pooled across the strata.

The goodness-of-fit was checked by comparing the semi-parametric estimates of the baseline cummulative hazard functions to the baseline cummulative hazards from the Weibull baseline hazards, separately within some other strata. The results are presented in Figure 5.7 for the Weibull baseline hazard across age groups, case-control status, and gender. The Weibull distribution describes the mortality really well up to at least 20 years. The difference (residuals) between the semi-parametric cummulative baseline and the fitted parametric were plotted in Figure 5.8. Deviations only were observed after 20 years. Overall, the Weibull distribution fitted the survival data pretty well.

Consequently, a Weibull double-Cox model was fitted to the prior model's covariates. The covariates which were violating the PH assumption were allowed to have shape effects. Significant shape and scale parameters of the significance level of 0.05 were retained.

5.3.2 Overview of Full - Case model: TIA

The final model is presented in Table 5.4. Birth cohorts of earlier decades had raised hazards. The birth cohort covariates had both shape and scale effects. Patients from 1941-1960 birth cohorts had significantly better survival outcomes than others. This may be a result of medical advances and better public health measures which have beneficially affected the life expectancy of patients.

The IMD category level 1 (Most deprived) had higher hazards compared to the other groups indicating the association of social deprivation with all-cause mortality. IMD Quintile 2 had HR of 0.88 with 95% CI (0.81 - 0.96) whereas IMD Quintile 5 had HR of 0.76 , 95% CI (0.69 - 0.84). Diabetic patients who were treated with anti-diabetics had HR of 1.82 with 95% CI (1.67 - 1.98) compared to those who were diabetes-free. Diabetic patients who were not treated with anti-diabetics had an HR of 1.32 with 95% CI (1.14 -1.53). This may be explained by the fact that Type 2 diabetes can sometimes initially be managed through lifestyle modification indicating lower severity.

Male patients were worse off than their female counterparts with an HR of 1.31 with 95% CI of (1.24 - 1.42). All the pre-morbid medical conditions were associated with a raised risk of all-cause mortality. For example, atrial fibrillation had an HR of 1.28 (1.15 - 1.413) and COPD, 1.71 (1.52 - 1.9). Anticoagulant therapy was associated with higher hazards, HR = 1.40, 95% CI (1.25 - 1.57). Smoking was associated with high hazards; ex-smokers had HR of 1.31 (1.23 - 1.39) and current smokers 1.90 (1.77 - 2.05), compared to non-smokers. Interestingly, current consumers of alcohol had better survival outcomes with HR of 0.9, 95% CI (0.85 - 0.95) compared to abstainers.
TIA diagnosis was associated with a high risk of all-cause mortality. TIA diagnosis had significant interactions in scale effects with age group and anti-hypertensive treatment and antiplatelet prescriptions. The model also had a significant frailty effect of the general practice with σ^2 of 0.077 with 95% CI (0.056 - 0.106).

TABLE 5.4: Parameter estimates of the Weibull double-Cox model for the full-case model.

Variables	Levels	Estimates	CI
Sample size		24,797	
Number of non-censored		6,188	
Weibull Parameters	$a(Scale)$	25.45	23.68−27.36
	$b(Shape)$	3.36	3.00−3.76
Shape parameters			
Birth Cohort	1900 to 1920	1	
	1921 to 1930	0.86	0.82−0.91
	1931 to 1940	0.7	0.65−0.75
	1941 to 1960	0.53	0.47−0.6
Heart Failure	No	1	
	Yes	0.82	0.78−0.87
BMI category	Normal	1	
	Obese + Overweight	1.07	1.03−1.11
Age Category	39−60	1	
	61−70	0.83	0.76−0.92
	71−76	0.78	0.7−0.87
	77+	0.66	0.59−0.74
Antiplatelet prescriptions	None	1	
	Aspirin only	0.91	0.87−0.95
	Dual therapy (Aspirin)	0.89	0.81−0.98
	Other Antiplatelets	0.85	0.76−0.94
Scale parameters			
Birth Cohort	1900 to 1920	1	
	1921 to 1930	0.54	0.49−0.6
	1931 to 1940	0.29	0.26−0.33
	1941 to 1960	0.12	0.1−0.15
Sex	Female	1	
	Male	1.31	1.24−1.42
IMD Quintile	1 (Most Deprived)	1	
	2	0.88	0.81−0.96
	3	0.88	0.8−0.96
	4	0.79	0.72−0.87
	5 (Least Deprived)	0.76	0.69−0.84
Case/Control diagnosis	Controls	1	
	Cases	2.29	1.89-2.77
Age Category	39-60	1	
	61−70	1.63	1.4−1.91
	71−76	2.34	1.97−2.76
	77+	3.73	3.16−4.4

Table 5.4 Continued

Variables	Levels	Estimates	CI
BMI	Normal + Underweight	1	
	Obese + Overweight	0.88	0.82−0.95
Asthma		1.26	1.16−1.37
COPD		1.71	1.53−1.91
CKD		1.18	0.99−1.4
Myocardial Infarction		1.33	1.22−1.45
Peripheral Arterial Disease		1.21	1.14−1.29
Atrial Fibrillation		1.28	1.15−1.43
Smoking	Non -smoker	1	
	Current smoker	1.9	1.77−2.05
	Ex-smoker	1.31	1.23−1.39
Alcohol	Non consumer	1	
	Consumer	0.9	0.85−0.95
Diabetes	No diagnosis/No treatment	1	
	Yes and Treated	1.82	1.67−1.98
	Yes and Untreated	1.32	1.14−1.53
Anticoagulant agents		1.4	1.25−1.57
Hypertension		1.45	1.34−1.57
Antiplatelet prescriptions	None	1	
	Aspirin only	0.966	0.91−1.03
	Dual therapy (Aspirin)	1.28	1.07−1.54
	Other Antiplatelets	0.919	0.76−1.12
Interactions (Scale effects)			
	Cases & Aspirin only	0.909	0.8−1.03
	Cases & Dual therapy (Aspirin)	0.646	0.41−1.01
	Cases & Other Antiplatelets	0.837	0.56−1.25
	Cases & Age 61−70	0.84	0.69−1.04
	Cases & Age 71−76	0.71	0.57−0.87
	Cases & Age 77+	0.59	0.48−0.73
	Cases & hypertension	0.91	0.81−1.02
Random Gamma frailty	σ^2	0.08	0.06−0.11
Concordance		79.2	

To assess the predictive value of our model, the Harrell's concordance index between the predicted and the observed survival was calculated. In the model with frailty, the estimate of the concordance was equal to 0.79.
We used graphical methods as well to check for the adequacy of the double-Cox Weibull model. The averages of the survival were calculated at various time points. We used different strata to check the fit of the expected survival functions with respect to the empirical survival functions. The graphical results in Figure 5.9 provide evidence of a good fit.The first plot shows that generally cases fit better than controls. Female cases are somewhat better fit than other groups. The first 3 birth cohorts are fitted well for the 15 years or so. The youngest cohort are somewhat worth.

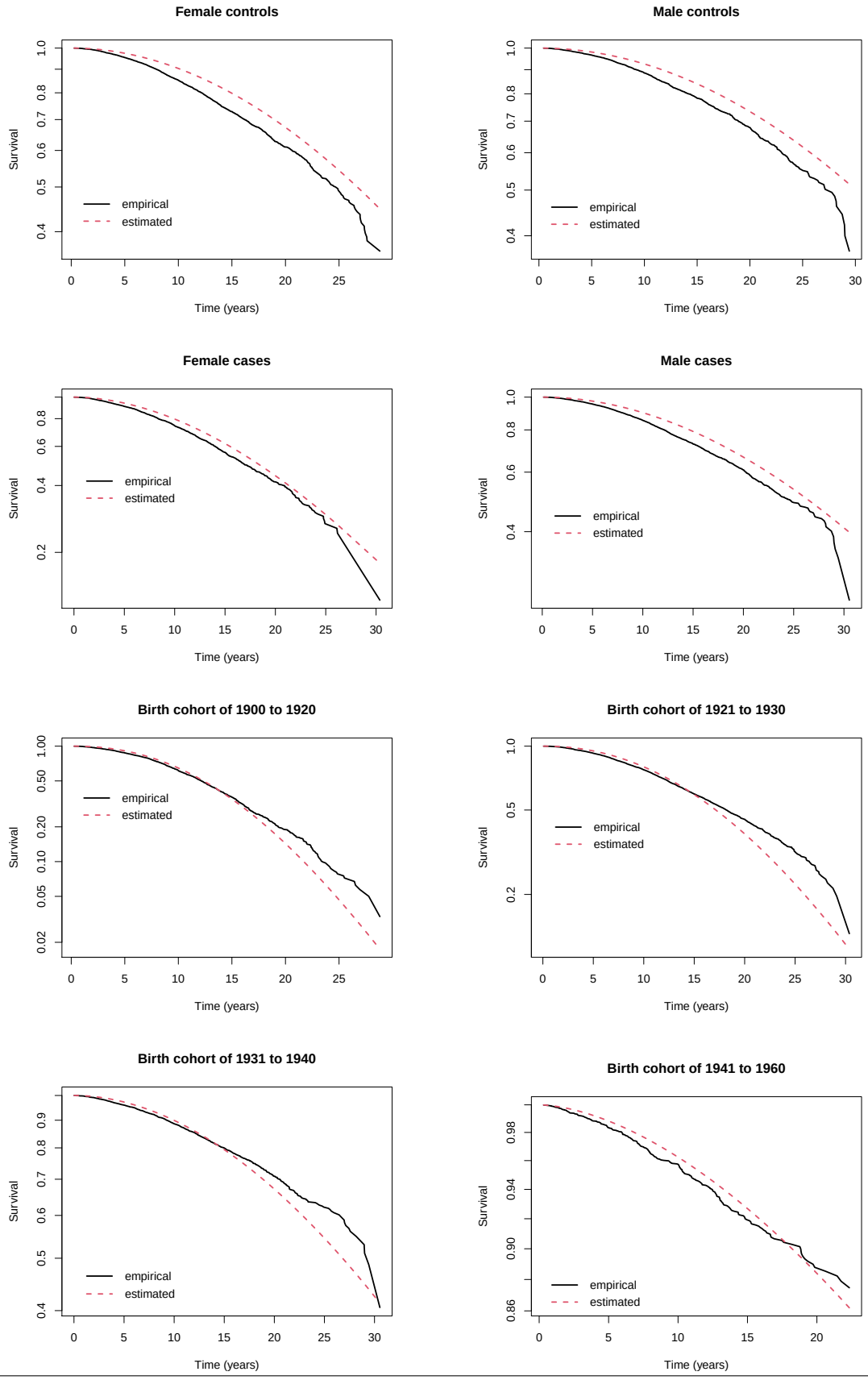


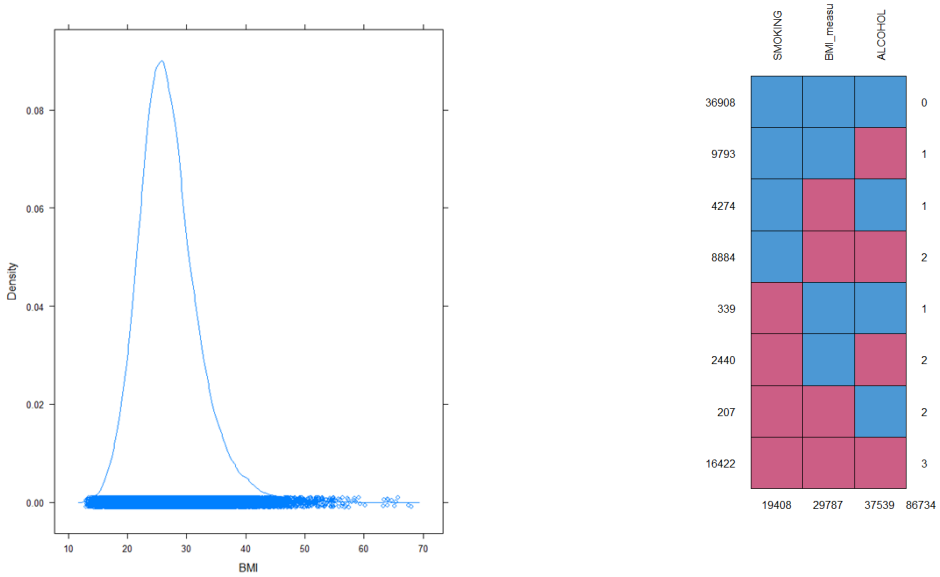
FIGURE 5.9: Comparison of fit of estimated survival to empirical across different strata.

5.4 Multiple Imputation

5.4.1 Exploring Pattern of Missingness.

The missing data mechanism was investigated by exploring patterns of missingness and by checking for associations between missing and observed data. The density of the BMI measurements with missing values is illustrated in Figure 5.10a. The density of the BMI was slightly right-skewed in full-case data. The missing data are depicted as blue dots below the kernel plot. This missing pattern is explored in Figure 5.10b and Table B.5. For example, there were 36,908 complete records. 9763 cases where Alcohol had missing values but BMI measurement and smoking were recorded. There were 164,422 total instances of missing values. The missing data pattern influenced the amount of information that could be transferred between variables). Variables may be missing on their own or together ("association") with other variables. A missing data pattern was observed which was *connected* (any observed data point can be reached from any other observed data point through a sequence of horizontal or vertical moves (like the rook in chess). This was suggestive of Missing at Random, *MAR* (Buuren & Groothuis-Oudshoorn, 2010).

We then tried to explore the correlation matrix of the missing variables with observed variables (see Table B.6). We detected a high correlation between missing variables, which again indicated MAR.



(A) Density plot with missing values in BMI measurement. ; the blue dots represents the missing values; the kernel plot is plotted on full records)

(B) Missingness pattern in the data, the blue box indicates full values and the pink box indicates missing values in the variables. The numbers of missing values per category is also explained in Table B.5

FIGURE 5.10: Missingness pattern in TIA dataset.

The number of missing values and the pattern of missingness helps to determine the likelihood of it being random rather than systematic. The missing data was assumed to be MAR and was handled using multiple imputation.

5.4.2 Multiple imputation model

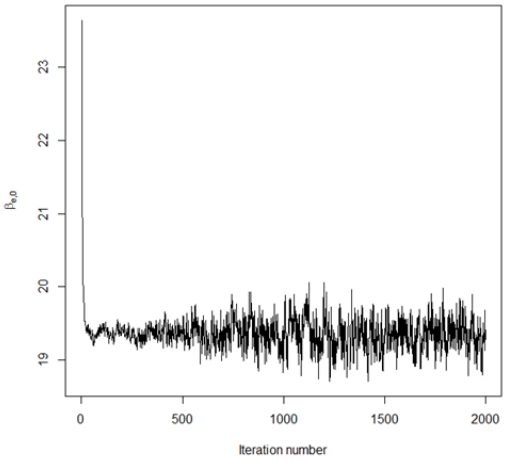
A multiple imputation model was built using the predictors from the full-case model. Multiple imputation with the Joint Modelling method was performed with the **R** package **JOMO** (Quartagno & Carpenter, 2016). The **JOMO** package is used to provide imputation for the clustered / multi-level data. The "*jomo1ranmix*" function helps to impute a clustered dataset with mixed data types as outcomes. A joint multivariate model for partially observed data is assumed and imputations are generated through the use of a Gibbs sampler where the covariance matrix is updated with a Metropolis-Hastings step. Fully observed categorical covariates may be considered as well, but they have to be included as dummy variables (Quartagno & Carpenter, 2016). The main entries are :

- A data frame with continuous responses of the joint imputation model: *BMI measurement* . After imputation, the BMI measurement was converted into levels.
- A data frame with categorical responses of the joint imputation model: *Smoking status and alcohol consumption.*
- A data frame with all other covariates: *groups, birth cohort, status, log(time), age, sex, asthma, atrial fibrillation, diabetes type 2 factor, CKD, COPD, heart failure, myocardial infarction, PVD, hypertension, antiplatelets, anticoagulants, the interaction of case-controls with age groups, antiplatelets, and anti-hypertensive drugs.* .
- cluster variable : *GP practice*

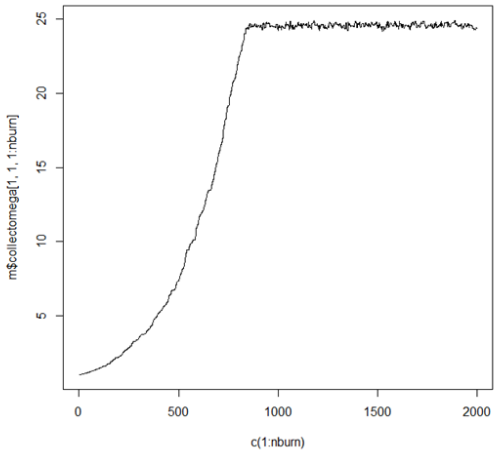
Replacement data for the missing values were generated in 10 imputations datasets. The number of iterations for the burn-in step was set to 2000. 100 iterations were run between each imputed dataset. *Jomo* function imputes by running Markov chain Monte Carlo (MCMC) and hence it is crucial to check convergence. As per Quartagno et al. (2019), convergence for elements was checked using :

- *collect_beta*: a three-dimensional array containing the fixed effect parameter draws at each of the n-burn iterations;
- *collect_omega*: a three-dimensional array containing the level-1 covariance matrix draws at each of the n-burn iterations;

The convergence of the first element of beta and the convergence of the first element of the level 2 covariance matrix was checked (Figure 5.11a and Figure 5.11b). In the current case, a burn-in of 500 - 900 was reasonable; the sampler converged very quickly. The frequency distributions of prognostic variables in the 10 datasets with imputed missing values were compared with the frequency distribution in the original dataset using histograms and basic summary statistics.



(A) convergence of the first element of beta

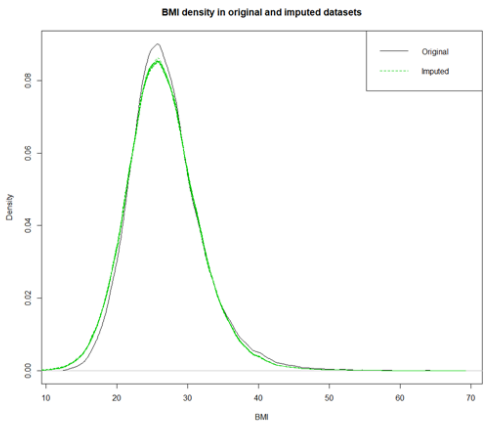


(B) convergence of any element of the level 2 covariance matrix.

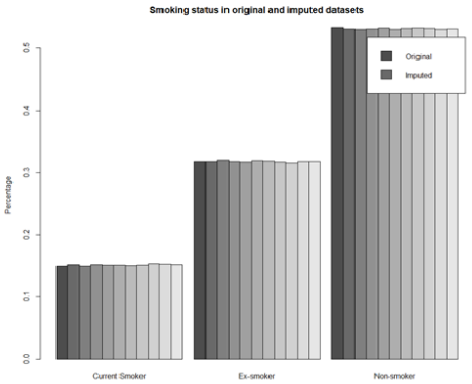
The density of the BMI imputed dataset is very similar to the full case density plot of Figure 5.12, with only a very slight shift to the left. The bar charts of the original and the imputed datasets with regards to alcohol consumption, smoking and IMD Quintiles were very similar; multiple imputation preserved the same distribution. The prevalences (%) of categorical factors in the original data (ignoring missing data) with the imputed dataset in Table 5.5) are provided.

TABLE 5.5: Comparing the distributions in complete case and completed dataset.

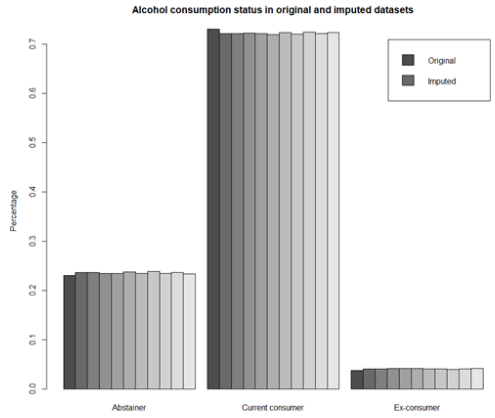
Covariates	Levels	**Original dataset** Number (%)	**Complete-case dataset** Number (%)	**Imputed dataset** Number (%)
Smoking status	Non-smoker	32140 (40.5%)	32140 (53.7%)	42213(53.3%)
	Ex-Smoker	19316 (24.4%)	19316 (32.3%)	25456 (32.1%)
	Current smoker	8403 (10.6%)	8403 (14%)	11598 (14.6%)
	Missing	19408 (24.5%)	-	-
BMI category	Underweight	1362 (1.7%)	1362 (2.8%)	2242 (2.8%)
	Normal	17678 (22.3%)	17678 (35.7%)	28708 (36.2%)
	Overweight	19680 (24.8%)	19680 (39.8%)	29966 (37.8%)
	Obese	10760 (13.6%)	10760 (21.7%)	18351 (23.2%)
	Missing	29787 (37.6%)	-	-
Alcohol consumption	Abstainer	8968 (11.3%)	8968 (21.5%)	17679 (22.3%)
	Ex-consumer	1616 (2%)	1616 (3.9%)	3325 (4.2%)
	Current consumer	31144 (39.3%)	31144 (74.6%)	58264 (73.5%)
	Missing	37539 (47.4%)	-	-



(A) BMI density, black : complete-case, green : Imputed dataset



(B) Bar chart for smoking status: the original dataset proportions of each level compared to the proportions of the other 10 imputed datasets.



(C) Bar chart for Alcohol status: the original dataset compared to the proportions of the other 10 imputed datasets.

FIGURE 5.12: Comparison of distribution of imputed variables with distribution of full case.

The Weibull Double-Cox model with parameters listed in Table 5.4 were fitted to the 10 imputed datasets and the parameter results were pooled using the Rubin's Rules. The estimates for the all-cause mortality hazard ratio in terms of scale and shape parameters from the full-case model were somewhat but not concerningly smaller than the estimates from the imputed data, except for very few estimates. Parameters from the two models had overlapping confidence intervals (See Table B.10). Hence, this suggests that the exclusion of full cases from the survival model may have caused an underestimation of the associated mortality risks. A significant frailty component with $\sigma^2 = 0.08$ for both the full-case and the multiple imputed model implied that the mortality varied strongly between GP practices. In terms of model performance, the full case model had a concordance c-index of 0.792 and the multiple imputed models had a similar high good measure of discrimination with a c-index of 0.792 (SE error 0.002).

5.5 Final models: Results from data from the imputed model.

TABLE 5.6: Final model : Description, parameter estimates and confidence intervals for the Weibull Double-Cox model with frailty terms.

Variables	Levels	Estimates	CI
Sample size		79,267	
Number of non-censored		24,176	
Weibull Parameters	$a(Scale)$	29.19	28.79-29.58
	$b(Shape)$	2.39	2.31-2.48
Shape parameters			
Age category	39-60 years	1	
	61-70 years	0.9	0.82-0.98
	71-76 years	0.77	0.7-0.85
	77+ years	0.59	0.51-0.67
Birth Cohort	1900 to 1920	1	
	1921 to 1930	0.84	0.82-0.86
	1931 to 1940	0.66	0.62-0.7
	1941 to 1960	0.48	0.4-0.56
Antiplatelet prescriptions	None	1	
	Aspirin only	0.96	0.94-0.98
	Dual therapy (Aspirin)	0.9	0.84-0.95
	Other Antiplatelets	0.89	0.83-0.95
Heart Failure		0.79	0.77-0.81
Scale parameters			
Birth Cohort	1900 to 1920	1	
	1921 to 1930	0.5	0.44-0.56
	1931 to 1940	0.23	0.15-0.31
	1941 to 1960	0.09	0.03-0.2
Sex	Female	1	
	Male	1.3	1.28-1.33
IMD Quintile	1 (Most Deprived)	1	
	2	0.91	0.87-0.97
	3	0.92	0.88-0.96
	4	0.87	0.81-0.93
	5 (Least Deprived)	0.81	0.75-0.87
Case/Control status	Controls	1	
	Cases	3.04	2.91-3.18
Age Category	39-60	1	
	61-70	1.77	1.66-1.89
	71-76	2.12	2-2.24
	77+	2.65	2.55-2.76

The results developed are presented in forest plot on Figure 5.13, in Table 5.7 and hazard curves on

Table 5.6 Continued

Variables	Levels	Estimates	CI
BMI	Normal + Underweight	1	
	Obese + Overweight	0.87	0.82-0.92
Asthma		1.16	1.12-1.2
COPD		1.73	1.67-1.79
CKD		1.04	0.96-1.11
Myocardial Infarction		1.2	1.16-1.24
Peripheral Arterial Disease		1.14	1.1-1.18
Atrial Fibrillation		1.27	1.21-1.32
Smoking	Non-smoker	1	
	Current smoker	2.04	1.96-2.11
	Ex-smoker	1.24	1.2-1.29
Alcohol Consumption	Non-current	1	
	Current consumer	0.89	0.85-0.93
Diabetes	No diagnosis/No treatment	1	
	Yes and Treated	1.52	1.48-1.56
	Yes and Untreated	1.12	1.04-1.2
Anticoagulant agents		1.23	1.17-1.29
Hypertension		1.41	1.18-1.59
Antiplatelet prescriptions	None	1	
	Aspirin only	1.01	0.95-1.07
	Dual therapy (Aspirin)	1.11	0.9-1.33
	Other Antiplatelets	0.99	0.79-1.18
Interactions (Scale effects)			
	Cases & Hypertension	0.92	0.86-0.97
	Cases & Aspirin only	0.88	0.83-0.94
	Cases & Dual therapy (Aspirin)	0.75	0.53-0.96
	Cases & Other Antiplatelets	0.86	0.66-1.06
	Cases & Age 61-70	0.65	0.51-0.79
	Cases & Age 71-76	0.54	0.41-0.68
	Cases & Age 77+	0.5	0.36-0.64
σ^2	Frailty	0.08	0.04-0.13
Concordance	C-index	0.79	

Figures 5.15 & 5.14, and hazard ratio time trend plots on Figures 5.16 & 5.17. The adjusted hazards had both constant and time-varying (shape) effects. The constant (scale effects only) hazard ratios during a follow-up of 20 years implied that the patients of a given medical condition or patients on a certain medical therapy seem to have a constant long-term worsened survival prospects relative to patients free of the given medical condition or patients from the no-treatment comparison group.

5.5.1 Time-invariant effects

The highest hazard ratio was the current smoker with HR = 2.04 $(1.96 - 2.11)$ relative to the non-smoking group. Ex-smokers, were associated with HR = 0.89 $(0.85 - 0.90)$. Current consumers of alcohol had a reduced HR = 0.89 $(0.85 - 0.93)$ compared to abstainers. Being obese or overweight

conferred lower hazards of death with associated HR = 0.87 (0.82 – 0.92) compared to the normal-weight group. This finding suggests the “obesity paradox" whereby obesity being a major risk factor for cardiovascular disease-related conditions such as myocardial infarction and heart failure, seems to provide survival benefits compared to non-obese individuals (Uretsky et al., 2007). Male patients were worse off with HR = 1.3(1.28 – 1.33).

There were mixed survival prospects associated with pharmacotherapy. The prescriptions of anticoagulants had an HR = 1.23 with 95% CI of (1.17 – 1.29) compared to patients with no prescriptions. This unusual result however could be a result of this factor acting as a proxy for atrial fibrillation. There were not many cases of under-prescribing of anticoagulants to patients if they had a clinical diagnosis of atrial fibrillation.

Most of the medical conditions were associated with worse survival prospects. The hazard ratios ranged from a maximum of HR = 1.73 (1.67 – 1.79) for patients diagnosed with chronic obstructive pulmonary disease to peripheral arterial disease with HR = 1.4(1.1 – 1.18).

The findings also showed that survival prospects differed by relative social deprivation measured by the Index of Multiple deprivations (IMD) from level 1 (the most deprived) to level 5 (the least deprived). The hazards of mortality ranged from HR = 0.81 (0.75 – 0.87) to 0.92 (0.88 – 0.96) from level 5 to level 2 compared to level 1 suggesting that patients have a worse long-term prognosis if they belonged to the most deprived area. This raises the question of a possibility of socio-economic gradient and perhaps barriers to proper health care services, lack of awareness of good health habits, food insecurity, higher levels of stress due to financial stress, poor housing, surrounding violence, and lack of physical activity prevailing in the in low-income neighbourhoods.

GP practices had significant survival differences with a significant gamma frailty effect of 0.08 (0.09 – 0.13) after adjusting for other factors. This may indicate substantial variation in the quality of care between GP practices.

5.5.2 Time-varying effects

The time-dependent factors were birth cohort, antiplatelets, heart failure, and age category of TIA diagnosis. The effects of these factors were dependent on the time of exposure. The birth cohort effect is explored in Figure 5.14 which depicts how the adjusted long-term hazards declined for patients born in later decades possibly due to the increased knowledge and improved management of risk factors over decades.
The study found significant two-way interactions of diagnosis of TIA with age at diagnosis of TIA, the types of prescriptions of antiplatelets and hypertension. The effect of antiplatelets is illustrated in Figures 5.16 and 5.17 using hazard ratio time trends with an associated 95% confidence interval. To estimate the 95% bootstrap confidence interval for the time-varying hazards estimates, 10,000 realisations of the estimated parameters were created using the expected value and variance of the estimators of the final model of the Weibull Double-Cox model (Model 5.6). Using hazard ratio curves with 95% confidence interval bands, the effects of the different types of antiplatelets were explored over 15 years of follow-up. The hazard ratios were calculated as the ratio of hazards of a given antiplatelet type compared to non-users of antiplatelets. Different panels refer to different age groups and birth cohorts. From Figure 5.16 (a), adjusted hazard ratio curves are plotted for TIA patients on aspirin and TIA-free patients on aspirin relative to their respective non-user counterparts. The second sub-Figure 5.16 (b) provides the plot of the dual therapy with aspirin and other antiplatelets options relative to non-users of antiplatelets. The findings are presented in Tables B.12 and B.13.

The results show that aspirin prescription was protective in the long term for patients diagnosed with TIA at $39-60$ years. Significant benefit can be observed after 7 years with HR = 0.9 $(0.82 - 0.98)$ and HR = $0.88(0.8 - 0.96)$ at 10 and 15 years respectively. Aspirin was associated with insignificantly reduced survival prospects in the TIA-free controls as that the survival confidence bands overlapped the threshold of HR = 1.

Modest insignificant survival benefits were also observed for dual therapy(aspirin with other options) and other antiplatelets options (dipyridamole and clopidogrel) for TIA patients. However, there was no significant survival reduction due to the uptake of dual therapy and other antiplatelets for the controls. Similar benefits were also observed for the other age groups.

TABLE 5.7: Adjusted estimated hazard ratios of all-cause mortality by age at diagnosis and case-control status (from the Weibull Double-Cox model on the imputed data).

Comparison group	Relative to:	HR (95% CI)
Cases aged 39 – 60 years	Controls aged 39 – 60 years	3.04 (2.91 – 3.18)
Cases aged 61 – 70 years	Controls aged 61 – 70 years	1.98 (1.55 – 2.30)
Cases aged 71 – 76 years	Controls aged 71 – 76 years	1.72 (1.20 – 2.07)
Cases aged 77+ years	Controls aged 77+ years	1.52 (1.15 – 1.97)

TABLE 5.8: Adjusted estimated hazard ratios of all-cause mortality by hypertension status and case-control status (from the Weibull Double-Cox model on the imputed data)

Comparison group	Relative to:	HR (95% CI)
Hypertensive controls	Non-hypertensive controls	1.41 (1.18 – 1.59)
Non-hypertensive cases	Non-hypertensive controls	3.04 (2.91 – 3.18)
Hypertensive cases	Non-hypertensive controls	3.94 (3.66 – 4.27)

Overall, it appears that aspirin uptake by patients with a diagnosis of TIA compared to the no-treatment group seems to have a beneficial effect on survival prognosis over the long term. Our findings highlight the comparative efficacy of aspirin for TIA patients in terms of mortality reduction and favours the latter for the long-term secondary preventive care as opposed to other options of antiplatelets.

Prior hypertension was associated with adverse mortality with HR = 1.41 (1.18 – 1.29) for controls and the risk was 4 fold higher with a diagnosis of TIA with HR = 3.94 (3.66 – 4.27) (see Table 5.8). The concomittant effect of hypertension and TIA was quite concerning.

The adjusted hazard ratios were calculated and their associated 95% confidence bands were calculated using the Bootstraps methods. Cases and controls were compared in each age group. It is worth noting that the shape effects were not apparent in Table 5.7. This is because while comparing same age categories and hazard ratios being a ratio of hazards leading to the ratio of same shape effect to be one; the comparisons were hence only scale effect difference. The relative hazards of mortality of cases diagnosed with TIA at 39 – 60 years were 3 times to their matched controls in the long term, HR = 3.04 (2.91 – 3.18. The mortality risks were 1.5 to 2 times for more elderly patients compared to their counterparts free of TIA. The diagnosis of TIA did not influence much the already existing risks associated with other medical conditions due to ageing.
Overall, the study showed that TIA patients experienced a significantly higher rate of mortality compared to their matched controls hence highlighting the need for managing modifiable lifestyle and medical risk factors in the long term after the first episode of transient ischaemic attack, the supposedly “harmless” neurological condition.

5.6 Morbidity

TIA patients experienced a higher risk of cardiovascular morbidity in the short term and in the long term (Table 5.9). 17.5% of TIA patients experienced at least one recurrent TIA event in the first year following the initial event. 3.2% of the cases had a recurrent TIA in the next 2 years. More concerning, 3.6% patients and 2.6% TIA patients had a stroke or a myocardial infarction respectively, in the first year after the index event compared to controls (0.3% and 1.3%). The results indicated significant difference in proportion of vascular events between cases and control with exception at 5 years and more.

TABLE 5.9: Vascular events at follow-up of the TIA cohort.

Morbidity at follow-up	**Follow-Up time**	**Cases (n = 20,633)**	**Controls (n = 58,634)**	**χ^2 test(p-value)**
Recurrent TIA event	0-1 year	3608(17.5%)	127(0.2%)	<0.001
	1-3 years	657(3.2%)	199(0.3%)	<0.001
	3-5 years	301(1.5%)	148(0.3%)	<0.001
	5 years +	490(2.4%)	249(0.4%)	<0.001
Stroke	0-1 year	746(3.6%)	174(0.3%)	<0.001
	1-3 years	232(1.1%)	275(0.5%)	<0.001
	3-5 years	119(0.6%)	192(0.3%)	<0.001
	5 years +	185(0.9%)	530(0.9%)	0.9582
Myocardial Infarction	0-1 year	531(2.6%)	740(1.3%)	<0.001
	1-3 years	571(2.8%)	1123(1.9%)	<0.001
	3-5 years	460(2.2%)	895(1.5%)	<0.001
	5 years +	910(4.4%)	2433(4.1%)	0.1132

5.7 Life Expectancy model for TIA

The life expectancy model is presented in Table B.11. The model was built using a Weibull Double-Cox model as it had time-varying effects on some variables.

The dataset used for the survival study after transient ischaemic attack was also used in this analysis for actuarial translation and the findings are presented in Table B.11. The reason of separate modelling was addressed in Section(3.15). This time, however, age was used on a continuous scale than age in categories.

A full case model was first considered and missing records were imputed using the Multiple Imputation method. An initial simple non-parametric Cox's proportional hazard model with all main effects and two-way interactions of all variables was used. The primary outcome was all-cause mortality. Selected variables were based on the literature review and NICE guidelines for the treatment and

management of transient ischaemic attacks (NICE, 2020). Diagnoses of medical conditions were defined by the standard list of clinical codes and drug prescriptions corresponding to the British National Formulary for lipid-lowering, anticoagulant, antiplatelet, antidiabetic, and antihypertensive drugs. Further details about levels and coding of covariates are provided in Table 1 (B.3 Appendix). The covariates consisted of gender, year of birth, case-control status, age at diagnosis, prior comorbidities such as asthma, chronic kidney disease (CKD), chronic obstructive pulmonary disorder (COPD), heart failure, hypothyroidism, myocardial infarction, peripheral vascular disease (PVD), hypertension, hypercholesterolemia, atrial fibrillation, diabetes type II, prescriptions of anticoagulants, antihypertensive, antiplatelet and antidiabetic drugs at baseline. Possible dependencies of the survival outcomes within a general practice were modelled by inclusion of a random frailty term. Backward elimination with $p < 0.05$ for fixed effects and $p < 0.01$ for interaction effects was used to obtain a parsimonious model. After a backward elimination process, the parsimonious model exhibited non-proportionality properties indicating time dependence of some covariates. Before fitting the double-Cox Weibull model, we first compared the semi-parametric estimates of the baseline hazards with different parametric models and the Weibull distribution was chosen. Since the same full-case cohort was used, so the percentage of missing values and the bivariate association of predictors with time were the same as in the full-case survival model in 5.4. The dataset had missing records on smoking status and BMI category which were imputed later using Multiple Imputation method. The final model is presented in Table B.11 with the estimates of the scale, shape effects, and their associated 95% CI. The full-case model and the model pooled by Rubin's rules after Multiple Imputation are quite comparable.

The birth cohort, gender, IMD quintile, BMI category, antiplatelets, COPD, heart failure, myocardial infarction, PAD, smoking status, alcohol consumption, anticoagulants, diabetes, case-control diagnosis, hypertension, heart failure, CKD, asthma, atrial fibrillation age as continuous and the interaction of case-control diagnosis with antiplatelet were retained by the model. Both models have similar parameter estimates and fit the data well, with the concordance of 63.1% and 66.3%, respectively. The frailty variance was highly significant but rather low at 0.07 and 0.08 in the imputed and the complete case models, respectively

The age of the patients, risk factors, the model parameters from Table B.11, time of follow-up after study entry and other covariate values are then used to find the relevant survival probabilities and the integral. They are calculated using Equation 3.84 to get estimates for life expectancies.

For illustration, the life expectancies of some key ages and follow-up times were explored for particular risk profiles. Table 5.10 shows the estimated life expectancy of the male patients aged 50, 60,70 and 80 years, respectively. Some comorbidities such as hypertension and diabetes are quite prevalent in the geriatric population. From the life expectancy model, the scale effects were 1.33,1.78, 2.00, and 1.64 for hypertension, diabetes Type II (treated), smoking, and diagnosis of TIA, respectively.

Our findings show that a male patient may lose an estimated 3.6, 2.48, 1.6, and 0.72 years if TIA occurs at ages 50, 60, 70 and 80, respectively.

Similar to the survival model after TIA, the model fit is tested through plots of the estimated and the empirical graphs. Figure 5.18 hence shows that cases and controls fit the data pretty well but a slightly better fit is observed for the controls. Both male and female groups fit the data well.

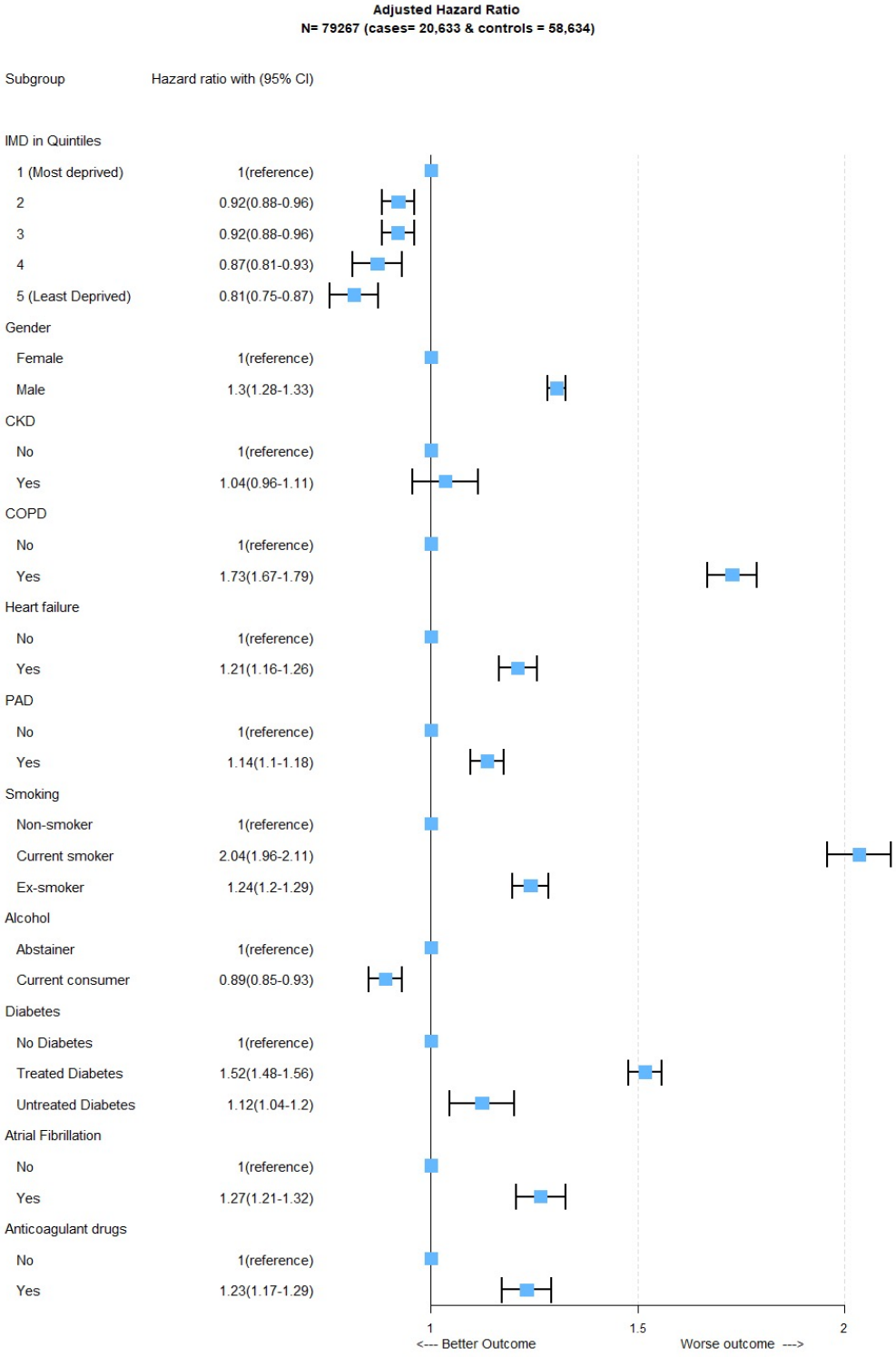


FIGURE 5.13: Forest plot of the adjusted hazard ratios for the factors having scale effect only.

Factors with no interactions with case control status.

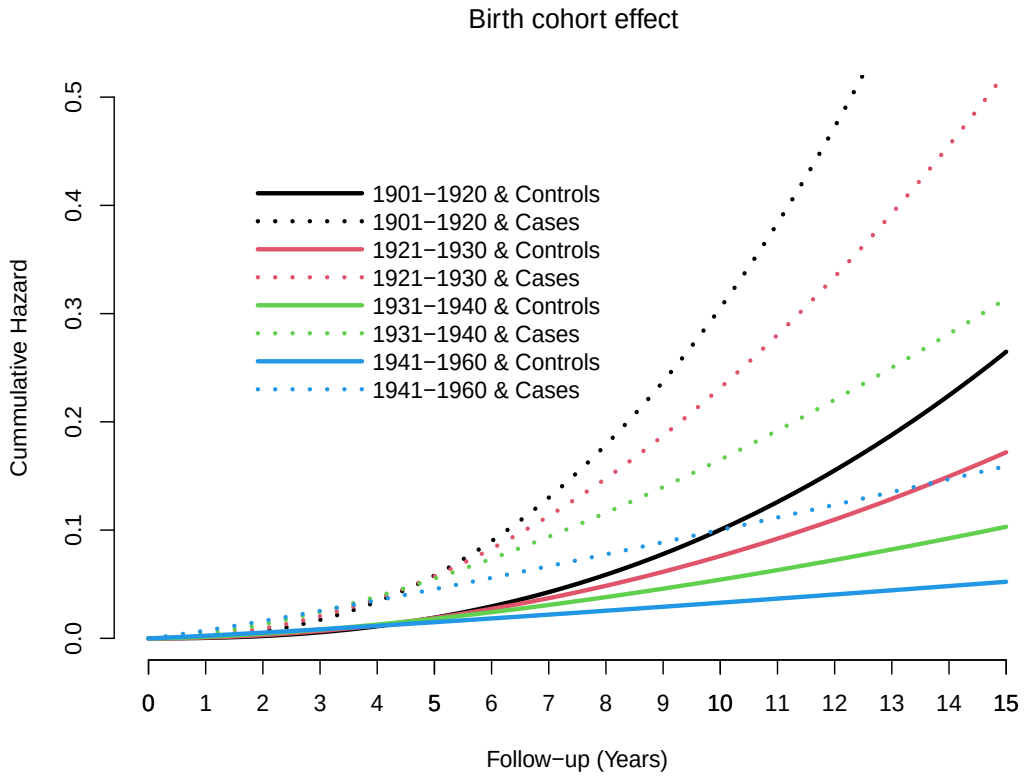


FIGURE 5.14: Cummulative hazard curves of the different birth cohorts for male cases and controls.

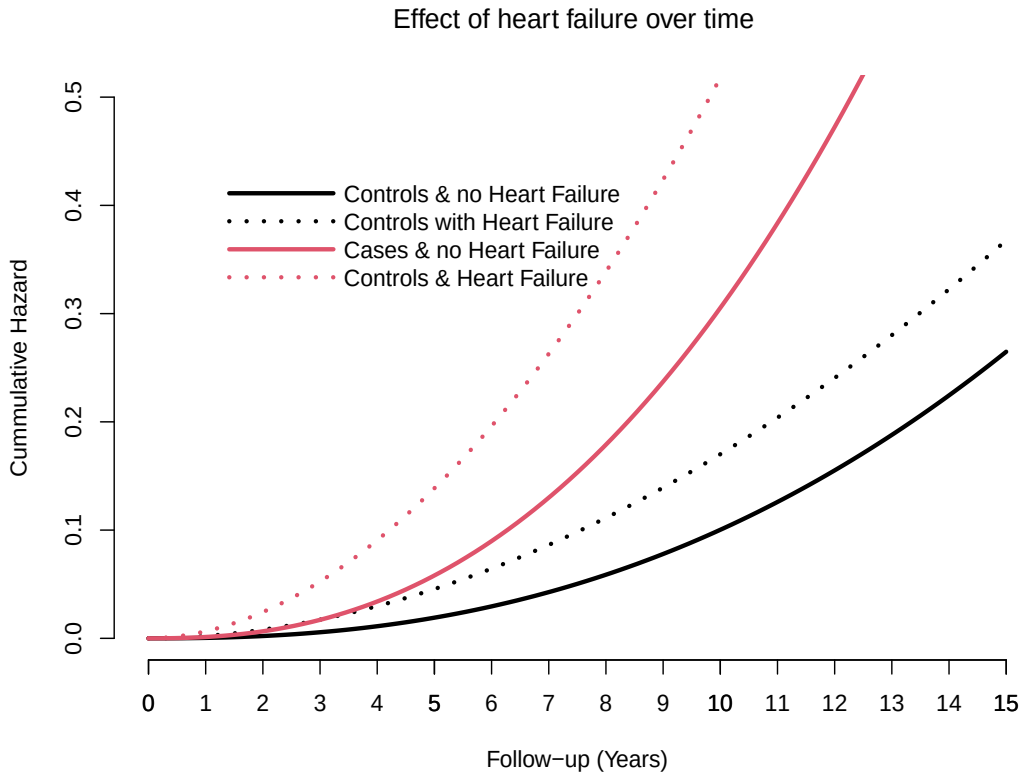


FIGURE 5.15: Cummulative hazard curves of the effect of heart failure on male cases and controls.

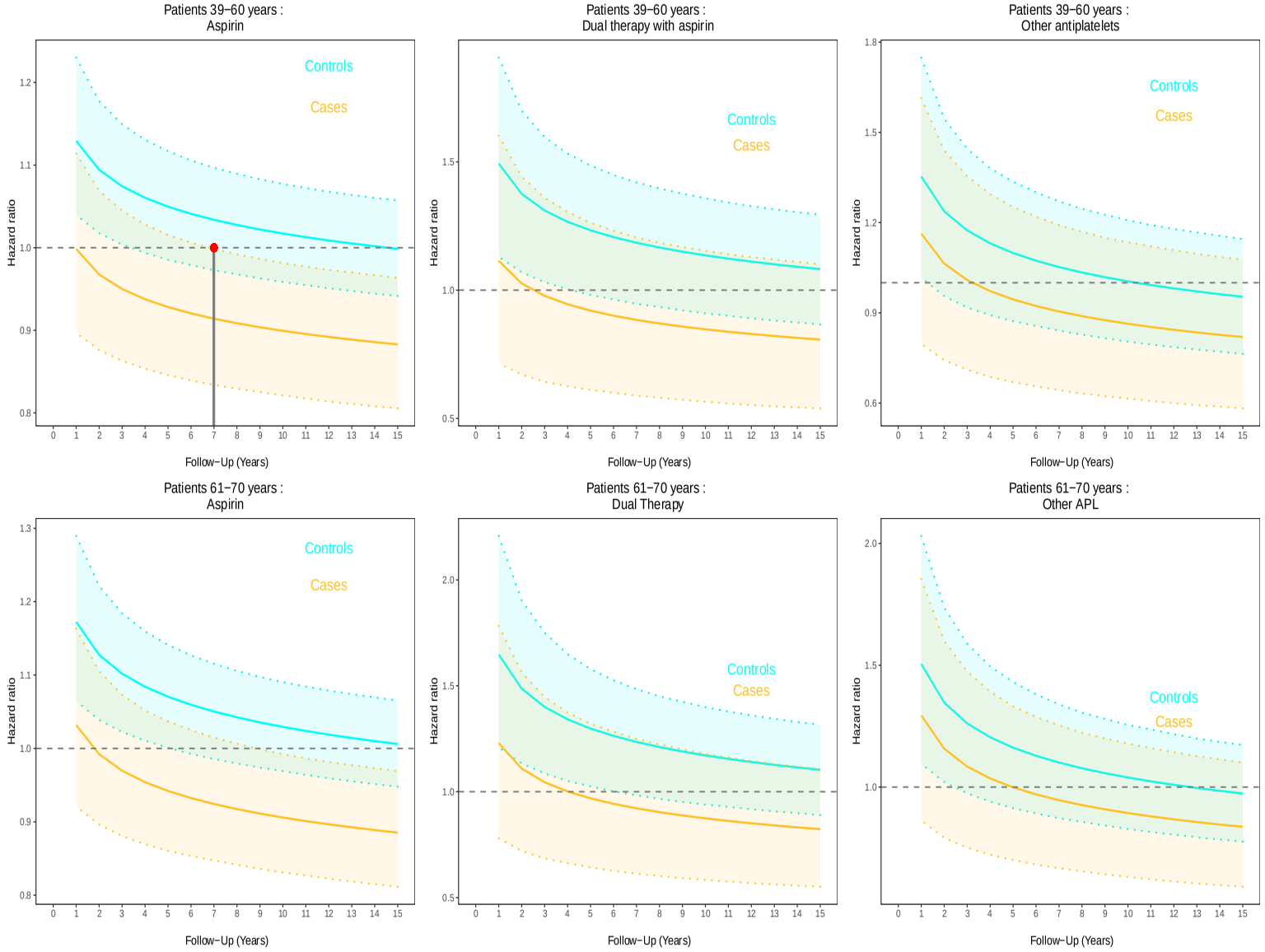


FIGURE 5.16: Hazard ratio curves with 95% confidence intervals for patients aged 39-60 years at entry (top panel) and aged 61-70 years at entry (bottom panel) different antiplatelets.

Cases on APL prescriptions were compared to cases who are non-users of APL. Controls on APL were compared to controls who are non-users of APL. *APL*: antiplatelet, *DAPT*: Combination of aspirin with other antiplatelet, *aspirin only*: aspirin monotherapy. *Other APL* include dipyridamole or/& clopidogrel).

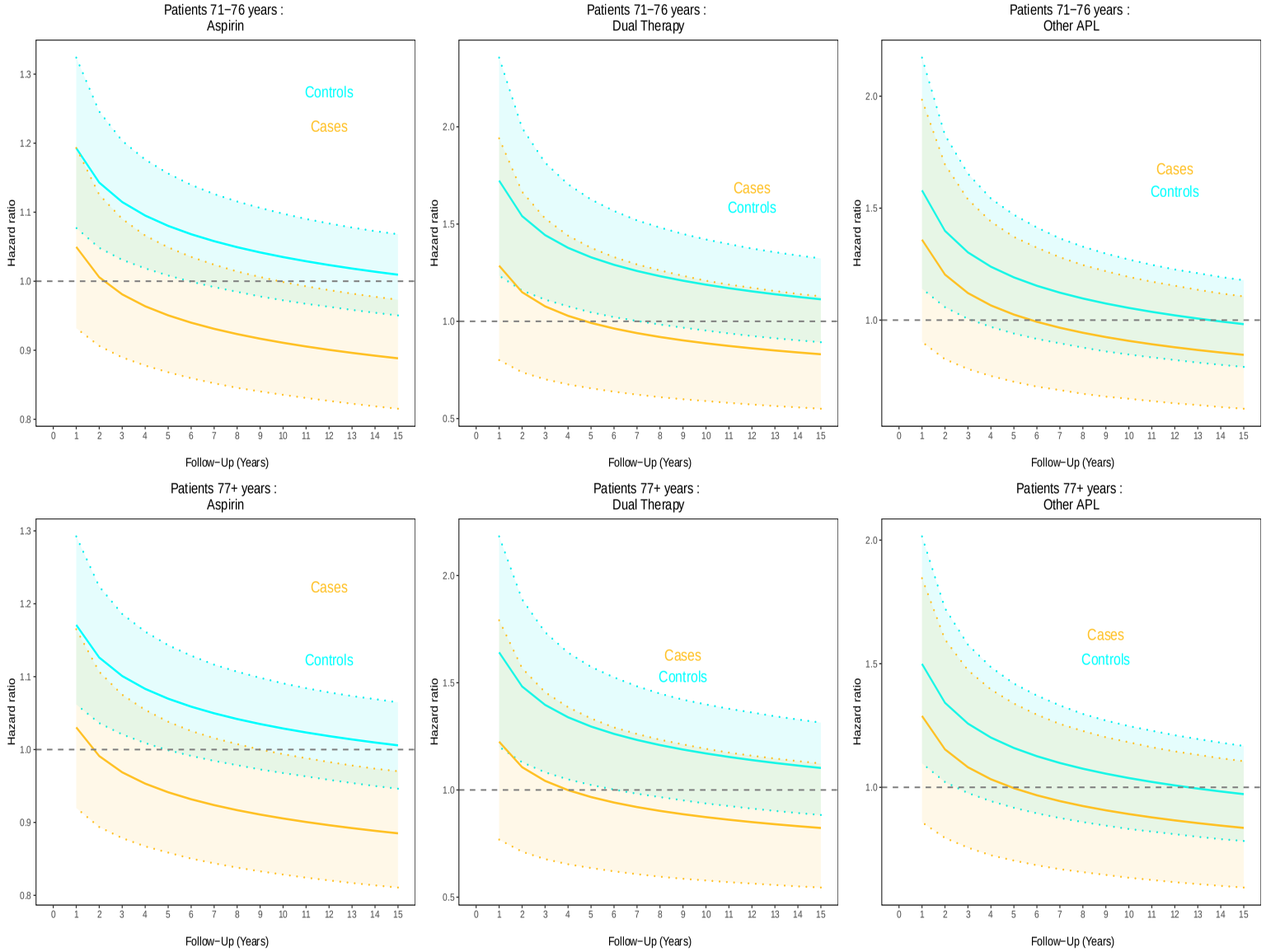


FIGURE 5.17: Hazard ratio curves with 95% confidence intervals for patients aged 71-76 years at entry(top panel) and aged 76 and above at entry(bottom panel) years at entry on different antiplatelets.

Cases on APL prescriptions were compared to cases who are non-users of APL. Controls on APL were compared to controls who are non-users of APL. *APL*: antiplatelet, *DAPT*: Combination of aspirin with other antiplatelet, *aspirin only*: aspirin monotherapy. *Other APL* include dipyridamole or/& clopidogrel).

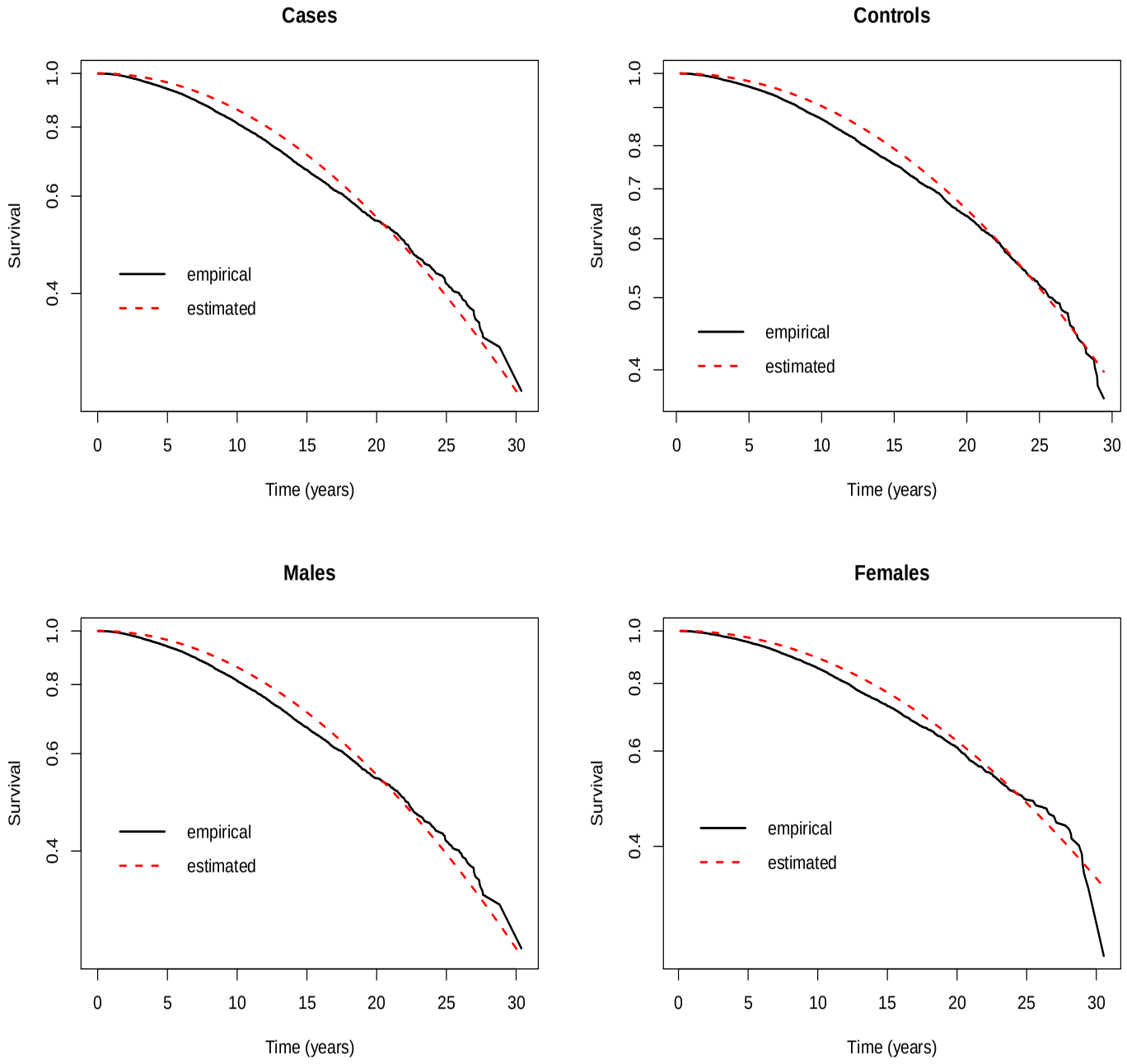


FIGURE 5.18: Comparison of empirical and estimated survival plots for different strata.

TABLE 5.10: Calculation of life expectancies of individuals in the TIA dataset at different ages.

Medical Condition	**50**	**60**	**70**	**80**
Scenario 1	15.60 (12.90 - 18.30)	10.15 (8.30 - 12.00)	6.40 (5.27 - 7.53)	4.75 (3.59 - 5.91)
Scenario 2: Scenario 1 + TIA diagnosis	12.0 (10.70 - 13.30)	7.67 (6.49 - 8.86)	4.80 (4.17 - 5.43)	3.32 (2.86 - 3.78)

Note. Scenario 1 implies that the individual has a diagnosis of hypertension and diabetes Type II and is a smoker.
Scenario 2 follows the Scenario 1 with the individual further experiencing TIA.
These results are estimates pertaining to 5 years after diagnosis.

5.8 Discussion

The study on survival after TIA is a large sample retrospective case-control study of patients adjusted for a wide variety of comorbidities and treatments and with a long follow-up. Our findings demonstrate a considerable excess risk of death in patients after the first event of TIA compared to patients free of TIA matched by age, sex and GP practices.

We found the hazard ratio of TIA diagnosis ranging from 1.52 to 3.04 compared to matched controls depending on the age groups at diagnosis. The findings are in accordance with (Edwards et al., 2017a; Gattellari et al., 2012; Gresham et al., 1998; Hankey, 2003; Hardie et al., 2003; Jacob et al., 2020) amongst others, reporting a higher risk of mortality of TIA patients as opposed to controls ranging from 2 to 4 fold more risk of death. Our adjusted hazards of mortality were slightly lower than previously estimated. This difference can be explained by short follow-up, the lack of adjustment of confounders by previous studies, and different data sources such as hospital data, registers, and community care data. The variation of risk of mortality was unlikely to be gender-related or birth cohorts related. There was no evidence of interactions of case-control status by age or sex in the current study, indicating the absence of gender-related mortality effects after a TIA event. Besides, the absence of interaction of the birth cohort with TIA indicated no improvements in health care management related to TIA. Nonetheless, our findings reinforce the conclusion that survival after a TIA event is an indicator of major long-term risk which warrants attention towards mitigating the risk. Patients diagnosed with TIA constitute a high-risk population for recurrent stroke, cardiovascular events, and hospitalisation with at least half of the patients being re-admitted to hospitals within the first year (Lichtman et al., 2009). The economic, social and financial burden and the unexpected deaths resulting from recurrent strokes or vascular events related to TIA aftermath could be prevented if TIA-associated risk of mortality is recognised by different individuals, carers, parents, doctors and policymakers.

Most of the studies on TIA have reported risk in the short-term period where the risk of adverse outcomes such as stroke, myocardial infarction, or death following a TIA is high. These studies had a limited follow-up ranging from 3 months to 3 years Anderson et al. (2004), Clark et al. (2003), and Gresham et al. (1998). Furthermore, their recruitment period ended many years ago when the clinical management of TIA and stroke, secondary management, and rehabilitation were not properly endorsed. Additionally, some of these studies had limited small sizes (Anderson et al., 2004; Carter et al., 2007; Clark et al., 2003; Gresham et al., 1998) while some studies reported on absolute mortality risk (Daffertshofer et al., 2004), or put more emphasis on morbidity (primary outcome being the occurrence of secondary vascular events following a TIA (Coull et al., 2004; Edwards et al., 2017a; Jacob et al., 2020) and hence not focusing on the mortality attributed to TIA.

Our analysis further examined the second-order interactions of TIA diagnosis. Significant interactions were detected with age at TIA events, hence providing new insights of how the long-term

hazards of mortality may vary by age groups in clinical care. The risk of mortality of patients diagnosed with TIA at $39-60$ years was 3 times higher compared to patients of the same age group but free of TIA. There is a dearth of evidence pertaining to the long-term mortality risk in different age groups. Gattellari et al. (2012) found increasing age of cases associated with an increased risk of death compared to controls ($p < 0.001$) in a study based on 22,257 adults hospitalised with a diagnosis of TIA in Australia. The team reported HR = 1.82 $(1.14-2.89)$, HR = 4.74 $(3.07-7.30)$ and HR = 7.77 $(5.08-11.89)$ for patients aged $50-64$ years, $64-74$ years and $75-84$ years relative to the younger groups ($<$ 50 years). However, they only adjusted for age, sex, year of discharge, and follow-up interval while our study adjusted for more variables.

In the current study, the relative risk of cases and controls declined for elderly patients. On a cautionary note, this does not translate to the risk of mortality decreasing for older patients but only implies that the effect of TIA diagnosis is not as pronounced as in younger patients. These findings suggest that older patients already have unfavourable medical conditions related to ageing and hence the impact of TIA increases the risk of mortality by only 1.5 times whereas the younger patients diagnosed with TIA, have a higher hazard ratio of almost 4 compared to their matched controls. It agrees with the findings of Rutten-Jacobs et al. (2013) indicating that stroke at a young age continues to contribute increased risk of death throughout patients' lives, even two decades after the index event and perhaps, the younger patients have the most to gain from intensive cardiovascular risk management even if they are stable after a TIA (Johnston et al., 2000).
The study also found second-order interaction of TIA with a clinical diagnosis of hypertension. The elevated mortality risk was 1.41 due to hypertension in controls but exacerbated in cases with HR = 3.94 $(3.66-4.27)$. Blood pressure is not only a risk factor for TIA and ischemic stroke (Khare, 2016) but our findings further show that the diagnosis of hypertension combined with a TIA diagnosis raises the risk of mortality considerably and are in line with Jacob et al. (2020)'s findings of HR = 1.95.

The study found prescription of aspirin to be protective of survival for cases but having negative though not significant survival effects on controls. Aspirin had long-term survival benefits in the younger age group of patients with HR = 0.93 $(0.84-1.01)$, HR = 0.9 $(0.82-0.98)$ and HR = 0.8 $(0.8-0.96)$ at 5, 10 and 15 years respectively. The benefits of the therapy were also observed for the other age groups. Other antiplatelets such as dipyridamole and clopidogrel had insignificant survival benefits in some cases. Dual therapy with aspirin also had uncertain effects in some cases. None of the antiplatelets showed any associated survival benefits in controls. Current practice recommends the administration of aspirin as first-line treatment for TIA at presentation but recommends modified-release dipyridamole with aspirin (dual therapy) for secondary management NICE (2021). The guidelines are based on the evidence from two randomized control trials, CHANCE (Y. Wang et al., 2015) and FASTER (Kennedy et al., 2007) comparing dual therapy of clopidogrel and aspirin

for 90 days in patients diagnosed with TIA or minor stroke.
It is worth noting that no trials have ever been done to assess the effect of clopidogrel monotherapy in TIA patients on long-term treatment. CHANCE (Y. Wang et al., 2015) found no statistically significant reduction in mortality with clopidogrel and aspirin therapy at 90 days ($n = 15,170$) based in China and no statistically significant difference in bleeding but the team found a decrease in risk of new stroke events. FASTER trial (Kennedy et al., 2007) with $n = 392$ patients found an increase in bleeding for dual therapy compared to aspirin. However, these two trials continued for less than longer than 90 days and have limited applicability to the long-term treatment. Furthermore, the CHANCE trial is based on the Chinese population and cannot be generalized to the mostly Caucasian population in the UK. Our findings hence lend support to the use of aspirin for the long-term management of TIA compared to other options of antiplatelets or dual therapy. Our study is large, has a long follow-up, reflects routine care in the UK, and also is the first in the UK to assess the relative efficacy of aspirin and other options of antiplatelets. The comparison of the efficacy of antiplatelets remains open for discussion in the stroke literature due to a dearth of data. Elderly patients and patients with high severity of vascular comorbidities are often excluded from trials; at least 75% stroke patients are excluded from real-world practice in randomised controlled trials (Vidyanti et al., 2019). With our simple and pragmatic study design with few exclusion criteria, probably a large number of patients with TIA were eligible for the study. Our study hence can be treated as real-world evidence highlighting the comparative advantage of prescribing aspirin and recommends the latter for secondary care management due to its efficacy in decreasing long-term hazards of mortality in lieu of the modified dipyridamole currently recommended by NICE (2021). This recommendation could present a change to the current guidelines. The choice of clopidogrel or dual therapy of aspirin plus modified-release dipyridamole as secondary management for TIA and ischaemic stroke is based on inadequate evidence. This new evidence raises the need to re-assess the therapeutic options in stroke care management and could be of value to GPs for deciding on prescriptions for better longer-term outcomes.

5.9 Conclusions

After adjusting for a large number of variables including medical conditions, lifestyle factors, socio-demographic factors, and pharmaceutical prescriptions, patients with a prior TIA and no prior stroke had a significantly higher risk of excess risk of death than matched controls. The hazards of mortality associated with TIA ranged from 1.5 to 3. For instance, a case diagnosed at $39-60$ years had HR = $3.09(2.91-3.18)$ compared a same-aged matched control.

The study shows outcomes after TIA are concerning with a high risk of long-term mortality and short-term risk of recurrence of stroke, 17.5% of TIA patients having a recurrent TIA and 3.6% developing a stroke within 1 year. Hence, TIA merits the same medical attention as stroke. Our findings demonstrate the relatively high long-term hazards attributable to TIA and highlight the

need for long-term risk management. The recurrence and mortality risks increase the toll of the economic, societal, and financial burden.
The survival prospects of TIA could be improved by antiplatelet prescriptions. Our study provides important insights into the mortality-reducing effect of aspirin compared to other antiplatelet options, for instance at 39-60 years at entry, TIA patients on aspirin had HR = $0.9(0.82-0.98)$ at 10 years and $0.88(0.8-0.96)$ at 15 years, respectively. The effects of other antiplatelets were insignificant. Further studies should evaluate the impact of different antiplatelet medications with regard to bleeding risks and recurrence of stroke events.
The study demonstrated that there was social deprivation survival difference in both cases and controls. However, we have not found any interaction between social deprivation and stroke, hence there were no barriers to care for TIA patients.

There was significant variation between GP practices.

In the next chapter, we explore the survival outcomes after the ischemic stroke, which is more severe in nature than TIA.

6 Survival analysis after ischaemic stroke

The earlier chapter explored the survival analysis after the first event of a transient ischaemic attack, known as the “mild" neurological condition but which according to our findings, conferred a considerable risk of mortality. This present chapter focused on the long-term survival model after a first ischaemic stroke event in England, the major and most common type of stroke. The data extraction was earlier described in Chapter 3. Firstly, a description of the medical history of the patients is provided. The procedures of the multiple imputations of the missing records are then described. The results of the survival models on the full-case and imputed models are illustrated through forest plots for time-invariant hazard ratios and hazard curves over time for time-varying covariates. The results are discussed and compared to existing research findings.

6.1 Study Design

The subsection describes the information used for the development of the survival models used to estimate the hazards of mortality associated with IS. The primary outcome of the model was all-cause mortality and the primary exposure was ischaemic stroke.

Patients with a first-ever ischaemic stroke event were extracted from THIN database. The cases were matched to three controls on age, gender and GP practice. At entry, patients’ prior medical conditions, prior medical therapies and lifestyle factors were obtained. The survival status of the patients was also recorded. We could not get information on the severity, sub-types and localisation of the ischaemic stroke because it was unavailable in the THIN database. We could not acquire data about psychological factors, living alone status, diet, adherence to medical therapy, and family history of heart disease, hypercholesterolaemia, hypertension or stroke either due to unavailability or due to low recording rates in the primary care database.

Missing values were imputed using the Multiple Imputation technique as explained in Chapter 5 where the data was assumed MAR. The extracted data were treated as multi-level whereby GP practices were considered as clusters of registered patients in the model. Patients within each cluster were assumed to be homogeneous. A random effect for GP practice was considered in the imputation model data which used a multi-level structure. Imputations were done using the **JOMO** package by Quartagno et al. (2019) in **R** using a joint modelling approach. The Monte Carlo Markov Chain

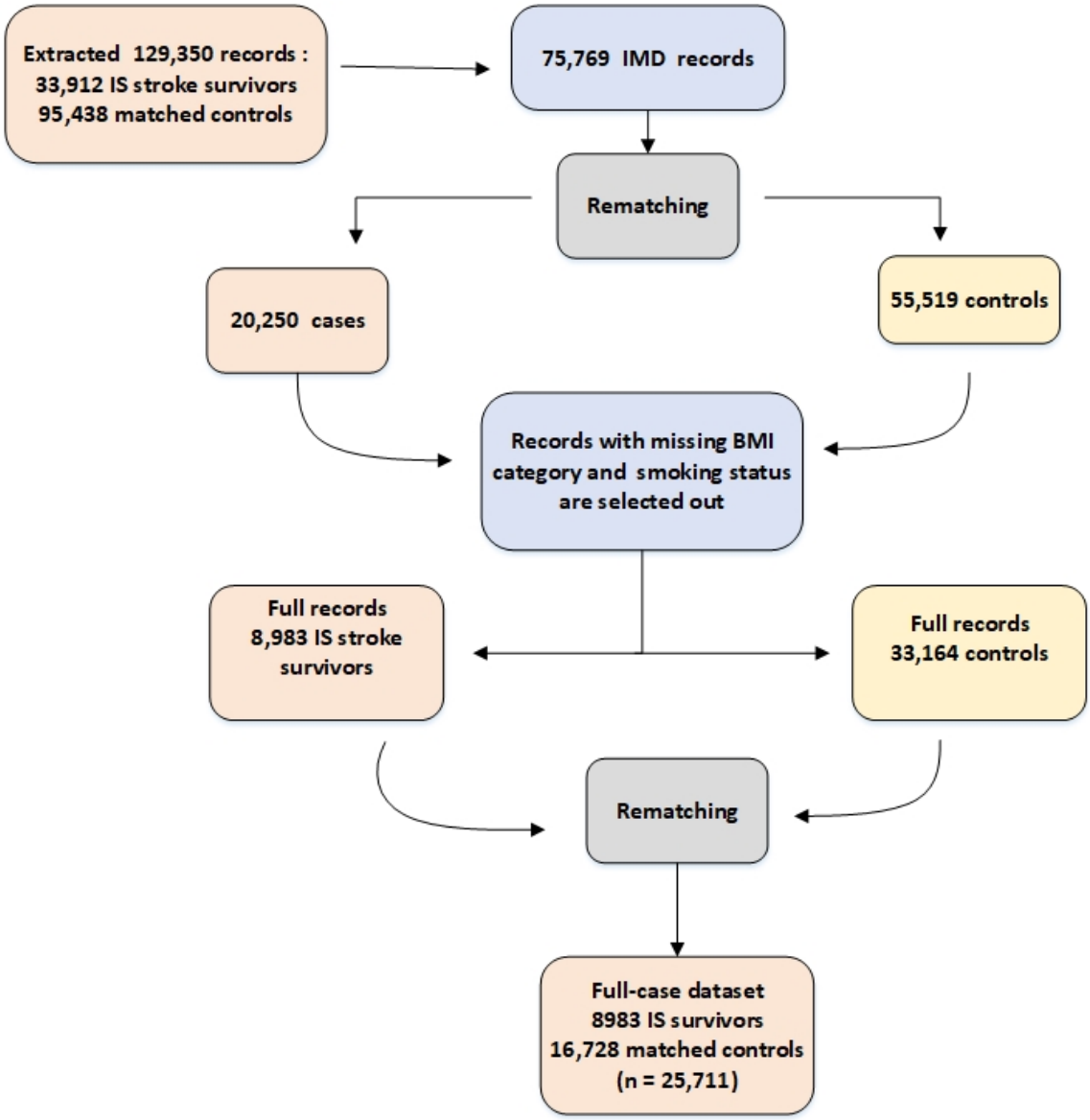


FIGURE 6.1: Data extraction in the IS dataset.

TABLE 6.1: The distribution of variables with missing values in the IS dataset before imputation procedures.

Variables	Levels	Number (%)[1]
Smoking	Non-smoker	27631 (36.5%)
	Ex-smoker	19097 (25.2%)
	Current Smoker	7520 (9.9%)
	Missing	21521 (28.4%)
BMI	Underweight	1365 (1.8%)
	Normal	15565 (20.5%)
	Overweight	16685 (22%)
	Obese	10061 (13.3%)
	Missing	32093 (42.4%)

(MCMC) method of imputation had a burn-in length of 2000 iterations and 100 iterations were run for each imputed dataset. Ten imputed datasets were obtained.

IMD in Quintiles was used as a factor of socioeconomic status. IMD values are available for patients living in England. 129,350 records of patients were extracted in total. They consisted of 33,912 first-ever IS stroke survivors where IS happened between the years 1986 to 2017 matched to 95,438 controls. An entry requirement for the cohort was a complete record of IMD deprivation. The data selection process to attain the complete case dataset (dataset without missing values) is depicted in the flowchart in Figure 6.1. To ensure a balanced subset, 8063 cases were matched to at least one control. The final full-case data consisted of 8983 cases and 16,728 controls. Statistical analysis was performed in three stages. In the first stage, a full-case analysis was performed to obtain a parsimonious model. The model was then used to inform the imputation regression model where the same predictors were used. Finally, the imputed datasets were analyzed and the results were combined using Rubin's rules.

[1]The proportions are determined out of a total of 75,769 records.

6.2 Univariate analysis of predictors

TABLE 6.2: Univariate analysis of predictors of the IS cohort

Study variable	Levels	Coefficient (Beta)	Exp(Beta)	SE	Z statistics	P Value
Birth cohort	1900 to 1920					
	1921 to 1930	-0.53	0.59	0.03	-15.98	0.00
	1931 to 1940	-1.34	0.26	0.04	-35.63	0.00
	1941 to 1960	-2.18	0.11	0.05	-41.43	0.00
Gender	Female					
	Male	0.11	1.12	0.03	4.41	0.00
Age	39-60					
	61-70	0.97	2.64	0.05	19.09	0.00
	71-76	1.63	5.12	0.05	31.9	0.00
	77+	2.38	10.8	0.05	47.53	0.00
Groups	cases					
	controls	0.88	2.4	0.03	34.8	0.00
IMD quintile	1					
	2	-0.13	0.88	0.04	-3.15	0.00
	3	-0.13	0.88	0.04	-3.13	0.00
	4	-0.33	0.72	0.04	-7.95	0.00
	5	-0.3	0.74	0.04	-7.19	0.00
BMI category	Normal					
	Obese +Overweight	-0.39	0.67	0.03	-15.61	0.00
Asthma	Present	0.26	1.29	0.04	6.39	0.00
COPD	Present	1.12	3.06	0.05	24.53	0.00
CKD	Present	0.87	2.38	0.06	14.86	0.00
Myocardial Infarction	Present	0.83	2.3	0.04	21.58	0.00
PAD	Present	0.48	1.62	0.03	16.2	0.00
Smoking	Non -smoker					
	Current smoker	0.3	1.34	0.04	8.32	0.00
	Ex-smoker	0.36	1.44	0.03	13.11	0.00
Atrial Fibrillation	Present	1.03	2.79	0.04	26	0.00
Diabetes	No					
	Yes and Treated	0.81	2.25	0.03	23.13	0.00
	Yes and Untreated	0.64	1.9	0.06	10.06	0.00
Anticoagulants	Yes	0.86	2.36	0.04	19.53	0.00
Hypertension	Yes	0.85	2.34	0.03	31.21	0.00
Antiplatelets	None					
	Aspirin	0.82	2.28	0.03	29.8	0.00
	Dual therapy	1.05	2.87	0.05	23.36	0.00
	Other agent	0.96	2.62	0.07	14.39	0.00

Table 6.2 shows the model-based estimates univariate analysis of the important predictors. All the predictors were significant ($p < 0.00$). The selected prognostic factors were then used for the multiple Cox regression.

6.3 Model Development and Diagnostic tests

A Cox proportional hazards regression model was fitted to estimate the effect of the first-ever IS and other factors on the hazard of all-cause mortality. The variables used to build the model are presented in Table A.2. The initial model included all the main effects and the second-order interactions of all variables. A random effect of the GP practice was adjusted in the model to account for the correlation between patients within a GP practice. Starting with a full model, the model was gradually reduced using a backward-elimination method. The covariates and the factors were removed based on the removal significance level of 0.05 for main effects and 0.01 for interactions using. The hazards of proportionality test were tested using the Grambsch and Therneau (1994)'s cox.zph procedure which is implemented in **R** survival package. The global test was highly significant with global p-value = 1.6e-14 (See Table C.1), providing evidence of non-proportionality.

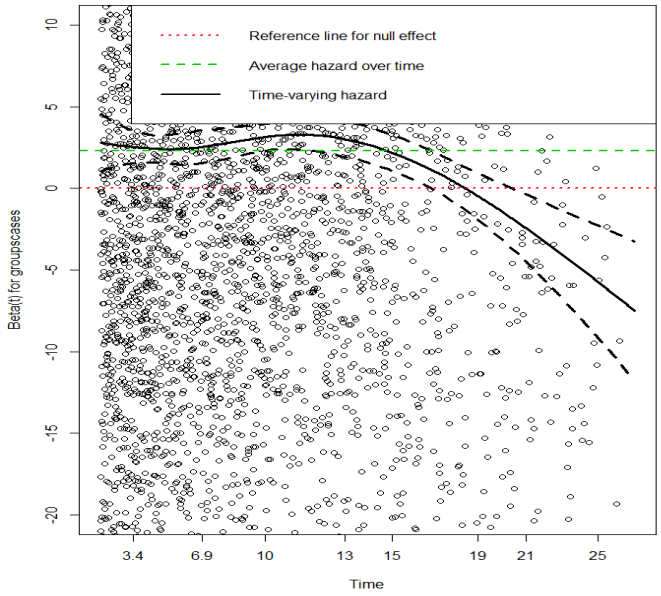


FIGURE 6.2: Schoenfeld residuals based on variable stroke diagnosis.

A general downward trend of the time-varying hazard. The hazards are constant from time $0-7$ years, increase from $7-13$ years, reach a peak at 13 years, and decrease after 13 years.

The scaled Schoenfeld residuals also showed the dependence of some variables on time. Violations to proportional hazard assumption were detected for birth cohort, age category, case-control status (See Figure 6.2) and antiplatelet effects. As previously described in Chapter 4, tests to confirm which variables have time-dependent effects were performed in the **timereg** package in **R**. The variables birth cohort, antiplatelet medications, and the case-control status exhibited time-varying effects and were assumed to have shape effects and additionally to scale effects considered for all

covariates.

6.4 Double-Cox Weibull model

The baseline hazards of the semi-parametric model were compared relative to different concurrent parametric distributions (exponential, logistic, log-logistic, Gompertz and log-normal distributions) across case-controls groups. The Weibull distribution fitted the baseline hazard reasonably well with the lowest AIC (AIC = 18,519 in cases and AIC = 20,218 in controls) as observed in Figure 6.3. Kindly note that the results were based on a model estimated with all individuals but only the baseline hazard form given.

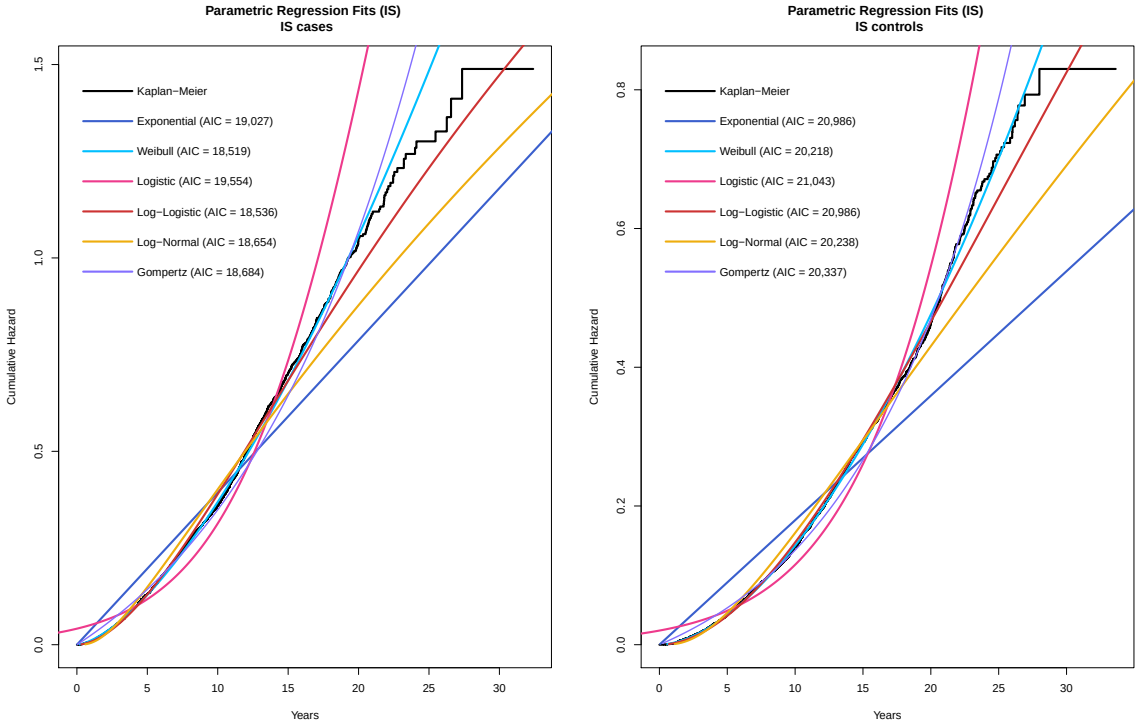


FIGURE 6.3: Graphical comparison of the cumulative hazards of several survival distributions with respective AIC:
Exponential, Weibull, Logistic, log-logistic, Log-Normal and Gompertz compared to KM cummulative hazards function.

The goodness-of-fit was checked by comparing the semi-parametric estimates of the baseline cummulative hazard functions to the baseline cummulative hazards from the Weibull baseline hazards, separately within some strata. The results are presented in Figure 6.4 for the Weibull baseline hazard across age groups, case-control status, and gender. The Weibull distribution describes the mortality really well up to at least 15-20 years. The difference (residuals) between the semi-parametric cummulative baseline and the fitted parametric were plotted. Deviations only were observed after 15 years in Figure 6.5. Overall, the Weibull distribution fitted the survival data pretty well.

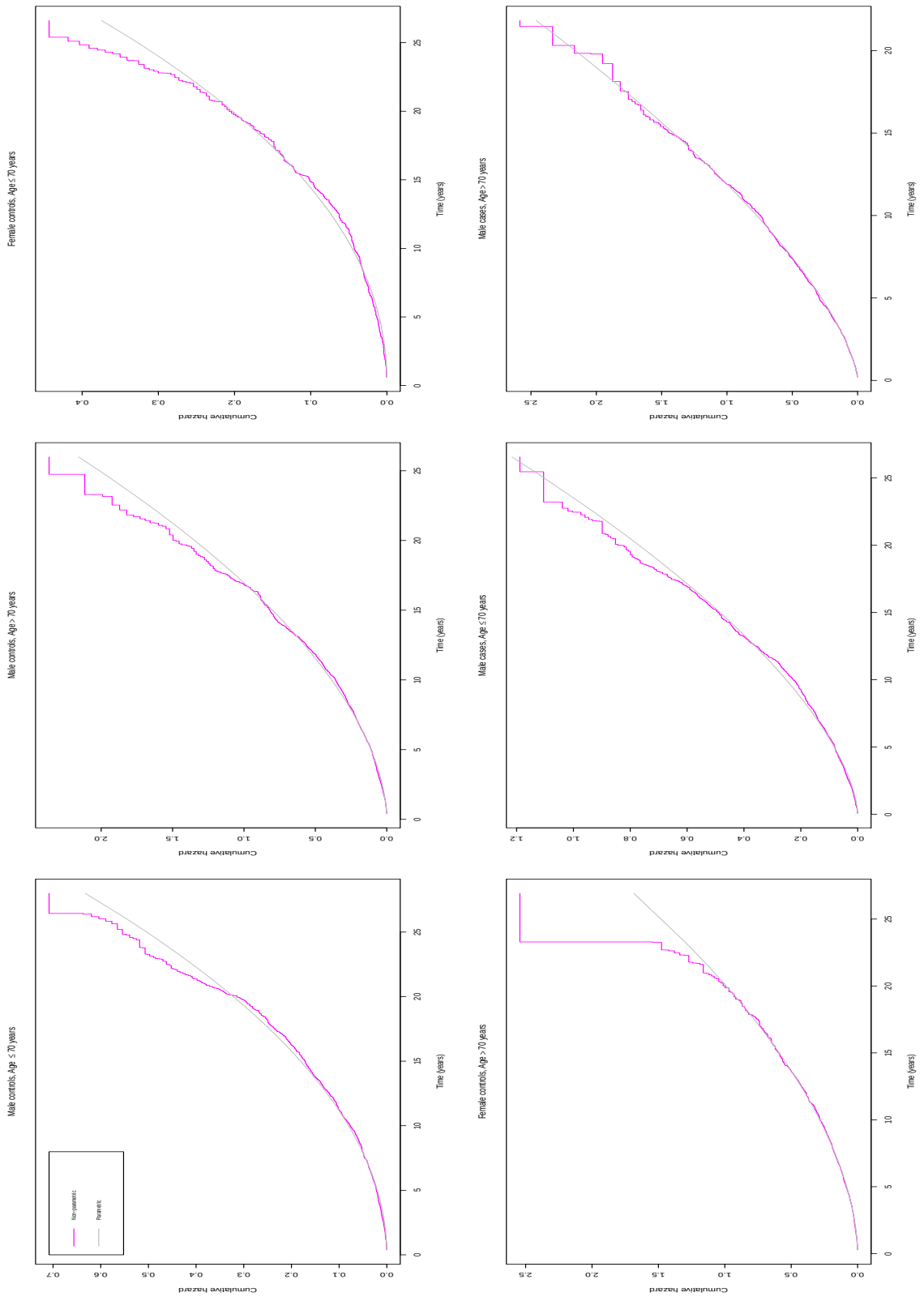


FIGURE 6.4: Comparison of the baseline cumulative hazard functions estimated using semi-parametric (magenta) and parametric (Weibull model, grey) methods. Age, sex and case-controls groups.

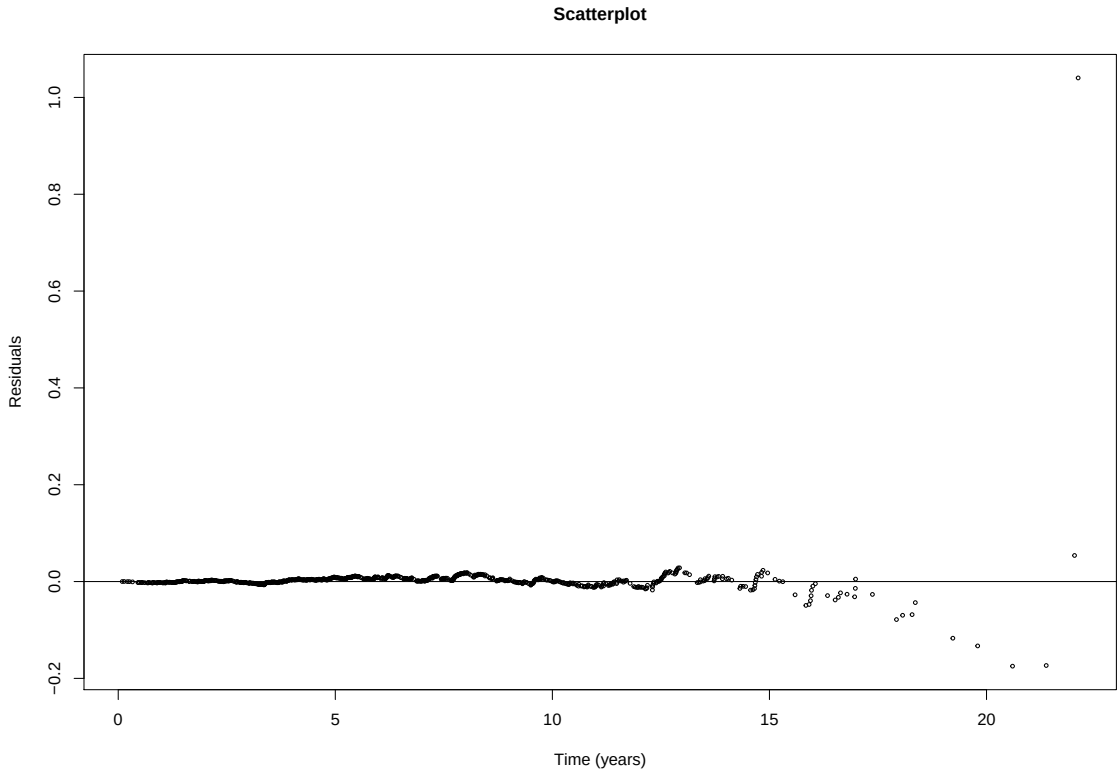


FIGURE 6.5: Plot of the residuals between the parametric and semi-parametric estimates of the baseline hazards pooled across the strata.

Hence, a Weibull Double-Cox model was fitted to the data, with the variables exhibiting time-varying effects set to have both shape effects and scale effects and time-invariant variables set to have only scale effects. To get a parsimonious model, shape and scale effects below 0.05 levels of significance were retained.

6.5 Overview of full-case of survival after ischaemic stroke

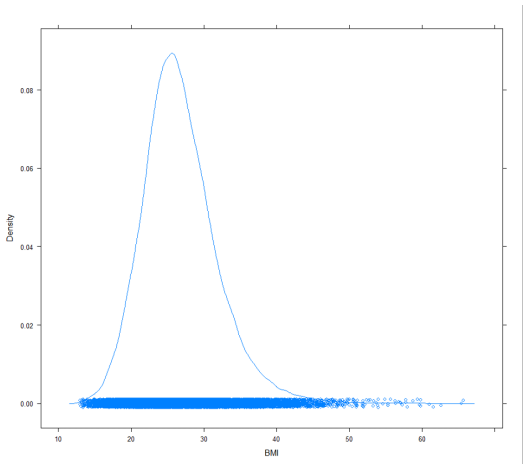
The model is presented in Table 6.4. Older birth cohorts have raised hazards of all-cause mortality suggesting effects of medical advances and improvement in healthcare. Prescriptions of antiplatelets had significant shape effects ranging from 0.90 to 0.84 compared to non-users, which was suggestive of a long-term protective effect. Some prior medical conditions were associated with poor survival prospects; diabetic patients on anti-diabetic medications had HR = 1.79 (1.68 – 1.92), COPD with HR = 1.96 (1.78 – 2.14), heart failure with HR = 1.59 (1.4 – 1.75) to quote a few. Prior prescriptions such as anticoagulants were associated with a risk of HR = 1.18 (1.06 – 1.34) compared to non-users. On the other hand, antiplatelets seem to be protective of survival. Smoking was associated with poor survival HR = 1.84(1.70 – 1.98). The diagnosis of ischaemic stroke has significant interactions with age category at diagnosis, gender, diagnosis of hypertension and antiplatelet prescription. The

model had significant frailty effect of GP practice with $\sigma^2 = 0.06(0.04 - 0.08)$. The model also had a reasonable predictive ability with a concordance of 0.75.

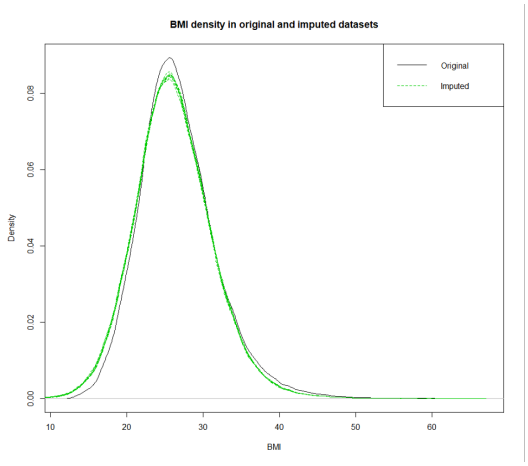
6.6 Multiple Imputation Procedures

The missing data mechanisms were investigated for missingness. The association of missing patterns is investigated in Figure B.1 showed the association of missing variables and the pattern of missingness with observed variables. The pattern confirmed MAR mechanism of missingness. The MAR mechanism was explained in more detail in Chapter 4.

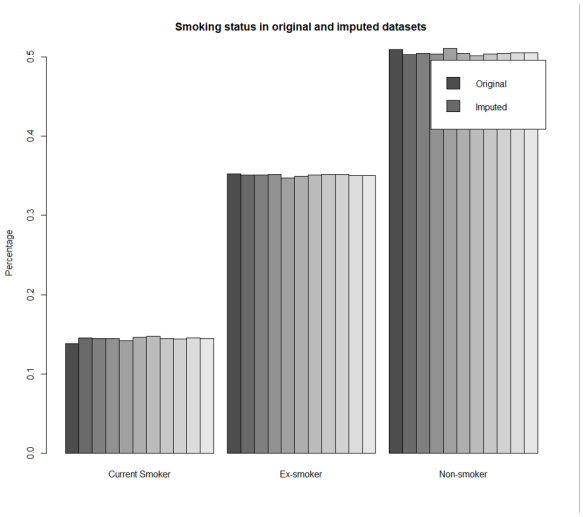
Assuming data to be MAR, the missing data was imputed using multiple imputation technique on BMI and smoking status. The multiple imputation model was built using previous predictors and models in Section 1.3. Using the **JOMO** package in **R**, 10 imputation datasets were generated using a burn-in step of 2000 and 100 in-between iterations. Figure 6.6(a) shows the density plot of the BMI distribution with the blue dots representing missing values prior to multiple imputation, Figure 6.6(b) shows the densities of the BMI of the full-case dataset and after multiple imputation and Figure 6.6(c) shows the distribution of smoking status categories. Multiple imputation preserved the distribution for the variables with missing data. The model in Table 6.4 shows the final results after the Weibull Double-Cox model was fitted to the 10 imputed datasets and the parameter estimates were pooled using Rubin's rules to derive the confidence intervals and associated p-values. The full-case model and the pooled multiple imputed models are compared in Table C.2. The parameters of the full-case model were somewhat lower but not very different from those of the multiple imputed models and their confidence intervals overlapped. This implies that the exclusion of some patients may have underestimated some risks of all-cause mortality.



(A) BMI distribution and missing distribution.



(B) BMI distribution with full-case values (black) and the distribution after multiple imputation (green).



(C) Bar chart for smoking status distribution across categories for the original dataset compared to the distribution of the other 10 imputed datasets.

FIGURE 6.6: Comparison of distribution of imputed data and full case data distribution in IS data.

TABLE 6.3: Baseline characteristics of IS dataset.

Variable	**Type of patients**	
	Cases ($n = 20,250$)	**Controls** ($n = 55,519$)
Death during FU (%)	8898(44%)	15061(27%)
Male(%)	10, 008 (49%)	27470(49%)
Age (mean(sd))	73.8(10.7)	73.2(10.6)
Birth cohort		
1901-1920	10844(20%)	10844(20%)
1921-1930	6627(33%)	18093(33%)
1931-1940	4942(24%)	14232(26%)
1941-1960	4264(21%)	12350(22%)
Year of Stroke diagnosis/ Year of Entry		
1986-1992	452(2%)	1226(2%)
1993-1999	3830(19%)	10340(19%)
2000-2006	7931(39%)	21816(39%)
2007-2016	8007(40%)	22137(40%)
IMD Quintile		
1	3032(15%)	6901(12%)
2	3925(19%)	10067(18%)
3	4414(22%)	11963(22%)
4	4542(22%)	13369(24%)
5	4337(21%)	13219(24%)
Pre-morbid conditions		
Asthma	1968(10%)	5020(9%)
Atrial fibrillation	2997(15%)	3201(6%)
Diabetes Type II	3357(17%)	5188(9%)
Coronary Heart Disease	4497(22%)	7866(14%)
Chronic Kidney Disease	1625(8%)	3587(6%)
COPD	1284(6%)	2559(5%)
Heart Failure	1694(8%)	2730(5%)
Hypertension	9951(49%)	20237(36%)
Hypercholesterolaemia	1341(7%)	3233(6%)
Myocardial Infarction	2064(10%)	3124(6%)
Peripherical vascular Disease	5243(26%)	11387(21%)
Pre-morbid prescriptions		
Anticoagulants	1923(9%)	3304(6%)

Table 6.3 *Continued from previous page*

Variable	Type of patients	
	Cases ($n = 20,250$)	**Controls** ($n = 55,519$)
Antiplatelets	9076(45%)	14078(25%)
Lipid lowering Drugs	6252(31%)	12813(23%)
Antidiabetic Drugs	2724(13%)	4159(7%)
Antihypertensives	14332(71%)	30504(55%)
Lifestyle factors		
BMI category		
Normal	3309(16%)	12256(22%)
Underweight	346(2%)	1019(2%)
Overweight	3617(18%)	13068(24%)
Obese	2372(12%)	7689(14%)
Missing	10606(52%)	21487(39%)
Smoking status		
Non-smoker	5981(30%)	21650(39%)
Ex-smoker	4941(24%)	14156(25%)
Current smoker	2118(10%)	5402(10%)
Missing	7210(36%)	1431(3%)

Table 6.3 shows the baseline characteristics of the study cohort before imputation. The average age at a first stroke was 73.8 years. 44% of stroke patients died during follow-up as opposed to 27% of the controls. The most commonly coded stroke risk factor was hypertension, recorded in 49% of patients. In addition, a considerably higher proportion of the stroke patients had the prior peripheral vascular disease (26% vs. 21% for controls), diabetes mellitus type II (17% vs. 9%) and coronary heart disease (22% vs. 14%). Compared to participants without stroke, cases had significantly higher levels of hypertension, hypercholesterolemia, MI, heart disease, diabetes, and atrial fibrillation ($p < 0.05$.). They were also more medicated than their counterparts. Antihypertensive and antiplatelet medications were commonly prescribed prior to IS stroke. The dataset had missing records on smoking status and BMI category which were imputed later using multiple imputation method.

6.7 Results of the survival models on the data from imputed dataset

The final survival model findings were presented in hazard curves and hazard ratio curves. Table 6.4 shows the final model with significant shape and scale effects.

TABLE 6.4: Final model on pooled data from imputed models: Description, parameter estimates and confidence intervals for the Double -Cox Weibull distribution model with frailty terms.

Variables	**Values/Levels**	**Estimates**	**CI**
Sample size		75,769	
Number of non-censored		23,959	
Weibull Parameters	$a(Scale)$	48.868	43.52−54.87
	$b(Shape)$	1.228	1.2−1.26
Shape parameters			
Birth Cohort	1900 to 1920	1	1
	1921 to 1930	1.015	0.99−1.04
	1931 to 1940	0.953	0.92−0.99
	1941 to 1960	0.768	0.73−0.81
Antiplatelet prescriptions	None	1	1
	Aspirin only	0.945	0.92−0.97
	Dual therapy (Aspirin)	0.982	0.94−1.03
	Other Antiplatelets	0.89	0.84−0.94
Hypertension	No	1	1
	Yes	0.834	0.82−0.85
Scale parameter			
Birth Cohort	1900 to 1920	1	1
	1921 to 1930	0.673	0.64−0.71
	1931 to 1940	0.348	0.32−0.38
	1941 to 1960	0.146	0.13−0.16
Sex	Female	1	1
	Male	1.222	1.18−1.27
IMD Quintile	1 (Most Deprived)	1	1
	2	0.974	0.93−1.02
	3	0.938	0.9−0.98
	4	0.876	0.84−0.92
	5 (Least Deprived)	0.843	0.8−0.89
Case/Control	Controls	1	1
	Cases	5.667	4.93−6.51
Age Category at diagnosis	39−60		
	61−70	1.861	1.65−2.1
	71−76	2.561	2.26−2.9
	77+	4.05	3.57−4.6

Table 6.4 Continued

Variables	Values/Levels	Estimates	CI
BMI	Normal + Underweight	1	1
	Obese + Overweight	0.845	0.82−0.87
COPD		1.619	1.54−1.7
CKD stage 3-5		1.145	1.08−1.22
Myocardial Infarction		1.142	1.09−1.2
Peripheral Arterial Disease		1.107	1.07−1.14
Atrial Fibrillation		1.182	1.13−1.24
Smoking	Non-smoker	1	1
	Current smoker	1.946	1.87−2.02
	Ex-smoker	1.282	1.24−1.32
Diabetes status	No diagnosis/No treatment	1	1
	Yes and Treated	1.354	1.3−1.41
	Yes and Untreated	1.096	1.01−1.19
Heart Failure		1.543	1.48−1.61
Anticoagulant agents		1.14	1.08−1.2
Hypertension		0.955	0.9−1.01
Antiplatelet prescriptions	None	1	1
	Aspirin only	0.966	0.91−1.03
	Dual therapy (Aspirin)	1.28	1.07−1.54
	Other Antiplatelets	0.919	0.76−1.12
Interactions			
	Cases & Aspirin only	0.716	0.67−0.76
	Cases & Dual therapy (Aspirin)	0.525	0.44−0.62
	Cases & Other Antiplatelets	0.661	0.55−0.79
	Cases & Age 61−70	0.668	0.58−0.77
	Cases & Age 71−76	0.533	0.46−0.62
	Cases & Age 77+	0.521	0.45−0.6
	Cases & sex_Male	0.891	0.85−0.94
σ^2	Frailty	0.073	
Concordance		0.81	

To assess the predictive value of our model, Harrell's concordance index between the predicted and the observed survival was calculated. In the model with frailty, the estimate of the concordance was equal to 0.81.
We used graphical methods as well to check for the adequacy of the double-Cox Weibull model. We used different strata to check the fit of the expected survival functions with respect to the empirical survival functions. The graphical results in Figure 6.7 provide evidence of a good fit. The first plot shows that cases fit better than controls for at least 15 years or so. A very good fit in age group 71-76 years is observed. For the older patients(77+ years), there is a good fit for the first 10 years. The departure after 10 years can be explained by the lack of data in later years.

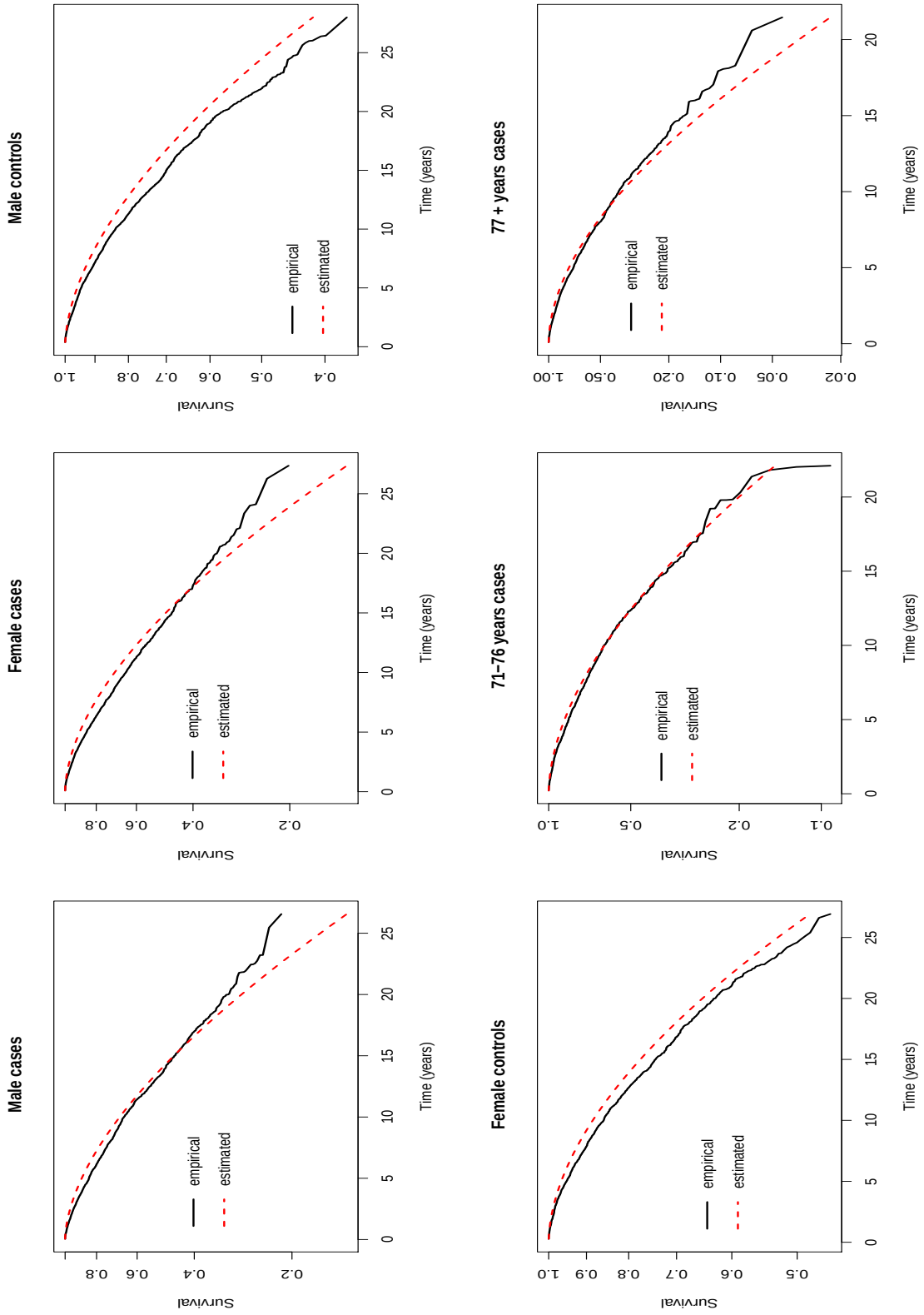


FIGURE 6.7: Comparison of fit of estimated survival to empirical across different strata.

6.7.1 Time-invariant effects of the model (6.4)

Male patients were worse off than females with an HR of all-cause mortality of 1.22 (1.18 – 1.27). There were socio-demographical survival differences; affluent patients from higher Quintiles 4 – 5 corresponding to the less deprivation had HRs of 0.84 – 0.88 relative to the most deprived category 1. The diagnosis of IS stroke was associated to at least 5 times the risk of mortality for IS survivors diagnosed of IS at 39 – 60 years compared to age-matched controls. The adjusted hazard ratios were approximately 3 to 4 times higher for patients diagnosed with an ischaemic stroke at 61 – 70, 71 – 76 and 77+ compared to the same age controls (Table 6.5). The findings on the long-term risk of all-cause mortality attributable to ischaemic stroke are in accordance with extensive stroke literature.

With respect to the factors for Body Mass Index, being overweight and obese conferred a lower risk of mortality with HR = 0.85 (0.82 – 0.87) relative to those of normal and underweight categories. This "obesity paradox" was also observed in the case of TIA survival study. Current smokers had an HR of 1.95 (1.87 – 2.02) and ex-smokers with an HR of 1.28 (1.24 – 1.32) relative to non-smokers.

TABLE 6.5: Adjusted estimated hazard ratios of all-cause mortality by age at diagnosis and case-control status (from the Weibull Double-Cox model on the imputed data.)

Comparison group	Relative to :	HR (95% CI)
Cases aged 39 – 60 years	Controls aged 39 – 60 years	5.67 (4.93 – 6.51)
Cases aged 61 – 70 years	Controls aged 61 – 70 years	3.77 (3.22 – 4.37)
Cases aged 71 – 76 years	Controls aged 71 – 76 years	3.02 (2.60 – 3.52)
Cases aged 77+ years	Controls aged 77+ years	2.95(2.55 – 3.40)

Survival prospects were worsened in patients who had existing cardio-vascular morbidities relative to patients free of those conditions; atrial fibrillation was associated with mortality risk with HR = 1.18(1.13 – 1.24), myocardial infarction with HR of 1.14 (1.09 – 1.2), diabetes (on treatment) with HR of 1.34 (1.3 – 1.41) and peripheral arterial disease with HR of 1.11 (1.07 – 1.14).

6.7.2 Time-varying effects of the model (6.4)

Birth cohorts, types of antiplatelet therapy and diagnosis of hypertension had significant shape effects. The diagnosis of ischaemic stroke had significant interaction effects with antiplatelet therapy, age category at diagnosis and gender. The time-varying effects of the antiplatelet types and different age categories at entry/diagnosis for cases and controls are illustrated in Figure 6.8 and Figure 6.9. For the top panel in Figure 6.8 of patients aged 39 – 60 years at entry, the main findings can be explained as :

- *Aspirin uptake*: There was a clear survival benefit due to the uptake of aspirin in cases relative to cases who were non-users of antiplatelets. The hazard ratios were 0.74 $(0.64 - 0.84)$ and 0.69 $(0.72 - 0.77)$ at 5 and 15 years. However, there was no beneficial effect of aspirin in controls(patients free of stroke). The controls on aspirin therapy had HR of all-cause mortality of 1.03 $(0.93 - 1.15)$ and HR of 0.97 $(0.90 - 1.05)$ at 5 and 10 years, respectively.
- *Dual therapy* : The dual therapy involved the combination of aspirin with dipyridamole and aspirin with clopidogrel. There was a survival benefit associated with dual therapy in some cases. However, the hazard curve had wider 95% confidence bands compared to the aspirin only, reflecting the small number of patients who were prescribed dual therapy. It is noteworthy that the dual therapy is normally issued to select groups of patients, hence reflecting the wider confidence intervals. For controls, on the other hand, dual therapy was associated with an increased hazard of mortality.
- *Other antiplatelets*: Other antiplatelets groups comprised dipyridamole monotherapy or clopidogrel monotherapy. This was associated with modest survival benefits at the start of the therapy, followed by beneficial effects only after around 7 years for stroke survivors.

The effects of aspirin, dual therapy and other antiplatelet agents had quite similar effects for other age categories. Our findings indicate survival benefits related to the uptake of aspirin in cases compared to other antiplatelets. The survival benefits in IS patients on aspirin are similar to that in the study of survival after TIA in Chapter 4. The routine use of aspirin seems to be effective for long-term therapy among ischaemic stroke survivors.

6.8 Life Expectancy model for IS

A life expectancy model using a Weibull Double-Cox model was built to explore the impact of several risk factors, medical conditions and socio-demographic factors on life expectancy. The dataset used for the survival study after ischaemic stroke was also used in this analysis for actuarial translation whose findings are presented in Table C.7. The reason of separate modelling was addressed in Section(3.15). This time, however, age was used on a continuous scale than age categories than the model of survival after IS.

A full case model was first considered and missing records were imputed using the Multiple Imputation method. An initial simple non-parametric Cox's proportional hazard model with all main effects and two-way interactions of all variables was used. The primary outcome was all-cause mortality. Selected variables were based on the literature review and NICE guidelines for the treatment and management of ischemic stroke (NICE, 2020). Diagnoses of medical conditions were defined by the standard list of clinical codes and drug prescriptions corresponding to the British National Formulary for lipid-lowering, anticoagulant, antiplatelet, antidiabetic, and antihypertensive drugs. Further details about levels and coding of covariates are provided in Table 1 (B.3 Appendix).

The covariates consisted of gender, year of birth, case-control status, age at diagnosis, prior comorbidities such as asthma, chronic kidney disease (CKD), chronic obstructive pulmonary disorder (COPD), heart failure, hypothyroidism, myocardial infarction, peripheral vascular disease (PVD), hypertension, hypercholesterolemia, atrial fibrillation, diabetes type II, prescriptions of anticoagulants, antihypertensive, antiplatelet and antidiabetic drugs at baseline. Possible dependencies of the survival outcomes within a general practice were modelled by inclusion of a random frailty term. Backward elimination with $p < 0.05$ for fixed effects and $p < 0.01$ for interaction effects was used to obtain a parsimonious model. After a backward elimination process, the parsimonious model exhibited non-proportionality properties indicating time dependence of some covariates. Before applying the double- Cox –Weibull model, we first compared the semi-parametric estimates of the baseline hazards with different parametric models and the Weibull distribution was chosen. Since the same full-case cohort was used, so the percentage of missing values and the bivariate association of predictors with time were the same as in the full-case survival model in Table 6.2. The dataset had missing records on smoking status, alcohol consumption status and BMI category which were imputed later using Multiple Imputation method. The final model is presented in Table C.8 with the estimates of the scale, shape effects, and their associated 95% CI. The full-case model and the model pooled by Rubin's rules after Multiple Imputation are quite comparable.

The birth cohort, gender, IMD quintile, BMI category, antiplatelets, COPD, heart failure, myocardial infarction, PAD, smoking status, anticoagulants, diabetes, case-control diagnosis, hypertension, age as continuous and the interaction of group with hypertension were retained by the model. Both models have similar parameter estimates and fit the data well, with the concordance of 71.1% and 73.3%, respectively. The frailty variance was highly significant but rather low at 0.07 and 0.08. in

the imputed and the complete case models, respectively

For illustration, some key ages, follow-up time, medical conditions and the uptake of medications were explored. Table 6.6 shows the estimated life expectancy in years for female and male patients. Multiple medical conditions are more prevalent due to ageing population. Using the IS dataset, the prevalence of some selected medical conditions was plotted against age at entry/diagnosis in Figure 6.11. There was a high rise in diagnoses of hypertension and diabetes type II with increasing age. The selected scenarios try to reflect a "realistic" individual risk profile based on typical existing medical conditions.

The study found decrease in life expectancies associated with ischaemic stroke by 6.57, 4.29 and 2.65 years for male patients struck with IS at 50, 60 and 70 years. Similarly, the female patients will lose 7.2, 5.19 and 3.25 years if stroke happens at the same age points.

Being on antiplatelet therapy was associated with minor increase in average longevity of 0.65, 0.17 and 0.09 years in female stroke survivors at 50, 60 and 70 years. On the same note, estimated gains of 1, 0.22, 0.20 years were observed for the male counterparts.

To check for model adequacy, similar to the survival model of the IS cohort, the graph of the empirical and estimated survival graphs were plotted and compared for some strata in Figure 6.10. The cases are good fit for 20 years or so. The controls are better fit. The males (slightly better than the females) are a good fit for the 15 years or so.

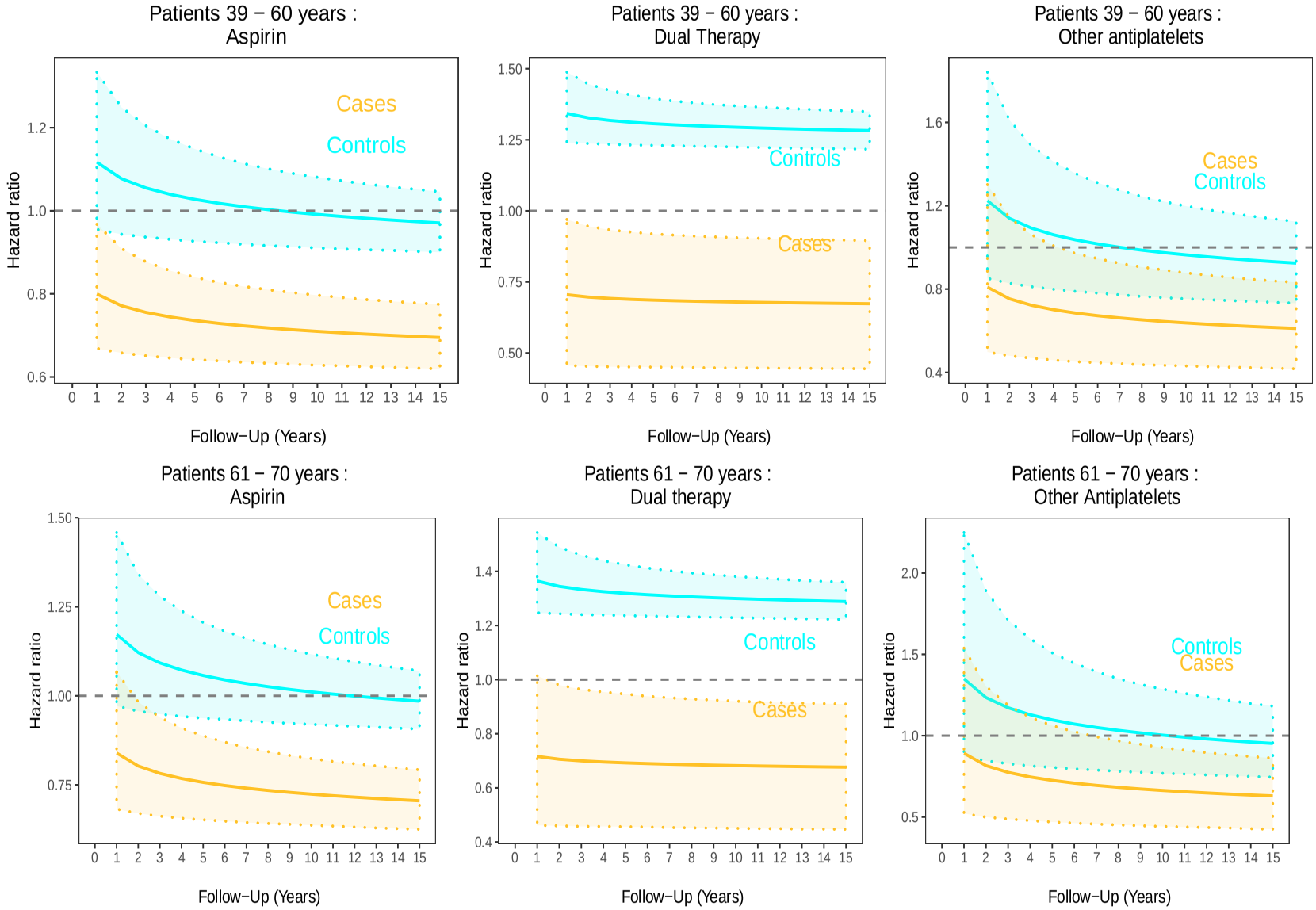


FIGURE 6.8: Hazard ratios of all-cause mortality curves with 95% confidence intervals for patients aged 39−60 and 61 − 70 years at entry treated with different antiplatelets.

Cases with APL prescriptions were compared to cases who were non-users of APL. Controls on APL were compared to controls who were non-users of APL. *APL*: antiplatelet drugs, *DAPT*: Combination of aspirin with other antiplatelet, *aspirin only*: aspirin monotherapy. *Other APLs* includes dipyridamole or/& clopidogrel).

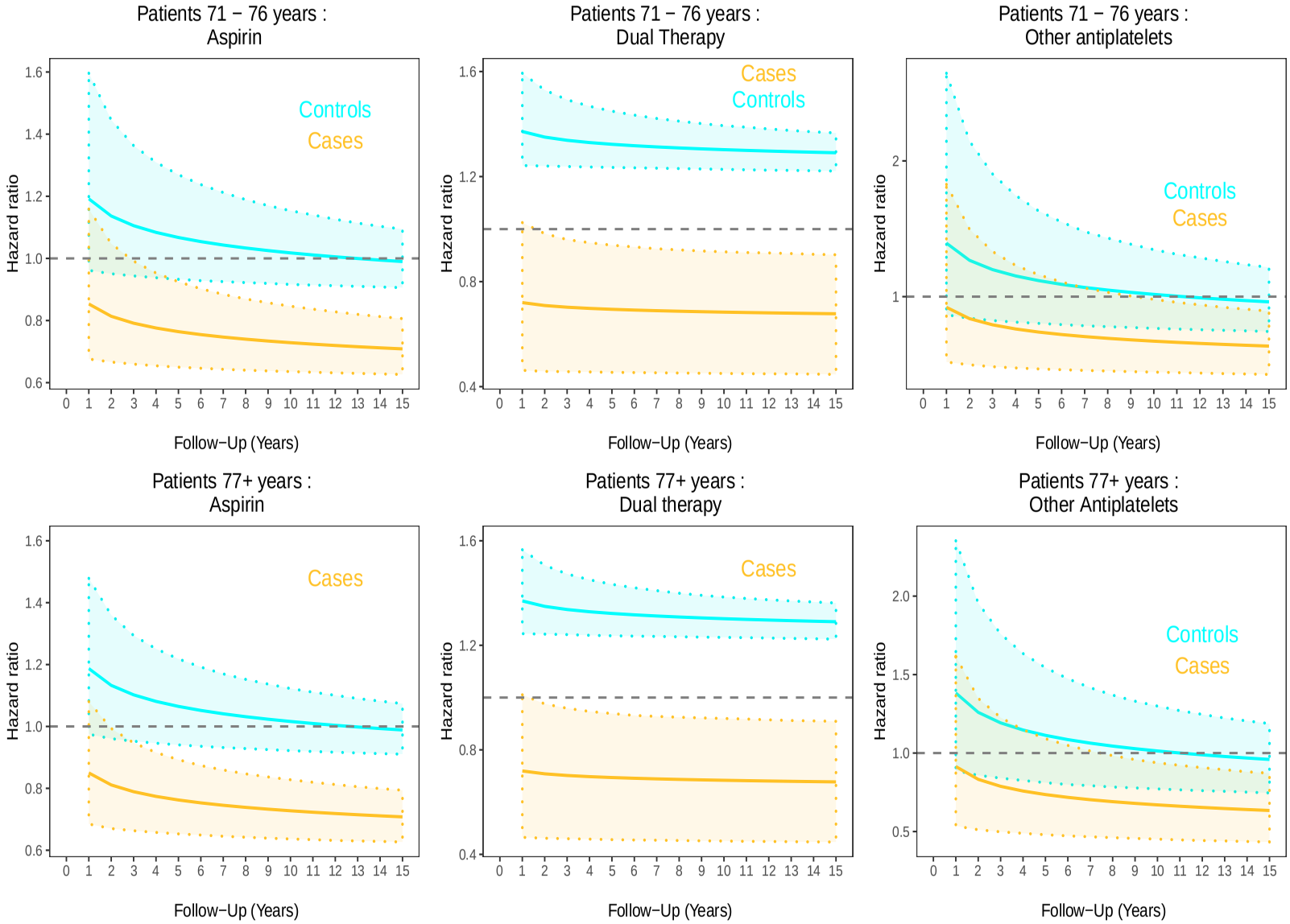


FIGURE 6.9: Hazard ratios of all-cause mortality curves with 95% confidence intervals for patients aged $71-76$ and $77+$ years at entry treated with different antiplatelets.

Cases with APL prescriptions were compared to cases who were non-users of APL. Controls on APL were compared to controls who were non-users of APL. *APL*: antiplatelet drugs, *DAPT*: Combination of aspirin with other antiplatelet, *aspirin only*: aspirin monotherapy. *Other APLs* includes dipyridamole or/& clopidogrel).

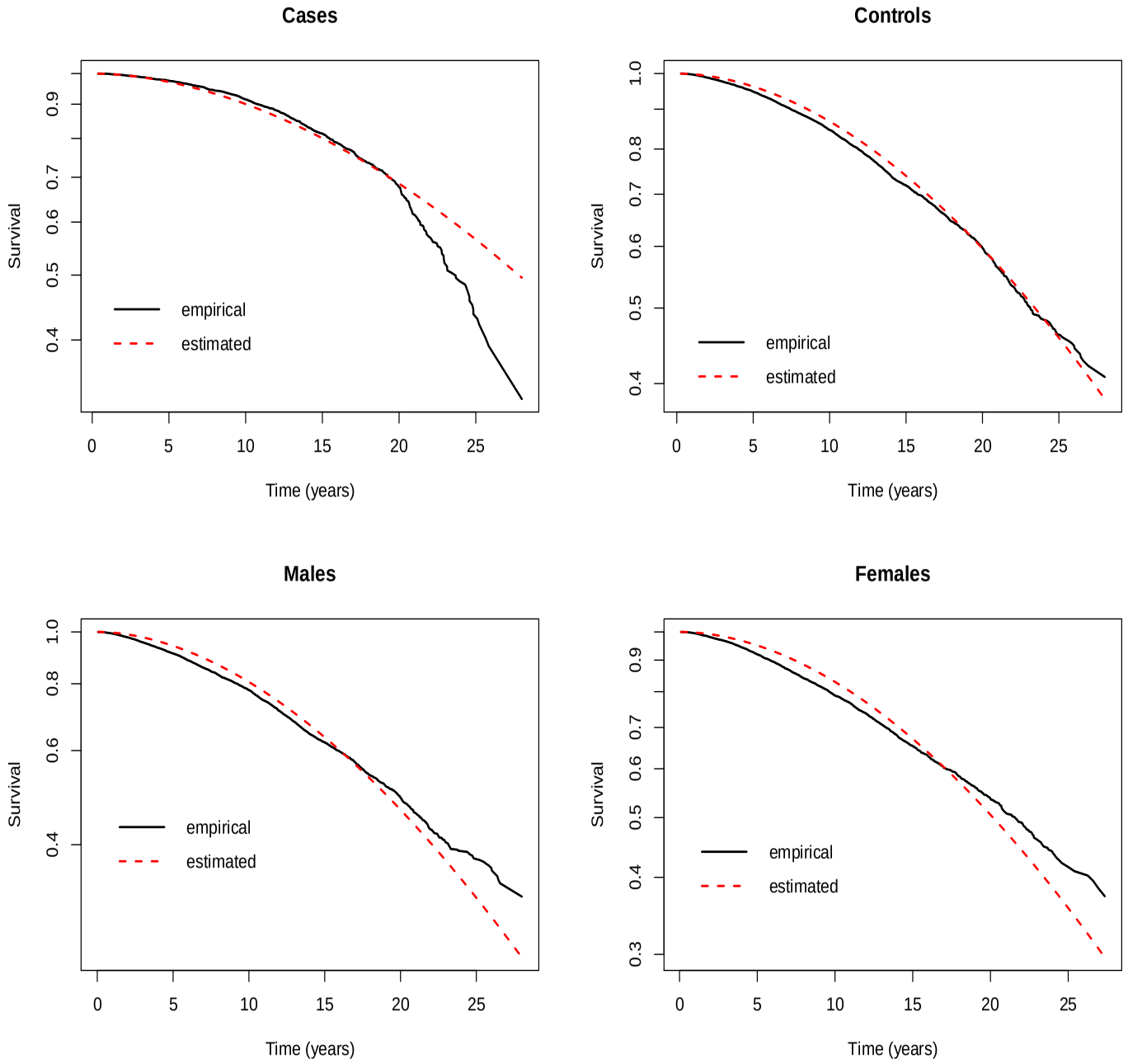


FIGURE 6.10: Comparison of empirical and estimated survival plots for different strata.

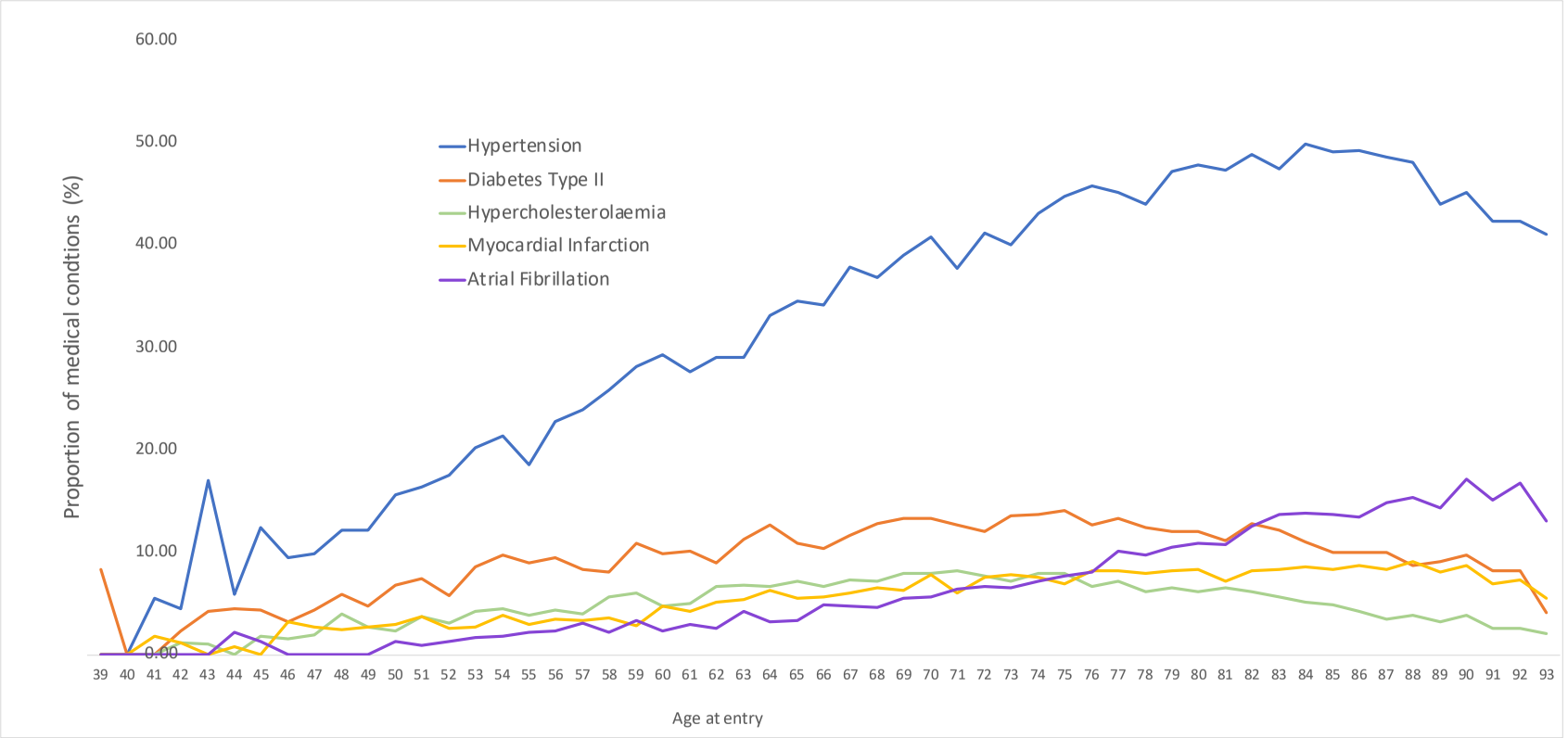


FIGURE 6.11: Prevalence of medical conditions in patients by age of entry.

IS dataset comprising both cases and controls.

TABLE 6.6: Calculation of life expectancies of individuals in the IS dataset at different ages.

Medical Condition	**50**		**60**		**70**	
	Female	Male	Female	Male	Female	Male
Scenario 1	20.19 (17.50 - 22.88)	17.23 (15.34 - 19.42)	13 (11.4 - 14.6)	10.9 (8.79 - 13.01)	8.13 (6.81 - 9.45)	6.81 (4.67 - 8.95)
Scenario 2 : Scenario 1 + IS diagnosis	12.66 (11.32 - 14.00)	10.7 (9.82 - 11.59)	7.81 (6.69 - 8.93)	6.61 (5.28 - 7.94)	4.88 (3.74 - 6.02)	4.16 (2.84 - 5.84)
Scenario 3 : Scenario 2 + Uptake of APL	13.31 (12.0 - 14.66)	11.06 (9.86 - 12.26)	7.98 (7.09 - 8.87)	6.82 (5.6 - 8.04)	4.92 (4.03 - 5.81)	4.40 (3.19 - 5.61)

Note. Scenario 1 implies that the individual has a diagnosis of hypertension and diabetes Type II and is a smoker.
Scenario 2 follows on the Scenario 1 with the individual further experiencing a stroke.
Scenario 3 follows on the Scenario 2 with the individual on antiplatelet therapy.
APL : Antiplatelet therapy.

6.9 Discussion

The current case-control study with 20,250 first-ever ischaemic stroke cases from 1986 to 2016 from selected GP practices in England, had a simple and pragmatic study design with few exclusion criteria. Hence, a maximum number of ischaemic stroke patients were eligible for the study. The findings of the trends in IS and survival modelling provided a detailed picture of the current disease IS burden in the English population.

Short-term survival after a first ischaemic stroke improved considerably from 1986 to 2016. The one-year and 1-month all-cause mortality significantly improved with a fast decline from 2008 to 2015 (- 0.5%) for 1-year fatality. The 1-month mortality's absolute change was - 0.3% for the same period. In post-year 2008, considerable improvements in one-year and 1-month all-cause mortality were observed which reflected advances in survival outcomes due to the improved primary care management of risk factors, better access to stroke centres, acute revascularisation and neuro-imaging, intensive care and secondary prevention and rehabilitation reflecting the initiatives based on the NICE guidelines in 2008 (NICE, 2019). The results are in line with Waziry et al. (2020)'s Rotterdam Study from 1991 to 2015 in the Netherlands, with the PERFECT study in Finland by Meretoja et al. (2011), Edwards et al. (2017b)'s longitudinal cohort registry study (2003-2013) in Canada and Y. Wang et al. (2013) who used the South London register over 1995 to 2010. The authors suggested that stroke unit care, dedicated rehabilitation and improved medical management, improved control of risk factors and counselling had a positive role in contributing to the decline in mortality.

In line with improved primary care management for risk factors, a rise in drug prescriptions such as anti-hypertensives, anticoagulant agents, antiplatelets and statins was observed in our study. This is in accordance with Lee et al. (2011) and Taylor et al. (2011) for trends in statins prescriptions and Falaschetti et al. (2009) for anti hypertensives prescriptions. The increased level of prescribing pharmaceutical therapies prior to first stroke is in accordance with the national guidelines aiming to reduce cardiovascular disease, NICE (2019). Our findings show that antiplatelet prescription prevalence rates increased from 1995 to 2001, and more rapidly from 2001 to 2010, however, dipped after 2010. A decline in antiplatelet prescriptions was also observed in Wales using retrospective analysis of anonymised data (Protty et al., 2018).

Our study estimated the adjusted hazards of all-cause mortality associated with a diagnosis of ischaemic stroke in English residents in routine care. Younger survivors were worse off than older survivors compared to their stroke-free controls. Patients diagnosed of IS at $39-60$ years had HR = $5.67(4.93-6.51)$ whereas cases diagnosed at above 77 years had HR = 2.95 $(2.55-3.40)$. The long-term risk of mortality ranged from 3-fold to 5-fold. As shown by the current and a plethora of other studies, the mortality in stroke survivors is greater than in their matched controls. The main important studies of long-term or short-term mortality following stroke are Rutten-Jacobs et

al. (2013), Carter et al. (2007), Hardie et al. (2003), Dennis et al. (1993), Gresham et al. (1998) and Brønnum-Hansen et al. (2001). Carter et al. (2007) reported a 3-fold risk of death for stroke subjects compared to the age and sex-matched reference population. While Hardie et al. (2003) and Dennis et al. (1993) showed 2-fold overall risk of death compared to the general population. Brønnum-Hansen et al. (2001) reported that stroke patients had a 5-fold increase in the risk of death between 4 weeks to 1 year and a 2-fold increase in risk subsequent to 1 year when compared to their matched controls. Our hazard of mortality is slightly different than previously estimated as the current research made use of primary care data, had a long follow-up duration and was adjusted for a wide range of confounders and had flexible inclusion-exclusion criteria, hence covering maximum patients.

In keeping with the published results, our model results in excess mortality of stroke patients compared to the stroke-free adults of the same age, sex and GP practices. Stroke is a medical emergency associated with a very high risk of death and with a continued excess of death throughout the life of patients. Secondary prevention should hence be a long-term process given the increased risk throughout decades after a stroke event.

The study found that antiplatelets could be very beneficial for ischaemic stroke survivors. Upon testing all the second-order interaction effects, significant interactions of prescriptions of antiplatelets and age category at entry with the diagnosis of ischaemic stroke were found. Aspirin was found to be the most effective treatment for stroke survivors in the long term. However, there was a potential concern for patients free of ischaemic stroke as the prescriptions of aspirin was associated with increased hazards. The benefits associated with dual therapy and other antiplatelets were only apparent after some years or were uncertain. Current NICE guidelines NICE (2020a) recommend the standard clopidogrel daily. Modified release dipyridamole may be used if clopidogrel and aspirin are not tolerated. Aspirin is only used if clopidogrel and dipyridamole are not tolerated. The current guidelines are based on review of two RCTs CHANCE (Y. Wang et al., 2015) and FASTER (Kennedy et al., 2007) with limited evidence with regards to the long-term impact on mortality of the different antiplatelets, with small size and one of them being based on a Chinese population. We hence believe that the restricted ethnic population and patterns of care in that trial limited the generalisability of the results and is debatable. We, therefore, recommend aspirin for secondary care management due to its efficacy in decreased long-term hazards of mortality in lieu of the modified dipyridamole. However, the best therapy should be individualised taking into account underlying comorbidities.

Our life expectancy calculations also showed that diagnosis of ischaemic stroke at 40 to 60 years even after 5 years after IS can contribute to a loss of 10 to 20 years of life expectancy but can be improved by a gain of 3-10 years with the uptake of antiplatelets.

The significant frailty term of the random GP practice effects shows practice variation. It implies

that heterogeneity in survival after stroke is partially due to the differences among GP practices. The identification of risk factors, after accounting for the random GP practices variation, provides useful information on how a patient's survival is affected. These effects could facilitate the comparison of practices performances in ischaemic stroke treatment and rehabilitation after adjustment for patient characteristics and clinical risk factors.

6.10 Conclusions

This large population-based study calculated the adjusted risk of all-cause mortality linked with ischaemic stroke after accounting for a variety of socio-demographic, lifestyle, and concomittant variables. The problem of the time dependence of some variables in the IS survival model was addressed using a Weibull Double-Cox model.

When compared to their matched controls, IS survivors exhibited lower survival chances, with HRs ranging from 3 to 5.

When compared to non-users, IS survivors on aspirin medication had a much lower mortality risk than those on other antiplatelet alternatives. For example, patients diagnosed with IS at $39-60$ years on aspirin therapy had an HR of $0.76(0.65-0.88)$ at 3 years of follow-up. The impact of dual therapy was likewise protective (albeit with a broader confidence interval), whereas it was not significant for other antiplatelets. On the other hand, there were no substantial benefits linked with antiplatelets in the control group. This finding is clinically significant because such prescriptions can improve the chances of survival for IS patients. In the TIA study, the positive effects of aspirin were seen in people of all ages. These findings could lead to changes in current guidelines, such as the recommendation of aspirin for secondary care therapy instead of other antiplatelet choices.

Recent trends in IS stroke suggest an encouraging decrease in the short-term all-cause mortality. This could be likely due to significantly improved vascular risk diagnosis and preventative therapy prescribed prior to and after stroke as proposed by Lee et al. (2011). Our findings back up this improvement in survival after a first-ever IS. Despite these encouraging findings, management appears to be lacking in several areas. Our findings suggest that antiplatelet prescriptions have had a protective impact on survival in the past, but that their use (pre-stroke) has been decreasing since 2010.

7 Conclusions

7.1 Introduction

This thesis pertains to the development of survival models to estimate the hazards of all-cause mortality risk following an ischaemic stroke and TIA event using English primary care data. This chapter will discuss the key findings, implications of findings, contributions to the existing clinical evidence, and the strengths and limitations. Finally, the overall conclusions are presented.

7.2 Main findings

Medical records from 1987 to 2017 from general practices contributing to The Health Improvement Network (THIN) database were used to develop two survival models. The first model was developed to estimate the hazards of all-cause mortality associated with a history of the first episode of transient ischaemic attack and the effects of related treatments while adjusting for socio-demographic, clinical risk and lifestyle factors.
The second model aimed to estimate the long-term survival after a first-ever ischaemic stroke, the most common stroke occurring in $80-85\%$ of cases of stroke in the adult population of England while adjusting for the above risk factors. In both retrospective case-control studies, the cases were matched to 3 controls and the studies have a long follow-up of up to 25 years. We found higher all-cause mortality hazards for TIA patients and IS survivors compared to their controls.

7.3 Survival study after TIA

There is a dearth of research focusing on the long-term survival after a TIA event with most studies either providing only short-term follow-up or focusing on associated morbidity, functional status or the recurrence of strokes in the short term. Thus, it was of particular interest to get an insight into the long-term impact of a TIA event on survival.
The important finding of the study was that a patient diagnosed with TIA had a substantially increased long-term hazard of all-cause mortality. TIA is often considered a 'minor' neurological condition but our research showed that it serves as a predictor for impaired long-term survival. Patients affected at younger ages and with a previous diagnosis of hypertension were worse off.
However, our study estimated lower hazards than some previous studies. The findings are in accordance with Edwards et al. (2017a), Gattellari et al. (2012), Gresham et al. (1998), Hankey (2003), Hardie et al. (2003), and Jacob et al. (2020) but differed from Brønnum-Hansen et al. (2001) and

Eriksson and Olsson (2001). Our survival model adjusted for a large number of risk factors, and had a larger sample size and follow-up period than other studies. We believe that our findings are more precise and reinforce the conclusion that the hazards of all-cause mortality following TIA are quite serious.

Our study also found considerable benefits associated with antiplatelet treatments. Antiplatelet therapy is one of the vital interventions to reduce vascular events or recurrent strokes, following a TIA event. However, our study found that the antiplatelets were hazardous for TIA-free patients (the "healthy" controls). This result agrees with McNeil et al. (2018)'s study who found no substantial difference in death between the aspirin group and the placebo group but noted a higher rate of bleeding in the aspirin group and therefore urged for the limitation of the use of aspirin in healthy individuals.

The current NICE guidelines recommend aspirin plus modified-release dipyridamole as the preferred antiplatelet therapy for secondary prevention following a TIA (NICE, 2021). However, their guidelines were based on two small randomised control trials (CHANCE and FASTER), both had short follow-ups, one had an ethnically different population (CHANCE trial) and were highly selective often excluding older patients.

This research found aspirin to be the most effective option for the long-term management of patients diagnosed with TIA relative to the other antiplatelet options, clear and consistent for all age groups. Our study hence contributes to the existing clinical evidence to show the effectiveness of aspirin in TIA patients and its beneficial effects in the long run. This finding encourages the use of aspirin for secondary intervention in adults with TIA. Our study also shows that aspirin may be causing more harm than good in individuals with no diagnosis of TIA.

Both at baseline and at follow-up, TIA patients were found to have a higher cardio-vascular burden which suggests the need to aggressively manage key risk factors, such as high blood pressure. They can also lower their risk of another stroke by managing behavioural choices, such as smoking cessation.

The findings of the high long-term hazards of all-cause mortality and shortened life expectancy associated with TIA imply that TIA patients need to be managed long-term and the diagnosis of TIA needs to be regarded as a medical emergency similar to a regular stroke.

7.4 Survival analysis after an IS

We considered the trends in the short-term mortality after ischaemic stroke; one month and 1-year all-cause mortality from 1985 to 2016 for ischaemic stroke patients, and found an encouraging decline over time. This reflected the rigourous primary care management of risk factors reflecting the national initiatives to reduce cardiovascular disease (Lee et al., 2011). The observed reduction can also be explained by better access to stroke centres, acute revascularisation and neuro-imaging, intensive care, and secondary prevention and rehabilitation (NICE, 2021). While prescriptions of

anti-hypertensive, anticoagulant, lipid-regulating drugs and anti-diabetic drugs were on the rise reflecting the active management of the IS survivors, the prescriptions of antiplatelets declined from 2008.

The study found a considerably raised hazards of all-cause mortality in ischaemic stroke patients, with younger IS survivors being more severely affected. The findings were in accordance with Rutten-Jacobs et al. (2013), Carter et al. (2007), Hardie et al. (2003), Dennis et al. (1993), Gresham et al. (1998) and Brønnum-Hansen et al. (2001). Carter et al. (2007). The adjusted hazard ratios were approximately in the range of the hazard ratios reported in the stroke literature. Any discrepancy in the hazard ratios can be accounted by the differences between primary care data and hospital and register-based data, greater sample size, longer follow-up and adjustment of a much larger number of risk factors than in published studies. Our modelling adjusted not only for sex and age but also for comorbidities, treatments, lifestyle choices, and socio-demographic factors, resulting in more accurate estimates.

Additionally, the study found survival benefits associated with antiplatelet prescription to ischaemic stroke patients for 15 years following the stroke event. Aspirin prescription was clearly beneficial for all age groups of stroke survivors. Our study hence suggests that younger patients have the most to gain from the aspirin monotherapy given their higher risk of all-cause mortality following a stroke event. Dual therapy (aspirin with dipyridamole or aspirin with clopidogrel) prescriptions were found to be associated with lower survival benefits in cases. Other antiplatelets (dipyridamole monotherapy or clopidogrel monotherapy) prescriptions were also found to be protective in cases, however the benefit was only significant after 7 years after the stroke event. The study found mixed survival prospects associated with the prescription of antiplatelets in controls. Insignificant survival benefit was associated with aspirin monotherapy and other antiplatelet therapy in controls while dual therapy was hazardous for all age groups of stroke-free controls.

This study contributed to the existing clinical evidence by showing the relative efficacy of aspirin for long-term survival relative to clopidogrel recommended by the NICE appraisal TA210 (NICE, 2010) for secondary prevention following IS.

To conclude, after surviving the first event of IS, the survivors remain at significant long-term risk of death, and improving prescriptions of antiplatelets has the potential of reducing long-term mortality.

7.5 Study Strengths

This study used routinely collected primary care data that were representative of the UK (Blak et al., 2011; Hall, 2009; Lewis et al., 2007). The studies had large sample sizes which provided high power in evaluating the difference in hazards of all-cause mortality among different groups. The electronic medical records of the patients from THIN are very rich in information on socio-demographic, clinical diagnoses of medical conditions, referrals, and medical and lifestyle interventions. Hence, our model could incorporate a large number of covariates.

Other studies of stroke based on data sources such as disease registers, hospital data, and mortality registers were mostly constrained by small sample sizes, were highly selective or have rigid inclusion-exclusion criteria, short follow-up period, with findings that are often non-generalisable, and models adjusted by limited risk factors. However, our research using primary care records provided a better way to model stroke survival.
Additionally, electronic medical records are regularly updated and are hence more precise. Both studies had a long follow-up time which allowed us to study the changes in lifestyle factors, medications, and interventions. The long follow-up permitted us to estimate the long-term hazards of death more accurately. This led to better precision in the calculation of the life expectancy which is useful for retirement planning and healthcare resource allocation and medical management.
The case-control design permitted the investigation of the comparative effect of stroke and TIA on survival unlike other designs, which usually lack a control group. This also helped to reduce the selection bias as both cases and controls were selected from the same pool of patients (Cole et al., 2011). Our pragmatic data modelling adjusted for a larger number of risk factors hence leading to better precision of the estimated hazards of mortality after stroke and TIA while also including second-order interactions. Furthermore, the study included variations due to GP practices by including a frailty term in the model.

Most survival models of stroke in the literature used the traditional Cox proportional hazards models often foregoing checking the proportionality assumption which is an underlying premise for Cox's survival models. If the proportionality test fails, the non-proportionality should be included in the model. In our study, we used a flexible parametric survival model, the Weibull Double-Cox model which could consider the time-varying nature of the hazard ratios. Our improved model hence provides better insight into the hazards associated with different risk factors.
Our model findings are important for the healthcare professionals for proper medical management of patients, for the stakeholders for proper resource allocation, for the individuals, caregivers and their doctors in identifying the risk factors that can be hazardous and interventions that can provide survival benefits. Additionally, life expectancy estimates resulting from the models can facilitate the choice of insurance products and retirement planning for the insureds, and product pricing for the insurer and financial providers. Lastly, the results are useful for the government and public health authorities in reviewing the UK pensions system and the national healthcare system to devise proper healthcare management, screening, awareness, and possible changes in guidelines.

7.6 Study Limitations

One of the limitations of the study was the lack of data on stroke severity. Stroke severity has been reported to be a strong predictor of survival in the literature, especially from hospital-based data (Andersen et al., 2005; Chang et al., 2010; Mogensen et al., 2013). We were unable to incorporate stroke severity in our analysis due to the THIN database not containing the stroke severity data. We

were also unable to adjust for stroke sub-types due to incomplete coding.
The risk of misclassification of patients' information remains a possibility. Patients may be recorded but not identified in the right category (Saxena et al., 2007). The record of the medical diagnosis is dependent on the accuracy of the administrative code. However, with QOF 2014 framework (Doran et al., 2014), this risk has been significantly reduced. Practices are remunerated on the results they achieve.

Furthermore, we had no data on the stroke outcomes such as quality of life, functional status and any psycho-social factors such as depression and stress following stroke. Hence, our study was unable to examine the effects of functional and cognitive impairments on survival. Selection bias could have taken place as the very mild or temporary TIAs might not have come to medical attention. It is however likely that these patients constituted a low-risk group.
Missing data was unavoidable in this study. There were records that were incomplete for BMI, smoking status and alcohol status. Multiple Imputation was used to impute the missing data, which might have reduced precision. However, the survival models on complete records and the imputed data estimated similar hazard ratios. The drug adherence was unknown and therefore survival prospects associated with prescriptions of drugs may not accurately reflect the effects of drugs on mortality. Finally, while major confounders of stroke were adjusted for, there could potentially be some residual confounding by a number of other factors such as the familial history of stroke or other vascular conditions, fruit and vegetable intake, and physical activity before a stroke.

7.7 Implications

The first aim of this research was to establish the list of the risk factors associated with the survival and longevity of stroke survivors. The list of variables is presented in Chapter 4 and most of the important variables were extracted from the THIN database.
The second aim was to estimate the survival benefits and hazards associated with survival after stroke and TIA. All medical conditions such as heart failure, PAD, COPD, atrial fibrillation, diabetes type II, and myocardial infarction were associated with higher hazards of all-cause mortality. Patients diagnosed of these conditions were generally worse off than their counterparts free of these conditions. High blood pressure had significant interaction with case-control status in the TIA study. Those important findings show the importance of screening of IS and TIA sufferers for high blood pressure, diabetes, atrial fibrillation in the long term. The timely diagnosis gives the opportunity for treatment. Control of cholesterol and blood pressure levels is important to avoid serious subsequent events.

The third aim was to tackle the problem of missing records. Multiple imputation was performed. The distribution of complete records was very close to that of the imputed ones. Furthermore, the survival models based on complete records estimated similar hazard ratios as the survival models

based on the multiple imputed datasets. The performance statistics, based on the concordance index were also similar.

The fourth aim was to estimate the effects of TIA and IS and associated treatments on all-cause survival. Anticoagulant drugs were found to be hazardous. However, interestingly, antiplatelets were found to be protective for TIA and stroke cases. Aspirin was found to be the best option for the long-term management of TIA and stroke patients as it was associated with reduced long-term all-cause mortality among all age groups. Our study hence recommends the use of aspirin for long-term secondary prevention as opposed to other antiplatelet options recommended by NICE guidelines. Stroke and TIA were associated with higher mortality than controls. TIA is usually considered to be a mild condition. Our study shows that it is a marker for long-term survival. The "warning" events hence serve as an opportunity for prevention, as early studies have underestimated the risk associated with a TIA. Insurance premiums, reserve and benefits calculations need to take account of the added risk due to TIA and stroke.

The fifth aim was to estimate the risk associated with socio-demographic, lifestyle and health factors and any associated interactions to help to inform public health measures. Survival prospects were poor for smokers, diabetic (untreated and treated) patients and for health conditions such as atrial fibrillation, heart failure for both cases and controls. There were no interactions between lifestyle choices and case-control status in both the IS and TIA studies. However, addressing modifiable risk factors can reduce the risk of subsequent cardiovascular and cerebrovascular events for both cases and controls. Equipped with this information, public-health officials should aim to encourage behaviour changes likely to improve the health of middle-aged and older people. The focus of prevention should not be short-term but a lifetime.

The sixth aim was to estimate the life expectancy based on survival after TIA and IS which was achieved by developing new models for life expectancy under different scenarios. The average longevity in years was calculated for stroke and TIA survivors. This can help individuals to make informed decisions on their lifestyle choices and retirement planning.

The seventh aim was to estimate the number of years lost due to medical conditions, treatments, lifestyle and socio-demographic factors and their interactions. We found that the uptake of antiplatelets can prolong life expectancy. Our life expectancy models can be applied for different scenarios and give a good overview of the gain and loss of expected years.

The eighth aim was to evaluate how the model can be used for different parties. Equipped with the information from the models, actuaries, and insurers might use the result to make better decisions in terms of pricing of products and adjustment and creation of products in view of reducing their basis risk.

7.8 General conclusions

The key objectives of this research were to investigate how a history of ischaemic stroke or transient ischemic attack influences survival and which treatments can help individuals live longer using English primary care records.

The research reported the survival prospects associated with a history of a TIA or IS, as well as how secondary medication therapy or concomittant medical conditions could affect those chances. Studies based on the survival after these neurological conditions were quite sparse and the available ones had design constraints, small sample sizes, poor adjustment of factors, shorter follow-up periods or made use of hospital or register data. There was a lack of research with a long-term follow-up period adjusting for a range of confounders. As a result, our large-scale study fills in the gaps in the literature and offers fresh insights on stroke and TIA survival in routine treatment. The findings show that stroke survivors and TIA patients are worse off compared to previously estimated. Furthermore, most Cox survival models in the stroke literature did not report checks for proportionality of hazards assumptions. In this work, we used a novel parametric model to handle the time dependency of any factors that violated the PH assumptions.

Antiplatelets prescribed in routine care were found to improve the survival of TIA and IS patients. Antiplatelet prescription guidelines from NICE are based on two RCTs that were small, had limited generalisability, and were not applicable to the English population. Our study which is the first large population-based in the UK fills in the gaps in clinical evidence with regard to antiplatelet therapy. In the long-term, aspirin is useful in TIA and IS patients, according to our research.
The research adds to the growing body of evidence that TIAs should be taken seriously. Not only are these patients at a higher risk of death than matched controls, but they are also at a higher risk of long-term sequelae such as recurrent stroke and myocardial infarction. Our findings emphasize the importance of bolstering secondary preventive strategies such as medication adherence.

IS survivors were also at a higher risk of death than matched controls. The increased morbidity and mortality risks following these conditions are quite worrying. Stroke is a lingering risk decades after the index occurrence, and strict adoption of measures is required to limit the number of premature deaths and disabilities caused by this debilitating medical condition.

A Readcodes

TABLE A.1: Readcodes for TIA diagnosis

Medcode	Description
G65..00	Transient cerebral ischaemia
G65..11	Drop attack
G65..12	Transient ischaemic attack
G65..13	Vertebro-basilar insufficiency
G650.00	Basilar artery syndrome
G650.11	Insufficiency - basilar artery
G651.00	Vertebral artery syndrome
G651000	Vertebro-basilar artery syndrome
G652.00	Subclavian steal syndrome
G653.00	Carotid artery syndrome hemispheric
G654.00	Multiple and bilateral precerebral artery syndromes
G655.00	Transient global amnesia
G656.00	Vertebrobasilar insufficiency
G657.00	Carotid territory transient ischaemic attack
G65y.00	Other transient cerebral ischaemia
G65z.00	Transient cerebral ischaemia NOS
G65z000	Impending cerebral ischaemia
G65z100	Intermittent cerebral ischaemia
G65zz00	Transient cerebral ischaemia NOS

TABLE A.2: Readcodes for IS diagnosis

Code	Description
G630.00	Basilar artery occlusion
G632.00	Vertebral artery occlusion
G63y000	Cerebral infarct due to thrombosis of precerebral arteries
G63y100	Cerebral infarction due to embolism of precerebral arteries
G64..00	Cerebral arterial occlusion
G64..11	CVA - cerebral artery occlusion
G64..12	Infarction - cerebral
G64..13	Stroke due to cerebral arterial occlusion
G640.00	Cerebral thrombosis
G640000	Cerebral infarction due to thrombosis of cerebral arteries
G641.00	Cerebral embolism
G641.11	Cerebral embolus
G641000	Cerebral infarction due to embolism of cerebral arteries
G64z.00	Cerebral infarction NOS
G64z.11	Brainstem infarction NOS
G64z.12	Cerebellar infarction
G64z000	Brainstem infarction
G64z200	Left sided cerebral infarction
G64z300	Right sided cerebral infarction
G64z400	Infarction of basal ganglia
G66..00	Stroke and cerebrovascular accident unspecified
G66..11	CVA unspecified
G66..12	Stroke unspecified
G66..13	CVA - Cerebrovascular accident unspecified
G660.00	Middle cerebral artery syndrome
G661.00	Anterior cerebral artery syndrome
G662.00	Posterior cerebral artery syndrome
G663.00	Brain stem stroke syndrome
G664.00	Cerebellar stroke syndrome
G665.00	Pure motor lacunar syndrome
G666.00	Pure sensory lacunar syndrome
G667.00	Left sided CVA
G668.00	Right sided CVA
G676000	Cereb infarct due cerebral venous thrombosis; nonpyogenic
G677000	Occlusion and stenosis of middle cerebral artery
G677100	Occlusion and stenosis of anterior cerebral artery
G677200	Occlusion and stenosis of posterior cerebral artery
G677300	Occlusion and stenosis of cerebellar arteries
G677400	Occlusion+stenosis of multiple and bilat cerebral arteries
G6W..00	Cereb infarct due unsp occlus/stenos precerebr arteries
G6X..00	Cerebrl infarctn due/unspcf occlusn or sten/cerebrl artrs
Gyu6300	[X]Cerebrl infarctn due/unspcf occlusn or sten/cerebrl artrs
Gyu6400	[X]Other cerebral infarction
Gyu6500	[X]Occlusion and stenosis of other precerebral arteries
Gyu6600	[X]Occlusion and stenosis of other cerebral arteries
Gyu6G00	[X]Cereb infarct due unsp occlus/stenos precerebr arteries

B Supplementary materials : Survival model after TIA

B.1 Tables

TABLE B.1: Characteristics of the cases and controls: Original dataset.

Variable	Type of patients		‡
	Cases (n = 20,633)	Controls (n = 58,634)	
Demographical variables			
Age (mean (SD)) (years)	72.3 (10.7)	71.6 (10.7)	
39-60	3209 (15.6%)	9669 (16.5%)	
61-70	5229 (25.3%)	15608 (26.6%)	
71-76	4103 (19.9%)	11975 (20.4%)	
77+	8092 (39.2%)	21382 (36.5%)	
Sex: Males	9582 (46.4 %)	27269 (46.5 %)	
Year of entry			
1986-1992	971 (4.7 %)	2642 (4.5 %)	
1993-1999	5175 (25.1 %)	14530 (24.8 %)	
2000-2006	8474 (41.1 %)	24052 (41.0 %)	
2007-2016	6013 (29.1%)	17410 (29.7 %)	
IMD Quintile §			
1(Most deprived)	2739 (13.3 %)	6988 (11.9%)	
2	3706 (18.0 %)	10312 (17.6 %)	
3	4385 (21.2 %)	12506 (21.4%)	
4	4983 (24.2%)	14468 (24.6%)	
5 (Least Deprived)	4820 (23.4%)	14360 (24.5%)	
Pre-TIA conditions			
Asthma	2116 (10.3%)	5142 (8.8%)	
Atrial Fibrillation	2000 (9.7%)	2900 (4.9%)	*
Diabetes II			
No Diabetes	18211 (88.3%)	53726 (91.6%)	
Treated Diabetes	1965 (9.5%)	3867 (6.6%)	
Untreated Diabetes	457 (2.2%)	1041 (1.8%)	

Table B.1 *Continued from previous page*

Variable	Type of patients ‡		
	Cases (n = 20,633)	Controls (n = 58,634)	
COPD	1100 (5.3%)	2293 (3.9%)	
Heart Failure	1419(6.9%)	2557 (4.4%)	*
Hypertension	8960 (43.4%)	19282 (32.9%)	*
Hypercholesterolaemia	1447 (7%)	3118 (5.3%)	*
Myocardial Infarction	1632 (7.9%)	2916 (5%)	*
Peripheral Arterial Disease (PAD)	4936 (23.9%)	11244 (19.2%)	*
Premorbid prescriptions			
Anticoagulant agents	1640 (7.9%)	2965 (5.1%)	
Lipid lowering agents	5840 (28.3%)	10670 (18.2%)	
Antihypertensive agents	13772 (66.7%)	29820 (50.9%)	
Antiplatelet therapy within 6 months			
None	6373 (31%)	48959 (83.5%)	
Aspirin monotherapy	9282 (45%)	8924 (15.2%)	
Combination with aspirin(dual)	2584 (12.5%)	321 (0.5%)	
Other APL agents ††	2424 (11.7%)	430 (0.7%)	
Lifestyle factors			
Body Mass Index (BMI)			
Underweight (<= 18.5)	338 (1.6%)	1024 (1.7%)	
Healthy weight (18.5 to 24.9)	4470 (21.7%)	13208 (22.5%)	
Overweight (25 to 29.9)	5048 (24.5%)	14632 (25%)	
Obese (30 to 40)	2956 (14.3%)	7804 (13.3%)	
Missing	7821 (37.9%)	21966 (37.5%)	
Smoking status			
Non-smoker	7936 (38.5%)	24204 (41.3%)	
Ex-Smoker	5394 (26.1%)	13922 (23.7%)	
Current smoker	2363 (11.5%)	6040 (10.3%)	
Missing	4940 (23.9%)	14468 (24.7%)	
Alcohol consumption			
Abstainer	2489 (12.1%)	6479 (11%)	
Ex-consumer	530 (2.6%)	1086 (1.9%)	
Current consumer	7659 (37.1%)	23485 (40.1%)	
Missing	9955 (48.2%)	27584 (47%)	

Note. This table is based on the big dataset where the missing values of factors were imputed using the full-case model.
‡ Cases: TIA patients, Controls: non-TIA patients.

TABLE B.2: Cases and controls lost to follow-up by age category, social deprivation and TIA diagnosis.

Note : It displays the number of patients who were transferred out for all cases and controls.

Number of transferred out patients	**TIA diagnosis**	
Age category/IMD deprivation	*cases*	*controls*
39-62	**814**	**1027**
1 (most deprived)	152	194
2	168	186
3	157	204
4	175	224
5 (least deprived)	162	219
63-69	**981**	**1349**
1 (most deprived)	182	
2	187	251
3	206	298
4	207	333
5 (least deprived)	199	254
70-76	**1373**	**1871**
1 (most deprived)	178	272
2	242	369
3	274	413
4	355	439
5 (least deprived)	324	378
77+	**1718**	**1814**
1 (most deprived)	219	236
2	341	341
3	371	398
4	400	445
5 (least deprived)	387	394
Grand Total	**4886**	**6061**

§ IMD stands for Index of Multiple Deprivation in England. Quintile 1 is the most deprived group and Quintile 5 is least depriv

†† Other antiplatelet (APL) agents included dipyridamole or/& clopidogrel.

* $p < .05$. for proportion difference (χ^2 test).

TABLE B.3: Description and coding of variables in the study

Category	**Variable Description**	**Levels**
Demographic	Groups	cases, controls
	Death indicator	0=Censored,1=Dead
	Identity of GP practice	Frailty random term
	Gender	1= Male,2= Female
	Age at diagnosis	1="39-60",2="61-70", 3="71-76",4="77+"
	Birth cohort	1="1900-1920", 2="1921-1930" , 3="1931-1940", 4="1941-1960"
	Time of follow-up	Continuous variable
Social Deprivation	IMD(Index of Multiple Deprivation)	level 1(Most Deprived), 2,3,4,5 (Most Deprived)

Table B.3 continued from previous page

Category	Variable Description	Levels
Pre-stroke medical conditions and medical therapy	Asthma	0=Not diagnosed, 1=Diagnosed
	CKD	0=Not diagnosed, 1=Diagnosed (stages 1-2)
	COPD	0=Not diagnosed, 1=Diagnosed
	Heart Failure	0=Not diagnosed, 1=Diagnosed
	Hypothyroidism	0=Not diagnosed, 1=Diagnosed
	Myocardial Infarction	0=Not diagnosed, 1=Diagnosed
	PVD	0=Not diagnosed, 1=Diagnosed
	Hypertension Factor	1= Not diagnosed and not treated 2= Diagnosed and Treated 3= Diagnosed and Not Treated
	Hypercholesterolemia Factor	1= Not diagnosed and not treated 2= Diagnosed and Treated 3= Diagnosed and Not Treated
	Atrial Fibrillation Factor	1= Not diagnosed and not treated 2= Diagnosed and Treated 3= Diagnosed and Not Treated
	Diabetes II Factor	1= Not diagnosed and not treated 2= Diagnosed and Treated 3= Diagnosed and Not Treated
	Antiplatelet agents	0=No, 1= Yes
Lifestyle factors	BMI category	1=Underweight, 2=Normal,3=Overweight,4=Obese
	Smoking Status	1=Non-smoker,2=Ex-smoker and 3=Current smoker
	Alcohol Status	1=Non-current, 2=Current

TABLE B.4: Results of the Grambsch and Therneau (1994)'s test of the assumption of proportional hazards

Factors	**Rho**	**Chi-square**	***p*-value**
Birth_Cohort	52.1609	3	2.80E-11
Age_category	5.5745	3	0.008
Sex	0.6545	1	0.4185
Case-control status	9.4538	1	0.0021
IMD_Quintile	4.8205	4	0.3062
BMI_category	7.0349	1	0.008
Antiplatelets	21.4079	3	8.70E-05
Asthma	0.0647	1	0.7991
CKD	3.1115	1	0.0777
COPD	0.8182	1	0.3657
Heart_failure	8.1735	1	0.0043
Myocardial_infarction	6.3373	1	0.0118
PVD_PAD	3.1927	1	0.074
SMOKING	4.4481	2	0.1082
Alcohol_category	0.0224	1	0.8812
Diabetes_factor	1.9	2	0.3867
Atrial_fibrillation	0.5074	1	0.4762
Hypertension	0.6778	1	0.4104
Anticoagulant_agents	5.9703	1	0.0145
Hypertension	17.3865	1	3.00E-05
Age_cat:groups	4.27	3	0.2337
Groups:Hypertension	11.7695	1	0.0006
Groups:DUAL	11.2871	3	0.0103
GLOBAL	**152.7032**	**40**	**4.50E-15**

TABLE B.5: Table explaining the patterns of missingness between variables.

SMOKING	**BMI measurement**	**ALCOHOL**	**Missing Pattern**	**Number missing**
✓	✓	✓	0	36908
✓	✓	✗	1	9763
✓	✗	✓	1	4274
✓	✗	✗	2	8884
✗	✓	✓	1	339
✗	✓	✗	2	2440
✗	✗	✓	2	207
✗	✗	✗	3	16422
19408	29787	37539	86734	

Note. The table row total shows the number of values missing in different missing pattern. The column subtotal sh
values in each variables.
✓ = observed. ✗ = missing.

TABLE B.6: Correlation matrix jointly missing variables.

	SMOKING	**ALCOHOL**	**BMI_measurement**
SMOKING	1	0.5682581	0.5655395
ALCOHOL	0.5682581	1	0.5842658
BMI_measurement	0.5655395	0.5842658	1

TABLE B.7: Multiplicative hazard model : test for time-invariant effects (Kolmogorov-Smirnov test)

	Test for time invariant effects	
	Kolmogorov-Smirnov test	**p−value** H_0 : constant effect
(Intercept)	10.9	0
birth_cohort1921 to 1930	1.02	0.126
birth_cohort1931 to 1940	2.89	0
birth_cohort1941 to 1960	5.88	0
age_cat2	2.81	0.005
age_cat3	3.72	0
age_cat4	4.1	0.001
sexMale	0.437	0.558
groupscases	2.24	0.099
IMD_Quintile2	0.684	0.569
IMD_Quintile3	0.611	0.65
IMD_Quintile4	0.548	0.744
IMD_Quintile5	0.779	0.459
BMI_categoryObese+Overweight	0.907	0.011
antiplatelet_drugs_drugs1	0.807	0.001
asthma1	0.663	0.672
CKD1	4.64	0.489
COPD1	1.02	0.73
heart_failure1	1.88	0.075
myocardial_infarction1	0.821	0.54
PVD_PAD1	0.609	0.466
SMOKINGCurrent smoker	0.572	0.584
SMOKINGEx-Smoker	0.91	0.02
alcohol_catYes	0.426	0.704
Diabetes_factorYes and Treated	0.579	0.896
Diabetes_factorYesand Untreated	1.12	0.657
atrial_fibrillation1	1.2	0.377
hypertension1	0.593	0.397
anticoagulant_agents1	1.07	0.63
antihypertensive_agents1	1.24	0.034
age_cat2:groupscases	2.36	0.116
age_cat3:groupscases	2.25	0.166
age_cat4:groupscases	2.35	0.173
groupscases:antihypertensive_agents1	1.01	0.369

TABLE B.8: Multiplicative hazard model : test for time-invariant effects (Cramer von Mises test.)

	Cramer von Mises test	p**-value** H_0 : constant effect
(Intercept)	975	0
birth_cohort1921 to 1930	3.85	0.154
birth_cohort1931 to 1940	59.1	0
birth_cohort1941 to 1960	244	0
age_cat2	48.8	0.002
age_cat3	97.7	0
age_cat4	114	0
sexMale	0.33	0.751
groupscases	21.1	0.093
IMD_Quintile2	1.58	0.512
IMD_Quintile3	1.14	0.593
IMD_Quintile4	0.621	0.825
IMD_Quintile5	1.92	0.437
BMI_categoryObese+Overweight	3.18	0.092
antiplatelet_drugs_drugs1	5.25	0.006
asthma1	1.17	0.691
CKD1	83.5	0.427
COPD1	3.46	0.632
heart_failure1	23	0.027
myocardial_infarction1	2.1	0.5
PVD_PAD1	0.737	0.626
SMOKINGCurrent smoker	1.71	0.317
SMOKINGEx-Smoker	2.59	0.086
alcohol_catYes	0.602	0.61
Diabetes_factorYes and treated	0.417	0.978
Diabetes_factorYes and Untreated	2.5	0.803
atrial_fibrillation1	6.24	0.269
hypertension1	1.68	0.243
anticoagulant_agents1	5.27	0.423
antihypertensive_agents1	10.6	0.009
age_cat2:groupscases	14	0.269
age_cat3:groupscases	20	0.173
age_cat4:groupscases	25	0.136
groupscases:antihypertensive_agents1	3.18	0.405

TABLE B.9: Life expectancy model for the TIA database.

Coefficients	**Estimates**
$a(Scale)$	225.137
$b(Shape)$	2.412
exp(birth_cohort1.shape)	0.934
exp(birth_cohort2.shape)	0.864
exp(birth_cohort3.shape)	0.772
exp(heart_failure1.shape)	0.936
exp(APL1.shape)	0.894
exp(birth_cohort1.scale)	0.524
exp(birth_cohort2.scale)	0.274
exp(birth_cohort3.scale)	0.12
exp(female.scale)	0.673
exp(IMD_Quintile2.scale)	0.852
exp(IMD_Quintile3.scale)	0.833
exp(IMD_Quintile4.scale)	0.755
exp(IMD_Quintile5.scale)	0.719
exp(BMI_category1.scale)	0.825
exp(asthma1.scale)	1.241
exp(COPD1.scale)	1.696
exp(CKD1.scale)	1.183
exp(myocardial_infarction1.scale)	1.311
exp(PVD_PAD1.scale)	1.201
exp(SMOKING1.scale)	1.995
exp(SMOKING2.scale)	1.3
exp(alcohol_cat1.scale)	0.902
exp(atrial_fibrillation1.scale)	1.213
exp(Diabetes_factor1.scale)	1.777
exp(Diabetes_factor2.scale)	1.269
exp(anticoagulant_agents1.scale)	1.395
exp(groupscases1.scale)	1.641
exp(hypertension.scale)	1.328
exp(APL1.scale)	0.561
exp(age_c.scale)	1.085
exp(groupscases1:APL1.scale)	0.803
σ^2	0.066
Loglik	-24813.79
AIC	49695.58

TABLE B.10: Estimated coefficients of full-case model and model on the imputed data and their significance.

Variable	Full case			Multiple Imputed dataset		
	Estimate[§]	**95% confidence Interval**	p-**value**	**Estimate** [§]	**95% confidence Interval**	p-**value**
Sample size	24,797	-	-	79,267	-	-
Number of non-censored	6,188	-	-	24,176	-	-
$a(Scale)$	25.45	23.68 – 27.36	0	29.19	28.79 – 29.59	0
$b(Shape)$	3.36	3-3.76	0	2.39	2.31 – 2.48	0
exp(birth_cohort1921 to 1930.shape)	0.86	0.82 – 0.91	0	0.84	0.82 – 0.86	0
exp(birth_cohort1931 to 1940.shape)	0.70	0.65 – 0.75	0	0.66	0.62 – 0.7	0
exp(birth_cohort1941 to 1960.shape)	0.53	0.47 – 0.6	0	0.48	0.4 – 0.56	0
exp(BMI_categoryObese+Overweight.shape)	1.07	1.03 – 1.11	0	1.02	0.98 – 1.05	0.68
exp(heart_failure1.shape)	0.82	0.78 – 0.87	0	0.79	0.77 – 0.81	0
exp(age_cat2.shape)	0.83	0.76 – 0.92	0	0.90	0.82 – 0.98	0
exp(age_cat3.shape)	0.78	0.7 – 0.87	0	0.77	0.7 – 0.85	0
exp(age_cat4.shape)	0.66	0.59 – 0.7	0	0.59	0.51 – 0.67	0
exp(APL_Aspirin only.shape)	0.91	0.87 – 0.95	0	0.96	0.94 – 0.98	0
exp(APL_DUAL Therapy(with Aspirin).shape)	0.89	0.81 – 0.98	0.01	0.90	0.84 – 0.95	0
exp(APL_Other Non Aspirin.shape)	0.85	0.76 – 0.94	0.00	0.89	0.83 – 0.95	0
exp(birth_cohort1921 to 1930.scale)	0.54	0.49 – 0.6	0	0.50	0.44 – 0.56	0
exp(birth_cohort1931 to 1940.scale)	0.29	0.26 – 0.33	0	0.23	0.15 – 0.31	0
exp(birth_cohort1941 to 1960.scale)	0.12	0.1 – 0.15	0	0.09	0.03 – 0.2	0
exp(sexMale.scale)	1.31	1.24 – 1.42	0	1.30	1.28 – 1.33	0
exp(IMD_Quintile2.scale)	0.88	0.81 – 0.96	0.00	0.92	0.88 – 0.96	0
exp(IMD_Quintile3.scale)	0.88	0.8 – 0.96	0.00	0.92	0.88 – 0.96	0
exp(IMD_Quintile4.scale)	0.79	0.72 – 0.87	0	0.87	0.81 – 0.93	0
exp(IMD_Quintile5.scale)	0.76	0.69 – 0.84	0	0.81	0.75 – 0.87	0
exp(BMI_categoryObese+Overweight.scale)	0.88	0.82 – 0.95	0	0.87	0.82 – 0.92	0

[§] Estimated coefficients of full-case model and model on the imputed data and their significance.

... Continued from Table B.10

Variable	Full case			Multiple Imputed dataset		
	Estimate[§]	95% confidence Interval	*p*-value	Estimate [§]	95% confidence Interval	*p*-value
exp(asthma1.scale)	1.26	1.16 – 1.37	0	1.16	1.12 – 1.2	0
exp(COPD1.scale)	1.71	1.53 – 1.91	0	1.73	1.67 – 1.79	0
exp(CKD1.scale)	1.18	0.99 – 1.4	0.06	1.04	0.96 – 1.11	0.07
exp(myocardial_infarction1.scale)	1.33	1.22 – 1.45	0	1.20	1.16 – 1.24	0
exp(PVD_PAD1.scale)	1.21	1.14 – 1.29	0	1.14	1.1 – 1.18	0
exp(SMOKING_Current smoker.scale)	1.90	1.77 – 2.05	0	2.04	1.96 – 2.11	0
exp(SMOKING_Ex-Smoker.scale)	1.31	1.23 – 1.39	0	1.24	1.2 – 1.29	0
exp(alcohol_cat_Yes.scale)	0.90	0.85 – 0.95	0	0.89	0.85 – 0.93	0
exp(atrial_fibrillation1.scale)	1.28	1.15 – 1.43	0	1.27	1.21 – 1.32	0
exp(Diabetes_factor_Yes and Treated.scale)	1.82	1.67 – 1.98	0	1.52	1.48 – 1.56	0
exp(Diabetes_factor_Yes and Untreated.scale)	1.32	1.14 – 1.53	0	1.12	1.04 – 1.2	0
exp(anticoagulant_agents1.scale)	1.40	1.25 – 1.57	0	1.23	1.17 – 1.29	0
exp(groups_cases.scale)	2.29	1.89 – 2.77	0	3.04	2.91 – 3.18	0
exp(hypertension.scale)	1.45	1.34 – 1.57	0	1.41	1.37 – 1.45	0
exp(age_cat2.scale)	1.63	1.4 – 1.91	0	1.77	1.66 – 1.89	0
exp(age_cat3.scale)	2.34	1.97 – 2.76	0	2.12	2 – 2.24	0
exp(age_cat4.scale)	3.73	3.16 – 4.4	0	2.65	2.55 – 2.76	0
exp(APL_Aspirin only.scale)	0.96	0.85 – 1.08	0.47	1.01	0.95 – 1.07	0.447
exp(APL_DUAL Therapy .scale)	1.35	0.85 – 2.15	0.21	1.11	0.9 – 1.33	0.23
exp(APL_Other Non Aspirin.scale)	0.93	0.62 – 1.41	0.74	0.99	0.79 – 1.18	0.78
exp(groups_cases:APLA_spirin only.scale)	0.91	0.8 – 1.03	0.14	0.88	0.83 – 0.94	0.04
exp(groups_cases:APL_DUAL Therapy.scale)	0.65	0.41 – 1.01	0.05	0.75	0.53 – 0.96	0.044
exp(groups_cases:APL_Other Non Aspirin.scale)	0.84	0.56 – 1.25	0.39	0.86	0.66 – 1.06	0.06
exp(groups_cases:hypertension.scale)	0.91	0.81 – 1.02	0.09	0.92	0.86 – 0.97	0
exp(groups_cases:age_cat2.scale)	0.84	0.69 – 1.04	0.10	0.65	0.51 – 0.79	0
exp(groups_cases:age_cat3.scale)	0.71	0.57 – 0.87	0.00	0.54	0.41 – 0.68	0
exp(groups_cases:age_cat4.scale)	0.59	0.48 – 0.73	0	0.50	0.36 – 0.64	0
σ^2	0.08	0.06 – 0.11	0	0.08	0.04 – 0.13	0

[§] Estimated coefficients for Weibull Double-Cox model fitted to the full case records and imputed datasets.

TABLE B.11: Estimated coefficients of full-case model and model on the imputed data and their significance: Survival model for Life expectancy.

		Full Case model			Model on Imputed model		
Definition of parameter	**Parameter**	**Estimates**	**95% CI**	***p*-value**	**Estimates**	**95% CI**	***p*-value**
a	Scale	215.65	169.92-273.04	0	225.14	179.41-282.52	0
b	Shape	2.13	2.03-2.23	0	2.41	2.31-2.51	0
Born in 1921-1930	Shape	0.92	0.88-0.97	0.0035	0.93	0.89-0.98	0.0044
Born in 1931-1940	Shape	0.82	0.77-0.87	0	0.86	0.82-0.91	0
Born in 1941-1960	Shape	0.87	0.8-0.95	0	0.77	0.7-0.85	0
Heart-failure - Present	Shape	0.93	0.92-0.95	0	0.94	0.92-0.95	0
Antiplatelet therapy	Shape	0.89	0.86-0.93	0	0.89	0.86-0.93	0
Born in 1921-1930	Scale	0.53	0.4-0.71	0	0.52	0.39-0.71	0
Born in 1931-1940	Scale	0.28	0.21-0.39	0	0.27	0.2-0.38	0
Born in 1941-1960	Scale	0.14	0.09-0.22	0	0.12	0.07-0.2	0
Females	Scale	0.57	0.53-0.61	0	0.67	0.64-0.71	0
IMD Quintile 2	Scale	0.69	0.61-0.77	4.00E-04	0.85	0.78-0.93	5.00E-04
IMD Quintile 3	Scale	0.73	0.66-0.81	1.00E-04	0.83	0.76-0.91	1.00E-04
IMD Quintile 4	Scale	0.66	0.59-0.73	0	0.76	0.69-0.83	0
IMD Quintile 5	Scale	0.73	0.67-0.81	0	0.72	0.65-0.79	0
BMI (Overweight & Obese)	Scale	0.84	0.8-0.89	0	0.83	0.78-0.87	0
Asthma - Present	Scale	1.26	1.16-1.38	0	1.24	1.14-1.35	0
Chronic obstructive pulmonary disease	Scale	1.77	1.59-1.97	0	1.70	1.52-1.9	0
Chronic Kidney Disease	Scale	1.17	0.98-1.39	0.0345	1.18	1-1.41	0.0545
Myocardial infarction	Scale	1.83	1.72-1.95	0	1.31	1.2-1.43	0
Peripheral Artery Disease/Vascular Disease	Scale	1.52	1.45-1.6	0	1.20	1.13-1.28	0
Current smoker	Scale	1.97	1.83-2.13	0	2.00	1.85-2.15	0
Ex-smoker	Scale	1.28	1.2-1.36	0	1.30	1.22-1.38	0
Alcohol intake - Yes	Scale	0.93	0.88-0.99	7.50E-04	0.90	0.85-0.96	7.00E-04
Atrial fibrillation - Present	Scale	1.24	1.11-1.38	3.94E-04	1.21	1.09-1.35	4.00E-04
Treated Diabetes	Scale	1.97	1.83-2.13	0	1.78	1.63-1.93	0
Untreated Diabetes	Scale	1.12	0.95-1.32	0.0018	1.27	1.09-1.47	0.0016
Anticoagulant agents	Scale	1.93	1.78-2.1	0	1.40	1.24-1.56	0
Cases-group	Scale	1.74	1.61-1.88	0	1.64	1.52-1.78	0
Antihypertensive agents	Scale	1.31	1.23-1.39	0	1.33	1.25-1.41	0
Antiplatelet therapy	Scale	0.58	0.45-0.73	0	0.56	0.44-0.72	0
Age as continuous	Scale	1.08	1.07-1.09	0	1.09	1.08-1.09	0
Interaction of cases & Antiplatelet Therapy	Scale	0.87	0.78-0.97	2.00E-04	0.80	0.71-0.91	3.00E-04
σ^2		0.08	0.06-0.1	0	0.07	0.05-0.09	0

TABLE B.12: Hazard ratios over different years associated with uptake of different types of antiplatelets for TIA patients by different age groups.

Age 39-60 years			
Time (years)	**Aspirin only**	**DAPT**	**Other APL**
1	1(0.89-1.11)	1.16(0.79-1.62)	1.16(0.79-1.62)
5	0.93(0.84-1.01)	0.92(0.61-1.26)	0.94(0.67-1.25)
10	0.9(0.82-0.98)	0.85(0.56-1.15)	0.86(0.61-1.14)
15	0.88(0.8-0.96)	0.81(0.54-1.1)	0.82(0.58-1.09)

Age 61- 70 years			
Time (years)	**Aspirin only**	**DAPT**	**Other APL**
1	0.98(0.88-1.09)	1.07(0.69-1.51)	1.2(0.82-1.68)
5	0.92(0.84-1)	0.9(0.6-1.23)	0.96(0.68-1.28)
10	0.89(0.82-0.98)	0.83(0.56-1.14)	0.87(0.62-1.16)
15	0.88(0.8-0.96)	0.8(0.53-1.09)	0.82(0.59-1.09)

Age 71-76 years			
Time (years)	**Aspirin only**	**DAPT**	**Other APL**
1	1.01(0.9-1.13)	1.14(0.74-1.65)	1.2(0.82-1.68)
5	0.93(0.85-1.02)	0.93(0.62-1.29)	0.96(0.68-1.28)
10	0.9(0.82-0.99)	0.85(0.57-1.17)	0.87(0.62-1.16)
15	0.88(0.81-0.97)	0.81(0.54-1.11)	0.82(0.59-1.09)

Age 77+ years			
Time (years)	**Aspirin only**	**DAPT**	**Other APL**
1	1.04(0.92-1.16)	1.22(0.77-1.81)	1.29(0.85-1.86)
5	0.95(0.86-1.04)	0.96(0.64-1.33)	0.99(0.7-1.34)
10	0.91(0.83-0.99)	0.87(0.58-1.19)	0.89(0.63-1.18)
15	0.89(0.81-0.97)	0.82(0.55-1.12)	0.83(0.59-1.11)

TABLE B.13: Hazard ratios over different years associated with uptake of different types of antiplatelets for TIA controls by different age groups.

Age 39-60 years			
Time (years)	**Aspirin only**	**DAPT**	**Other APL**
1	1.13(1.04-1.21)	1.49(1.13-2.06)	1.35(1.07-1.68)
5	1.05(0.99-1.1)	1.23(0.95-1.51)	1.1(0.89-1.28)
10	1.02(0.96-1.07)	1.14(0.9-1.33)	1(0.82-1.16)
15	1(0.94-1.05)	1.08(0.85-1.28)	0.95(0.78-1.1)

Age 61- 70 years			
Time (years)	**Aspirin only**	**DAPT**	**Other APL**
1	1.12(1.02-1.20)	1.43(1.22-1.69)	1.29(0.99-1.66)
5	1.04(0.98-1.1)	1.21(1.11-1.26)	1.07(0.85-1.30)
10	1.01(0.95-1.07)	1.12(1.05-1.18)	0.99(0.79-1.19)
15	1(0.94-1.05)	1.07(1.01-1.13)	0.95(0.75-1.13)

Age 71-76 years			
Time (years)	**Aspirin only**	**DAPT**	**Other APL**
1	1.14(1.05-1.25)	1.54(1.15-1.99)	1.40(1.04-1.8)
5	1.06(0.99-1.12)	1.25(0.99-1.51)	1.12(0.88-1.35)
10	1.02(0.96-1.078)	1.14(0.91-1.37)	1(0.81-1.21)
15	1(0.94-1.06)	1.08(0.87-1.30)	0.96(0.77-1.15)

Age 77+ years			
Time (years)	**Aspirin only**	**DAPT**	**Other APL**
1	1.10(1.02-1.18)	1.38(1.08-1.73)	1.25(0.97-1.57)
5	1.03(0.97-1.11)	1.18(0.95-1.42)	1.05(0.84-1.27)
10	1.01(0.94-1.06)	1.11(0.89-1.32)	0.98(0.78-1.19)
15	1(0.94-1.05)	1.06(0.85-1.28)	0.94(0.75-1.13)

B.2 Figures

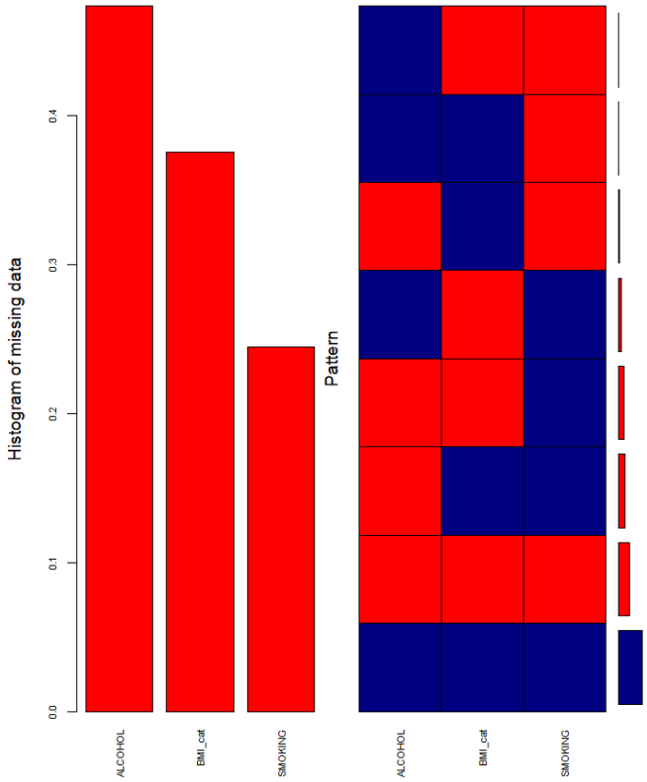


FIGURE B.1: The missing pattern in the TIA dataset and the association of missing records.

C Supplementary materials : Survival model after IS

C.1 Tables

TABLE C.1: Results of the Grambsch and Therneau (1994)'s test of the assumption of proportional hazards.

Variables	**Chi-square**	**df**	***p*-value**
Birth_cohort	21.9523	3	6.70E-05
Age_cat	11.9221	3	0.00765
Sex	0.7516	1	0.38596
Groups	43.4684	1	4.30E-11
IMD_Quintile	4.1569	4	0.38519
BMI_cat	2.5118	3	0.47317
APL	47.4093	3	2.80E-10
asthma	0.908	1	0.34064
CKD	10.3074	1	0.00132
COPD	1.5238	1	0.21705
Heart_failure	6.3192	1	0.01194
Myocardial_infarction	0.6413	1	0.42323
PAD	1.5223	1	0.21727
SMOKING	2.832	2	0.24269
Alcohol_cat	0.4474	1	0.5036
Diabetes_factor	6.4835	2	0.0391
Atrial_fibrillation	0.0225	1	0.88081
Hypertension	22.3166	1	2.30E-06
Age_cat:groups	25.9859	3	9.60E-06
Sex:groups	13.9863	1	0.00018
Groups:hypertension	36.0279	1	1.90E-09
GLOBAL	**142.0436**	**36**	**1.60E-14**

TABLE C.2: Estimated coefficients and associated significance level of full-case model and data on the imputed data for survival after IS.

Variable	Full case model			Model on the imputed data		
	Estimate[§]	**95% confidence Interval**	***p*-value**	**Estimate** [§]	**95% confidence Interval**	***p*-value**
Sample size	25,711	-	-	75,769	-	-
Number of non-censored	6,372	-	-	23,959	-	-
$a(Scale)$	36.521	33.027 − 40.385	‘ 0	48.868	43.523 − 54.869	0
$b(Shape)$	2.377	2.258 − 2.503	0	1.228	1.2 − 1.256	0
exp(birth_cohort1921 to 1930.shape)	0.937	0.891 − 0.985	0.0103	1.015	0.992 − 1.039	0.2054
exp(birth_cohort1931 to 1940.shape)	0.906	0.856 − 0.959	7.00E-04	0.953	0.923 − 0.985	0.004
exp(birth_cohort1941 to 1960.shape)	0.764	0.705 − 0.828	0	0.768	0.729 − 0.809	0
exp(APLAspirin.shape)	0.903	0.866 − 0.941	0	0.945	0.923 − 0.967	0
exp(APLDual therapy.shape)	0.791	0.739 − 0.847	0	0.982	0.938 − 1.027	0.4249
exp(APLother.shape)	0.837	0.761 − 0.922	3.00E-04	0.89	0.84 − 0.943	1.00E-04
exp(Hypertension.shape)	0.872	0.837 − 0.909	0	0.834	0.816 − 0.853	0
exp(birth_cohort1921 to 1930.scale)	0.655	0.572 − 0.749	0	0.673	0.635 − 0.712	0
exp(birth_cohort1931 to 1940.scale)	0.435	0.37 − 0.511	0	0.348	0.321 − 0.378	0
exp(birth_cohort1941 to 1960.scale)	0.213	0.171 − 0.264	0	0.146	0.129 − 0.164	0
exp(IMD_Quintile2.scale)	0.944	0.866 − 1.03	0.1941	0.974	0.929 − 1.02	0.2667
exp(IMD_Quintile3.scale)	0.949	0.87 − 1.036	0.2422	0.938	0.895 − 0.984	0.0083
exp(IMD_Quintile4.scale)	0.825	0.755 − 0.903	0	0.876	0.835 − 0.92	0
exp(IMD_Quintile5.scale)	0.826	0.752 − 0.907	1.00E-04	0.843	0.8 − 0.888	0
exp(BMI_categoryObese +Overweight .scale)	0.773	0.734 − 0.814	0	0.845	0.823 − 0.868	0
exp(anticoagulant_agents1.scale)	1.18	1.064 − 1.308	0.0016	1.14	1.081 − 1.202	0
exp(APLAspirin.scale)	0.888	0.778 − 1.013	0.0778	0.966	0.909 − 1.027	0.2654
exp(APLDual therapy.scale)	1.092	0.778 − 1.532	0.6113	1.28	1.065 − 1.538	0.0086
exp(APLother.scale)	0.706	0.465 − 1.072	0.1024	0.919	0.758 − 1.115	0.3944
exp(atrial_fibrillation1.scale)	1.332	1.216 − 1.459	0	1.182	1.129 − 1.238	0
exp(Diabetes_factorYes and Treated.scale)	1.787	1.662 − 1.921	0	1.354	1.296 − 1.414	0
exp(Diabetes_factorYes and Untreated.scale)	1.378	1.213 − 1.565	0	1.096	1.014 − 1.185	0

[§] Estimated coefficients for Weibull double-Cox model fitted to the full case records and imputed datasets.

... Continued from Table C.2

Variable	Full case model			Model on the imputed data		
	Estimate[§]	**95% confidence Interval**	***p*-value**	**Estimate** [§]	**95% confidence Interval**	***p*-value**
exp(CKD1.scale)	1.244	1.102 − 1.404	4.00E-04	1.145	1.078 − 1.215	0.0206
exp(COPD1.scale)	1.952	1.778 − 2.142	0	1.619	1.541 − 1.702	0
exp(heart_failure1.scale)	1.589	1.443 − 1.749	0	1.543	1.476 − 1.613	0
exp(myocardial_infarction1.scale)	1.234	1.137 − 1.339	0	1.142	1.091 − 1.195	0
exp(PVD_PAD1.scale)	1.156	1.088 − 1.229	0	1.107	1.073 − 1.142	0
exp(SMOKINGCurrent smoker.scale)	1.836	1.704 − 1.978	0	1.946	1.872 − 2.022	0
exp(SMOKINGEx−Smoker.scale)	1.263	1.192 − 1.337	0	1.282	1.244 − 1.32	0
exp(groupscases.scale)	3.818	3.116 − 4.679	0	5.667	4.932 − 6.511	0
exp(sexMale.scale)	1.399	1.3 − 1.505	0	1.222	1.181 − 1.265	0
exp(Hypertension.scale)	1.122	1 − 1.26	0.0502	0.955	0.902 − 1.011	0.1139
exp(age_cat2.scale)	2.064	1.74 − 2.449	0	1.861	1.652 − 2.097	0
exp(age_cat3.scale)	3.412	2.833 − 4.109	0	2.561	2.262 − 2.901	0
exp(age_cat4.scale)	5.818	4.785 − 7.073	0	4.05	3.569 − 4.596	0
exp(groupscases:age_cat2.scale)	0.759	0.619 − 0.93	0.0078	0.668	0.577 − 0.772	0
exp(groupscases:age_cat3.scale)	0.575	0.468 − 0.706	0	0.533	0.462 − 0.615	0
exp(groupscases:age_cat4.scale)	0.564	0.461 − 0.69	0	0.521	0.454 − 0.597	0
exp(groupscases:sexMale.scale)	0.879	0.794 − 0.974	0.0136	0.891	0.845 − 0.941	0
exp(groupscases:Hypertension)	0.856	0.763 − 0.96	0.0078	0.958	0.902 − 1.018	0.171
exp(APLAspirin:groupscases.scale)	0.911	0.803 − 1.035	0.1511	0.716	0.672 − 0.763	0
exp(APLDual therapy:groupscases.scale)	0.562	0.413 − 0.766	3.00E-04	0.525	0.443 − 0.622	0
exp(APLother:groupscases.scale)	1.029	0.704 − 1.503	0.8843	0.661	0.552 − 0.793	0
σ^2	0.058	0.041 − 0.081	0	0.073	0.059 − 0.091	0

[§] Estimated coefficients for Weibull double-Cox model fitted to the full case records and imputed datasets.

TABLE C.3: Hazard ratios over different years associated with uptake of different types of antiplatelets for patients aged 39−60 years at entry (*IS dataset).*

	Time(years)	Aspirin	Dual Therapy	Other Antiplatelets
Cases	1	0.8(0.67 − 0.97)	0.7(0.46−0.97)	0.81(0.5−1.3)
	3	0.76(0.65−0.88)	0.69(0.46−0.94)	0.72(0.47−1.07)
	5	0.74(0.64−0.84)	0.69(0.45−0.92)	0.68(0.45−0.98)
	10	0.71(0.63−0.8)	0.68(0.45−0.91)	0.64(0.43−0.88)
	15	0.69(0.62−0.77)	0.67(0.59−0.76)	0.61(0.42−0.84)
	Time (years)	**Aspirin**	**Dual Therapy**	**Other Antiplatelets**
Controls	1	1.12(0.96−1.33)	1.34(1.24−1.49)	1.22(0.85−1.85)
	3	1.06(0.94−1.2)	1.30(1.23−1.40)	1.09(0.81−1.48)
	5	1.03(0.93−1.15)	1.310(1.23−1.39)	1.04(0.79−1.35)
	10	0.99(0.91−1.08)	1.29(1.23−1.36)	0.96(0.76−1.2)
	15	0.97(0.9−1.05)	1.28(1.22−1.35)	0.92(0.74−1.13)

TABLE C.4: Hazard ratios over different years associated with uptake of different types of antiplatelets for patients aged 61−70 years at entry (*IS dataset).*

	Time(years)	Aspirin	Dual Therapy	Other Antiplatelets
Cases	1	0.84(0.68−1.05)	0.72(0.46−1.01)	0.89(0.53−1.53)
	3	0.78(0.66−0.93)	0.7(0.45−0.95)	0.77(0.49−1.19)
	5	0.76(0.65−0.88)	0.69(0.45−0.94)	0.73(0.47−1.07)
	10	0.72(0.64−0.82)	0.68(0.45−0.92)	0.66(0.44−0.93)
	15	0.71(0.63−0.79)	0.68(0.45−0.91)	0.63(0.43−0.86)
	Time(years)	**Aspirin**	**Dual Therapy**	**Other Antiplatelets**
Controls	1	1.17(0.97−1.45)	1.36(1.24−1.54)	1.35(0.88−2.26)
	3	1.09(0.95−1.28)	1.33(1.24−1.46)	1.17(0.84−1.7)
	5	1.06(0.94−1.2)	1.32(1.23−1.42)	1.1(0.81−1.5)
	10	1.01(0.92−1.12)	1.3(1.23−1.38)	1(0.77−1.28)
	15	0.98(0.91−1.07)	1.29(1.22−1.36)	0.95(0.75−1.18)

TABLE C.5: Hazard ratios over different years associated with uptake of different types of antiplatelets for patients aged 71−76 years at entry (*IS dataset).*

Cases	**Time(years)**	**Aspirin**	**Dual Therapy**	**Other Antiplatelets**
	1	0.85(0.68−1.15)	0.72(0.46−1.03)	0.92(0.52−1.86)
	3	0.79(0.66−0.99)	0.7(0.46−0.97)	0.79(0.48−1.35)
	5	0.76(0.65−0.92)	0.69(0.46−0.96)	0.74(0.46−1.17)
	10	0.73(0.64−0.85)	0.68(0.45−0.92)	0.67(0.44−0.97)
	15	0.71(0.63−0.81)	0.68(0.59−0.76)	0.68(0.45−0.91)
Controls	**Time(years)**	**Aspirin**	**Dual Therapy**	**Other Antiplatelets**
	1	1.19(0.96−1.61)	1.37(1.25−1.59)	1.39(0.88−2.66)
	3	1.11(0.94−1.37)	1.34(1.24−1.49)	1.2(0.84−1.91)
	5	1.07(0.93−1.27)	1.32(1.24−1.45)	1.12(0.81−1.64)
	10	1.02(0.92−1.16)	1.3(1.23−1.39)	1.02(0.78−1.36)
	15	0.99(0.91−1.09)	1.29(1.22−1.36)	0.96(0.75−1.22)

TABLE C.6: Hazard ratios over different years associated with uptake of different types of antiplatelets for patients aged 77+ years at entry (*IS dataset)*.

	Time(years)	**Aspirin**	**Dual Therapy**	**Other Antiplatelets**
Cases	1	0.85(0.68−1.08)	0.72(0.46−1.03)	0.91(0.54−1.59)
	3	0.79(0.66−0.95)	0.69(0.46−0.95)	0.79(0.5−1.22)
	5	0.76(0.65−0.89)	0.69(0.45−0.95)	0.74(0.48−1.08)
	10	0.73(0.64−0.83)	0.68(0.45−0.93)	0.67(0.45−0.94)
	15	0.71(0.63−0.79)	0.68(0.44−0.91)	0.63(0.43−0.87)
	Time(years)	**Aspirin**	**Dual Therapy**	**Other Antiplatelets**
Controls	1	1.19(0.97−1.48)	1.37(1.25−1.56)	1.38(0.89−2.34)
	3	1.1(0.95−1.3)	1.34(1.24−1.47)	1.19(0.84−1.74)
	5	1.06(0.94−1.22)	1.32(1.24−1.43)	1.11(0.82−1.53)
	10	1.02(0.92−1.13)	1.3(1.23−1.39)	1.01(0.78−1.29)
	15	0.99(0.91−1.08)	1.29(1.23−1.36)	0.96(0.75−1.19)

TABLE C.7: Life expectancy model using the IS dataset fitted to a Weibull double-Cox model.

Coefficients	**Estimates**
a	338.339
b	2.254
exp(birth_cohort1.shape)	0.903
exp(birth_cohort2.shape)	0.885
exp(birth_cohort3.shape)	0.782
exp(antiplatelet_drugs_drugs1.shape)	0.866
exp(birth_cohort1.scale)	0.476
exp(birth_cohort2.scale)	0.351
exp(birth_cohort3.scale)	0.162
exp(IMD_Quintile2.scale)	0.951
exp(IMD_Quintile3.scale)	0.96
exp(IMD_Quintile4.scale)	0.825
exp(IMD_Quintile5.scale)	0.827
exp(BMI_category1 .scale)	0.797
exp(antiplatelet_drugs_drugs1.scale)	0.409
exp(COPD1.scale)	1.884
exp(heart_failure1.scale)	1.639
exp(myocardial_infarction1.scale)	1.238
exp(PVD_PAD1.scale)	1.14
exp(SMOKING1.scale)	1.941
exp(SMOKING2.scale)	1.27
exp(anticoagulant_agents1.scale)	1.313
exp(Diabetes_factor1.scale)	1.78
exp(Diabetes_factor2.scale)	1.329
exp(sex1.scale)	1.343
exp(groupscases1.scale)	2.345
exp(Hypertersion.scale)	1.483
exp(age_c.scale)	1.084
exp(groupscases1:Hypertension.scale)	0.803
σ^2	0.059
Loglik	-24987.97
AIC	50035.94

TABLE C.8: Estimated coefficients of full-case model and model on the imputed data and their significance: Survival model for Life expectancy.

		Full case model			Model on Imputed model		
Definition of parameter	**Parameter**	**Estimates**	**95% CI**	***p*-value**	**Estimates**	**95% CI**	***p*-value**
a	Scale	332.45	260.19-424.34	0	338.34	266.08-430.23	0
b	Shape	2.15	2.05-2.25	0	2.25	2.16-2.36	0
Born in 1921-1930	Shape	0.87	0.83-0.92	0	0.90	0.86-0.95	0
Born in 1931-1940	Shape	0.87	0.82-0.92	0	0.89	0.84-0.94	0
Born in 1941-1960	Shape	0.75	0.69-0.81	0	0.78	0.72-0.85	0
Antiplatelet drugs	Shape	0.82	0.79-0.85	0	0.87	0.84-0.9	0
Born in 1921-1930	Scale	0.47	0.34-0.66	0	0.48	0.34-0.66	0
Born in 1931-1940	Scale	0.32	0.21-0.47	0	0.35	0.24-0.51	0
Born in 1941-1960	Scale	0.16	0.1-0.26	0	0.16	0.1-0.26	0
IMD Quintile 2	Scale	0.93	0.85-1.02	0.2659	0.95	0.87-1.04	0.253
IMD Quintile 3	Scale	0.98	0.9-1.07	0.3258	0.96	0.88-1.05	0.3549
IMD Quintile 4	Scale	0.81	0.74-0.89	0	0.83	0.75-0.9	0
IMD Quintile 5	Scale	0.83	0.75-0.91	0.00E+00	0.83	0.75-0.91	1.00E-04
BMI (Overweight + Obese)	Scale	0.80	0.76-0.84	0	0.80	0.76-0.84	0
Antiplatelet drugs	Scale	0.34	0.25-0.46	0	0.41	0.32-0.53	0
Chronic obstructive pulmonary disease	Scale	1.49	1.32-1.67	0	1.88	1.72-2.07	0
Heart failure Present	Scale	1.36	1.21-1.53	0	1.64	1.49-1.8	0
Myocardial infarction Present	Scale	1.27	1.18-1.38	0	1.24	1.14-1.34	0
Peripheral Artery Disease	Scale	1.31	1.25-1.39	0	1.14	1.07-1.21	0
Current smoker	Scale	1.84	1.7-1.99	0	1.94	1.8-2.09	0
Ex-smoker	Scale	1.53	1.46-1.6	0	1.27	1.2-1.35	0
Anticoagulant agents Present	Scale	1.31	1.2-1.44	0	1.31	1.2-1.44	0
Treated Diabetes	Scale	1.68	1.55-1.81	0	1.78	1.66-1.91	0
Untreated Diabetes	Scale	1.33	1.17-1.51	0	1.33	1.17-1.51	0
Gender - Males	Scale	1.73	1.66-1.81	0	1.34	1.27-1.42	0
Cases group	Scale	2.15	1.94-2.37	0	2.35	2.14-2.57	0
Hypertension	Scale	1.68	1.57-1.8	0	1.48	1.37-1.6	0
Age as continuous	Scale	1.41	1.4-1.43	0	1.08	1.08-1.09	0
Interaction of cases& hypertension	Scale	0.83	0.75-0.93	0.00E+00	0.80	0.72-0.9	1.00E-04
Variance (Sigma2)		0.07	0.05-0.09	0	0.06	0.04-0.08	0

Bibliography

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

Aburto, J. M., Kashyap, R., Schöley, J., Angus, C., Ermisch, J., Mills, M. C., & Dowd, J. B. (2021). Estimating the burden of the covid-19 pandemic on mortality, life expectancy and lifespan inequality in england and wales: A population-level analysis. *J Epidemiol Community Health*.

Aburto, J. M., Schöley, J., Zhang, L., Kashnitsky, I., Rahal, C., Missov, T. I., Mills, M. C., Dowd, J. B., & Kashyap, R. (2021). Recent gains in life expectancy reversed by the covid-19 pandemic. *medRxiv*.

Adams Jr, H. P., Bendixen, B. H., Kappelle, L. J., Biller, J., Love, B. B., Gordon, D. L., & Marsh 3rd, E. (1993). Classification of subtype of acute ischemic stroke. definitions for use in a multicenter clinical trial. toast. trial of org 10172 in acute stroke treatment. *stroke*, *24*(1), 35–41.

Albers, G. W., Caplan, L. R., Easton, J. D., Fayad, P. B., Mohr, J., Saver, J. L., & Sherman, D. G. (2002). Transient ischemic attack—proposal for a new definition.

Allison, P. D. (2001). Missing data (vol. 136). *Thousand Oaks, CA, US: Sage publications*.

Allison, P. D. (2010). *Survival analysis using sas: A practical guide*. Sas Institute.

Andersen, Andersen, K. K., Kammersgaard, L. P., & Olsen, T. S. (2005). Sex differences in stroke survival: 10-year follow-up of the copenhagen stroke study cohort. *Journal of Stroke and Cerebrovascular Diseases*, *14*(5), 215–220.

Andersen & 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, C. S., Carter, K. N., Brownlee, W. J., Hackett, M. L., Broad, J. B., & Bonita, R. (2004). Very long-term outcome after stroke in auckland, new zealand. *Stroke*, *35*(8), 1920–1924.

Andrinopoulou. (2014). Introduction to joint models of longitudinal and survival data [Accessed: 2018-10-19].

Asthma & Foundation, B. L. (2019, August). *Asthma statistics*. https://statistics.blf.org.uk/asthma

Bamford, J., Sandercock, P., Dennis, M., Warlow, C., & Burn, J. (1991). Classification and natural history of clinically identifiable subtypes of cerebral infarction. *The Lancet*, *337*(8756), 1521–1526.

Barrett, K. M., Brott, T. G., Brown, R. D., Frankel, M. R., Worrall, B. B., Silliman, S. L., Case, L. D., Rich, S. S., Meschia, J. F., Group, I. S. G. S., et al. (2007). Sex differences in stroke

severity, symptoms, and deficits after first-ever ischemic stroke. *Journal of Stroke and Cerebrovascular Diseases*, *16*(1), 34–39.

Bates, B. E., Xie, D., Kwong, P. L., Kurichi, J. E., Ripley, D. C., & Stineman, M. G. (2014). One-year all-cause mortality after stroke: A prediction model. *PM&R*, *6*(6), 473–483.

Begun, A., & Kulinskaya, E. (2022). A simulation study of the estimation quality in the double-cox model with shared frailty for non-proportional hazards survival analysis. *arXiv preprint arXiv:2206.05141*.

Begun, A., Kulinskaya, E., & Macgregor, A. (2019). Risk-adjusted cusum control charts for shared frailty survival models with application to hip replacement outcomes: A study using the njr dataset. *BMC Medical Research Methodology*.

Béjot, Y., Bailly, H., Graber, M., Garnier, L., Laville, A., Dubourget, L., Mielle, N., Chevalier, C., Durier, J., & Giroud, M. (2019). Impact of the ageing population on the burden of stroke: The dijon stroke registry. *Neuroepidemiology*, *52*(1-2), 78–85.

Bewick, V., Cheek, L., & Ball, J. (2004). Statistics review 12: Survival analysis. *Critical care*, *8*(5), 389.

Bhatnagar, P., Wickramasinghe, K., Williams, J., Rayner, M., & Townsend, N. (2015). The epidemiology of cardiovascular disease in the uk 2014. *Heart*, *101*(15), 1182–1189.

Blak, B., Thompson, M., Dattani, H., & Bourke, A. (2011). Generalisability of the health improvement network (thin) database: Demographics, chronic disease prevalence and mortality rates. *Journal of Innovation in Health Informatics*, *19*(4), 251–255.

Bodner, T. E. (2008). What improves with increased missing data imputations? *Structural equation modeling: a multidisciplinary journal*, *15*(4), 651–675.

Bonita, R., Anderson, C. S., Broad, J. B., Jamrozik, K. D., Stewart-Wynne, E. G., & Anderson, N. E. (1994). Stroke incidence and case fatality in australasia. a comparison of the auckland and perth population-based stroke registers. *Stroke*, *25*(3), 552–557.

Boysen, G., Marott, J. L., Grønbæk, M., Hassanpour, H., & Truelsen, T. (2009). Long-term survival after stroke: 30 years of follow-up in a cohort, the copenhagen city heart study. *Neuroepidemiology*, *33*(3), 254–260.

Bracard, S., Ducrocq, X., Mas, J. L., Soudant, M., Oppenheim, C., Moulin, T., Guillemin, F., et al. (2016). Mechanical thrombectomy after intravenous alteplase versus alteplase alone after stroke (thrace): A randomised controlled trial. *The Lancet Neurology*, *15*(11), 1138–1147.

Brand, J. (1999). *Development, implementation and evaluation of multiple imputation strategies for the statistical analysis of incomplete data sets*.

British Heart, F. (2019, March). *Four million people are living with untreated high blood pressure, new estimates show*. https://www.bhf.org.uk/what-we-do/news-from-the-bhf/news-archive/2019/may/four-million-people-are-living-with-untreated-high-blood-pressure

British Heart Foundation. (2020). How does my ethnicity affect my risk of heart and circulatory diseases? [Accessed: 2020-12-06].

Brønnum-Hansen, H., Davidsen, M., Thorvaldsen, P., et al. (2001). Long-term survival and causes of death after stroke. *Stroke*, *32*(9), 2131–2136.

Broström, G. (2012). *Event history analysis with r*. CRC Press.

Brown, H. A., Sullivan, M. C., Gusberg, R. G., Dardik, A., Sosa, J. A., & Indes, J. E. (2013). Race as a predictor of morbidity, mortality, and neurologic events after carotid endarterectomy. *Journal of vascular surgery*, *57*(5), 1325–1330.

Burn, J., Dennis, M., Bamford, J., Sandercock, P., Wade, D., & Warlow, C. (1994). Long-term risk of recurrent stroke after a first-ever stroke. the oxfordshire community stroke project. *Stroke*, *25*(2), 333–337.

Buuren, S. v., & Groothuis-Oudshoorn, K. (2010). Mice: Multivariate imputation by chained equations in r. *Journal of statistical software*, 1–68.

Cancer Research UK, C. (2020, March). *Cancer incidence statistics*. https://www.cancerresearchuk.org/health-professional/cancer-statistics/incidence

Carpenter, J., Pocock, S., & Johan Lamm, C. (2002). Coping with missing data in clinical trials: A model-based approach applied to asthma trials. *Statistics in medicine*, *21*(8), 1043–1066.

Carter, A. M., Catto, A. J., Mansfield, M. W., Bamford, J. M., & Grant, P. J. (2007). Predictive variables for mortality after acute ischemic stroke. *Stroke*, *38*(6), 1873–1880.

Cea-Soriano, L., Fowkes, F. G. R., Johansson, S., Allum, A. M., & Rodriguez, L. A. G. (2018). Time trends in peripheral artery disease incidence, prevalence and secondary preventive therapy: A cohort study in the health improvement network in the uk. *BMJ open*, *8*(1), e018184.

Chang, K.-C., Lee, H.-C., Tseng, M.-C., & Huang, Y.-C. (2010). Three-year survival after first-ever ischemic stroke is predicted by initial stroke severity: A hospital-based study. *Clinical neurology and neurosurgery*, *112*(4), 296–301.

Chang, K.-C., Tan, T.-Y., Liou, C.-W., & Tseng, M.-C. (2006). Predicting 3-month mortality among patients hospitalized for first-ever acute ischemic stroke. *Journal of the Formosan Medical Association*, *105*(4), 310–317.

Chutoo, P. (n.d.). Survival analysis after a first ischaemic stroke event: A case-control study in the adult population of england. *Methodology*, *2*, 1.

Chutoo, P., Kulinskaya, E., Bakbergenuly, I., Steel, N., Pchejetski, D., & Brown, B. (2022). Long term survival after a first transient ischaemic attack in england: A retrospective matched cohort study. *Journal of Stroke and Cerebrovascular Diseases*, *31*(9), 106663.

Clark, T., Murphy, M., & Rothwell, P. (2003). Long term risks of stroke, myocardial infarction, and vascular death in "low risk" patients with a non-recent transient ischaemic attack. *Journal of Neurology, Neurosurgery & Psychiatry*, *74*(5), 577–580.

Cole, J. A., Taylor, J. S., Hangartner, T. N., Weinreb, N. J., Mistry, P. K., & Khan, A. (2011). Reducing selection bias in case-control studies from rare disease registries. *Orphanet Journal of Rare Diseases*, *6*(1), 1–7.

Coull, A., Lovett, J., & Rothwell, P. (2004). Population based study of early risk of stroke after transient ischaemic attack or minor stroke: Implications for public education and organisation of services. *Bmj*, *328*(7435), 326.

Cox, D. R. (1972). Models and life-tables regression. *JR Stat. Soc. Ser. B*, *34*, 187–220.

Cumbler, E. (2015). In-hospital ischemic stroke. *The Neurohospitalist*, *5*(3), 173–181.

Custodio Martinez, R. L. M. (2007). Diagnostics for choosing between log-rank and wilcoxon tests.

Daffertshofer, M., Mielke, O., Pullwitt, A., Felsenstein, M., & Hennerici, M. (2004). Transient ischemic attacks are more than "ministrokes". *Stroke*, *35*(11), 2453–2458.

Daniel, K., Wolfe, C. D., Busch, M. A., & McKevitt, C. (2009). What are the social consequences of stroke for working-aged adults? a systematic review. *Stroke*, *40*(6), e431–e440.

Davis, R. C., Hobbs, F. R., Kenkre, J. E., Roalfe, A. K., Iles, R., Lip, G. Y., & Davies, M. K. (2012). Prevalence of atrial fibrillation in the general population and in high-risk groups: The echoes study. *Europace*, *14*(11), 1553–1559.

De Jong, G., Van Raak, L., Kessels, F., & Lodder, J. (2003). Stroke subtype and mortality: A follow-up study in 998 patients with a first cerebral infarct. *Journal of clinical epidemiology*, *56*(3), 262–268.

Denburg, M. R., Haynes, K., Shults, J., Lewis, J. D., & Leonard, M. B. (2011). Validation of the health improvement network (thin) database for epidemiologic studies of chronic kidney disease. *Pharmacoepidemiology and drug safety*, *20*(11), 1138–1149.

Dennis, M. S., Burn, J. P., Sandercock, P. A., Bamford, J. M., Wade, D. T., & Warlow, C. P. (1993). Long-term survival after first-ever stroke: The oxfordshire community stroke project. *Stroke*, *24*(6), 796–800.

Devin, I. (2019). Parametric survival modeling.

Dhamoon, M. S., Moon, Y. P., Paik, M. C., Boden-Albala, B., Rundek, T., Sacco, R. L., & Elkind, M. S. (2009). Long-term functional recovery after first ischemic stroke: The northern manhattan study. *Stroke*, *40*(8), 2805–2811.

Di Carlo, A., Lamassa, M., Baldereschi, M., Pracucci, G., Consoli, D., Wolfe, C. D., Giroud, M., Rudd, A., Burger, I., Ghetti, A., et al. (2006). Risk factors and outcome of subtypes of ischemic stroke. data from a multicenter multinational hospital-based registry. the european community stroke project. *Journal of the neurological sciences*, *244*(1), 143–150.

Diabetes. (2018, March). *Diabetes in uk*. https://www.diabetes.org.uk/in_your_area/south_west/regional-news/nearly-300000-people-south-west-have-diabetes

Doran, T., Kontopantelis, E., Reeves, D., Sutton, M., & Ryan, A. M. (2014). Setting performance targets in pay for performance programmes: What can we learn from qof? *Bmj*, *348*, g1595.

Douiri, A., Rudd, A. G., & Wolfe, C. D. (2013). Prevalence of poststroke cognitive impairment. *Stroke*, *44*(1), 138–145.

Dworzynski, K., Ritchie, G., Fenu, E., MacDermott, K., & Playford, E. D. (2013). Rehabilitation after stroke: Summary of nice guidance. *Bmj*, *346*.

Dziura, J. D., Post, L. A., Zhao, Q., Fu, Z., & Peduzzi, P. (2013). Strategies for dealing with missing data in clinical trials: From design to analysis. *The Yale journal of biology and medicine*, *86*(3), 343.

Easton, J. D., Saver, J. L., Albers, G. W., Alberts, M. J., Chaturvedi, S., Feldmann, E., Hatsukami, T. S., Higashida, R. T., Johnston, S. C., Kidwell, C. S., et al. (2009). Definition and evaluation of transient ischemic attack: A scientific statement for healthcare professionals from the american heart association/american stroke association stroke council; council on cardiovascular surgery and anesthesia; council on cardiovascular radiology and intervention; council on cardiovascular nursing; and the interdisciplinary council on peripheral vascular disease: The american academy of neurology affirms the value of this statement as an educational tool for neurologists. *Stroke*, *40*(6), 2276–2293.

Edwards, J. D., Kapral, M. K., Fang, J., & Swartz, R. H. (2017a). Long-term morbidity and mortality in patients without early complications after stroke or transient ischemic attack. *Cmaj*, *189*(29), E954–E961.

Edwards, J. D., Kapral, M. K., Fang, J., & Swartz, R. H. (2017b). Trends in long-term mortality and morbidity in patients with no early complications after stroke and transient ischemic attack. *Journal of Stroke and Cerebrovascular Diseases*, *26*(7), 1641–1645.

Edwardson, M., Dromerick, A., Kasner, S. E., & Dashe, J. F. (2016). Ischemic stroke prognosis in adults. *UpToDate, Official Reprint, Topic*, *14086*.

Efron, B. (1977). The efficiency of cox's likelihood function for censored data. *Journal of the American statistical Association*, *72*(359), 557–565.

Ellekjær, H., Holmen, J., Krüger, Ø., & Terent, A. (1999). Identification of incident stroke in norway: Hospital discharge data compared with a population-based stroke register. *Stroke*, *30*(1), 56–60.

Eriksson, Carlberg, B., & Eliasson, M. (2012). The disparity in long-term survival after a first stroke in patients with and without diabetes persists: The northern sweden monica study. *Cerebrovascular Diseases*, *34*(2), 153–160.

Eriksson & Olsson, J.-E. (2001). Survival and recurrent strokes in patients with different subtypes of stroke: A fourteen-year follow-up study. *Cerebrovascular diseases*, *12*(3), 171–180.

Eriksson, S.-E. (2017). Secondary prophylactic treatment and long-term prognosis after tia and different subtypes of stroke. a 25-year follow-up hospital-based observational study. *Brain and behavior*, *7*(1), e00603.

Falaschetti, E., Chaudhury, M., Mindell, J., & Poulter, N. (2009). Continued improvement in hypertension management in england: Results from the health survey for england 2006. *Hypertension*, *53*(3), 480–486.

Feigin, V. L., & Krishnamurthi, R. (2011). Stroke prevention in the developing world. *Stroke*, STROKEAHA–110.

Feigin, V. L., Lawes, C. M., Bennett, D. A., Barker-Collo, S. L., & Parag, V. (2009). Worldwide stroke incidence and early case fatality reported in 56 population-based studies: A systematic review. *The Lancet Neurology*, *8*(4), 355–369.

Fernandes, T. G., Goulart, A. C., Campos, T. F., Lucena, N. M., Freitas, K. L., Trevisan, C. M., Benseñor, I. M., & Lotufo, P. A. (2012). Early stroke case-fatality rates in three hospital registries in the northeast and southeast of brazil. *Arquivos de neuro-psiquiatria*, *70*(11), 869–873.

Fleming, T. R., & Harrington, D. P. (2011). *Counting processes and survival analysis* (Vol. 169). John Wiley & Sons.

Foundation, B. L. (2020, August). *Chronic obstructive pulmonary disease (copd) statistics*. https://statistics.blf.org.uk/copd

Fox & Weinsberg. (2011). Cox proportional hazards regression in survival data in **R** (C. Sage thousand Oaks, Trans.). *An Appneix to An R Companion to Applied Regression, Second Edtion*.

Garcıéa-Rodrıéguez, L. A., Gaist, D., Morton, J., Cookson, C., & González-Pérez, A. (2013). Antithrombotic drugs and risk of hemorrhagic stroke in the general population. *Neurology*, *81*(6), 566–574.

Gattellari, M., Goumas, C., Biost, F. G. M., & Worthington, J. M. (2012). Relative survival after transient ischaemic attack: Results from the program of research informing stroke management (prism) study. *Stroke*, *43*(1), 79–85.

Giles, M. F., & Rothwell, P. M. (2007). Substantial underestimation of the need for outpatient services for tia and minor stroke. *Age and ageing*, *36*(6), 676–680.

Gillen, D. (2016). *Proportional hazards regression diagnostics* (I. Department of Statistics University of California, Ed.). https://www.ics.uci.edu/~dgillen/STAT255/Handouts/lecture10.pdf

Goldstein, H., Carpenter, J., Kenward, M. G., & Levin, K. A. (2009). Multilevel models with multivariate mixed response types. *Statistical modelling*, *9*(3), 173–197.

Goldstein, L., Bushnell, C., Adams, R., Appel, L., Braun, L., Chaturvedi, S., Creager, M., Culebras, A., Eckel, R., Hart, R., et al. (2011). American heart association stroke council; council on cardiovascular nursing; council on epidemiology and prevention; council for high blood pressure research; council on peripheral vascular disease, and interdisciplinary council on quality of care and outcomes research. guidelines for the primary prevention of stroke. *Headache*, *51*(6), 1011–1021.

Gomes, F., Hookway, C., & Weekes, C. (2014). Royal college of physicians intercollegiate stroke working party evidence-based guidelines for the nutritional support of patients who have had a stroke. *Journal of human nutrition and dietetics*, *27*(2), 107–121.

Gompertz, B. (1825). Xxiv. on the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies. in a letter to francis baily, esq. frs &c. *Philosophical transactions of the Royal Society of London*, *115*, 513–583.

Gonzalez-Perez, A., Gaist, D., Wallander, M.-A., McFeat, G., & Garca-Rodriguez, L. A. (2013). Mortality after hemorrhagic stroke: Data from general practice (the health improvement network). *Neurology*, *81*(6), 559–565.

Gorelick, P. B. (2019). The global burden of stroke: Persistent and disabling. *The Lancet Neurology*, *18*(5), 417–418.

Graham, J. W., et al. (2009). Missing data analysis: Making it work in the real world. *Annual review of psychology*, *60*(1), 549–576.

Grambsch, P. M., & Therneau, T. M. (1994). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, *81*(3), 515–526.

Greenwood, M. (1926). The" errors of sampling" of the survivorship tables. *Reports on public health and medical subjects*.

Gresham, G. E., Kelly-Hayes, M., Wolf, P. A., Beiser, A. S., Kase, C. S., & D'Agostino, R. B. (1998). Survival and functional status 20 or more years after first stroke: The framingham study. *Stroke*, *29*(4), 793–797.

Group, U.-T. S., et al. (1988). United kingdom transient ischaemic attack (uk-tia) aspirin trial: Interim results. *Br Med J (Clin Res Ed)*, *296*(6618), 316–320.

Hackam, D., Kapral, M., Wang, J., Fang, J., & Hachinski, V. (2009). Most stroke patients do not get a warning: A population-based cohort study. *Neurology*, *73*(13), 1074–1076.

Hagberg, G., Fure, B., Sandset, E. C., Thommessen, B., Ihle-Hansen, H., Øksengård, A. R., Nygård, S., Wyller, T. B., & Ihle-Hansen, H. (2019). Long-term effects on survival after a 1-year multifactorial vascular risk factor intervention after stroke or tia: Secondary analysis of a randomized controlled trial, a 7-year follow-up study. *Vascular health and risk management*, *15*, 11.

Hall, G. C. (2009). Validation of death and suicide recording on the thin uk primary care database. *Pharmacoepidemiology and drug safety*, *18*(2), 120–131.

Hanagal, D. D. (2011). *Modelling survival data using frailty models*. Chapman; Hall/CRC.

Hankey. (2003). Long-term outcome after ischaemic stroke/transient ischaemic attack. *Cerebrovascular diseases*, *16*(Suppl. 1), 14–19.

Hankey, G., Spiesser, J., Hakimi, Z., Bego, G., Carita, P., & Gabriel, S. (2007). Rate, degree, and predictors of recovery from disability following ischemic stroke. *Neurology*, *68*(19), 1583–1587.

Hankey, G. J. (2007). Clinical update: Management of stroke. *The Lancet*, *369*(9570), 1330.

Hankey, G. J., & Warlow, C. P. (1999). Treatment and secondary prevention of stroke: Evidence, costs, and effects on individuals and populations. *The Lancet*, *354*(9188), 1457–1463.

Hankey, Slattery, J. M., & Warlow, C. P. (1991). The prognosis of hospital-referred transient ischaemic attacks. *Journal of Neurology, Neurosurgery & Psychiatry*, *54*(9), 793–802.

Hannerz, H., & Nielsen, M. L. (2001). Life expectancies among survivors of acute cerebrovascular disease. *Stroke*, *32*(8), 1739–1744.

HannoverRE. (2018). Longevity risk [Last accessed 09 July 2021]. https://www.hannover-re.com/371848/longevity-risk-2018.pdf

Hardie, K., Hankey, G. J., Jamrozik, K., Broadhurst, R. J., & Anderson, C. (2003). Ten-year survival after first-ever stroke in the perth community stroke study. *Stroke*, *34*(8), 1842–1846.

Harrell, F. E., et al. (2001). *Regression modeling strategies: With applications to linear models, logistic regression, and survival analysis* (Vol. 608). Springer.

Harrington, D. P., & Fleming, T. R. (1982). A class of rank test procedures for censored survival data. *Biometrika*, *69*(3), 553–566.

Hartmann, A., Rundek, T., Mast, H., Paik, M., Boden–Albala, B., Mohr, J., & Sacco, R. (2001). Mortality and causes of death after first ischemic stroke: The northern manhattan stroke study. *Neurology*, *57*(11), 2000–2005.

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (Vol. 2). Springer.

Heneghan, C., Goldacre, B., & Mahtani, K. R. (2017). Why clinical trial outcomes fail to translate into benefits for patients. *Trials*, *18*(1), 1–7.

Hillborn, M. (1998). Alcohol consumption and stroke: Benefits and risks. *Alcoholism: Clinical and Experimental Research*, *22*(S7).

Hippisley-Cox, J., Coupland, C., Vinogradova, Y., Robson, J., May, M., & Brindle, P. (2007). Derivation and validation of qrisk, a new cardiovascular disease risk score for the united kingdom: Prospective open cohort study. *Bmj*, *335*(7611), 136.

Hosmer, D. W., & Lemeshow, S. (1999). Applied survival analysis: Regression modelling of time to event data (1999). *Eur Orthodontic Soc*, 561–2.

Hosmer, D. (2000). Exact methods for logistic regression models. in. hosmer dw, lemeshow s. applied logistic regression.

Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (Vol. 398). John Wiley & Sons.

IMD. (2014). *National statistics english indices of deprivation 2019*. https://www.gov.uk/government/statistics/english-indices-of-deprivation-2019

Ingeman, A., Andersen, G., Hundborg, H. H., Svendsen, M. L., & Johnsen, S. P. (2011). In-hospital medical complications, length of stay, and mortality among stroke unit patients. *Stroke*, *42*(11), 3214–3218.

Jacob, L., Tanislav, C., & Kostev, K. (2020). Long-term risk of stroke and its predictors in transient ischaemic attack patients in germany. *European journal of neurology*, *27*(4), 723–728.

Jeng, J.-S., Huang, S.-J., Tang, S.-C., & Yip, P.-K. (2008). Predictors of survival and functional outcome in acute stroke patients admitted to the stroke intensive care unit. *Journal of the neurological sciences*, *270*(1), 60–66.

Johnston, Gress, D. R., Browner, W. S., & Sidney, S. (2000). Short-term prognosis after emergency department diagnosis of tia. *Jama*, *284*(22), 2901–2906.

Johnston, Morrison, V., Macwalter, R., & Partridge, C. (1999). Perceived control, coping and recovery from disability following stroke. *Psychology and health*, *14*(2), 181–192.

Jorgensen, H. S., Nakayama, H., Raaschou, H. O., & Olsen, T. S. (1995). Recovery of walking function in stroke patients: The copenhagen stroke study. *Archives of physical medicine and rehabilitation*, *76*(1), 27–32.

Kalbfleisch, J. D., & Prentice, R. L. (2011). The statistical analysis of failure time data. *360*.

Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, *53*(282), 457–481.

Kennedy, J., Hill, M. D., Ryckborst, K. J., Eliasziw, M., Demchuk, A. M., Buchan, A. M., Investigators, F., et al. (2007). Fast assessment of stroke and transient ischaemic attack to prevent early recurrence (faster): A randomised controlled pilot trial. *The Lancet Neurology*, *6*(11), 961–969.

Khare, S. (2016). Risk factors of transient ischemic attack: An overview. *Journal of mid-life health*, *7*(1), 2.

Kifle, Y. G. (2013). *Modeling the effect of distance from a hydro-electric dam on malaria incidence based on frailty and mixed poisson regression models* (Doctoral dissertation). Ghent University.

Kim, A. S., & Johnston, S. C. (2011). Global variation in the relative burden of stroke and ischemic heart disease. *Circulation*, CIRCULATIONAHA–111.

King, D., Wittenberg, R., Patel, A., Quayyum, Z., Berdunov, V., & Knapp, M. (2020). The future incidence, prevalence and costs of stroke in the uk. *Age and ageing*, *49*(2), 277–282.

Kissela, B., Lindsell, C. J., Kleindorfer, D., Alwell, K., Moomaw, C. J., Woo, D., Flaherty, M. L., Air, E., Broderick, J., & Tsevat, J. (2009). Clinical prediction of functional outcome after ischemic stroke. *Stroke*, *40*(2), 530–536.

Kolominsky-Rabas, P. L., Weber, M., Gefeller, O., Neundoerfer, B., & Heuschmann, P. U. (2001). Epidemiology of ischemic stroke subtypes according to toast criteria: Incidence, recurrence, and long-term survival in ischemic stroke subtypes: A population-based study. *Stroke*, *32*(12), 2735–2740.

Kremers, W. K. (2007). Concordance for survival time data: Fixed and time-dependent covariates and possible ties in predictor and time. *Mayo Foundation*.

Krishnamurthi, R. V., Feigin, V. L., Forouzanfar, M. H., Mensah, G. A., Connor, M., Bennett, D. A., Moran, A. E., Sacco, R. L., Anderson, L. M., Truelsen, T., et al. (2013). Global and regional burden of first-ever ischaemic and haemorrhagic stroke during 1990–2010: Findings from the global burden of disease study 2010. *The Lancet Global Health*, *1*(5), e259–e281.

Larsen, T. B., & Lip, G. Y. (2014). Warfarin or novel oral anticoagulants for atrial fibrillation? *The Lancet*, *383*(9921), 931–933.

Lawes, C. M., Bennett, D. A., Feigin, V. L., & Rodgers, A. (2004). Blood pressure and stroke: An overview of published reviews. *Stroke*, *35*(3), 776–785.

Lawes, C. M., Vander Hoorn, S., Rodgers, A., et al. (2008). Global burden of blood-pressure-related disease, 2001. *The Lancet*, *371*(9623), 1513–1518.

Lee, S., Shafe, A. C., & 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.

Lewis, J. D., Schinnar, R., Bilker, W. B., Wang, X., & Strom, B. L. (2007). Validation studies of the health improvement network (thin) database for pharmacoepidemiology research. *Pharmacoepidemiology and drug safety*, *16*(4), 393–401.

Liao, D., Myers, R., Hunt, S., Shahar, E., Paton, C., Burke, G., Province, M., & Heiss, G. (1997). Familial history of stroke and stroke risk: The family heart study. *Stroke*, *28*(10), 1908–1912.

Lichtman, J. H., Jones, S. B., Watanabe, E., Allen, N. B., Wang, Y., Howard, V. J., & Goldstein, L. B. (2009). Elderly women have lower rates of stroke, cardiovascular events, and mortality after hospitalization for transient ischemic attack. *Stroke*, *40*(6), 2116–2122.

Little, R., Rubin, D., Little, R., et al. (2002). Missing data in experiments statistical analysis with missing data. hoboken.

Little, R. J., & Rubin, D. (1987). Statistical analysis with incomplete data. *New York*.

Lloyd-Jones, D., Adams, R., Brown, T., Carnethon, M., Dai, S., De Simone, G., Ferguson, T., Ford, E., Furie, K., Gillespie, C., et al. (2010). Heart disease and stroke statistics–2010 update: A report from the american heart association. *Circulation*, *121*(7), e46.

Machin, D., Cheung, Y. B., & Parmar, M. (2006). *Survival analysis: A practical approach*. John Wiley & Sons.

Mack, J. (2020, March). *Deprivation and poverty,pse*. https://www.poverty.ac.uk/definitions-poverty/deprivation-and-poverty

Maheswaran, R., Strong, M., Clifford, P., & Brewins, L. (2018). Socioeconomic deprivation, mortality and health of within-city migrants: A population cohort study. *J Epidemiol Community Health*, *72*(6), 519–525.

Mahoney, F. I., et al. (1965). Functional evaluation: The barthel index. *Maryland state medical journal*, *14*(2), 61–65.

Makeham, W. M. (1860). On the law of mortality and construction of annuity tables. *Journal of the Institute of Actuaries*, *8*(6), 301–310.

Manning, N. W., Chapot, R., & Meyers, P. M. (2016). Endovascular stroke management: Key elements of success. *Cerebrovascular Diseases*, *42*(3-4), 170–177.

Marston, L., Carpenter, J. R., Walters, K. R., Morris, R. W., Nazareth, I., & Petersen, I. (2010). Issues in multiple imputation of missing data for large general practice clinical databases. *Pharmacoepidemiology and drug safety*, *19*(6), 618–626.

Martinussen, T., & Scheike, T. H. (2007). *Dynamic regression models for survival data*. Springer Science & Business Media.

Mayer, S. A., & Rincon, F. (2005). Treatment of intracerebral haemorrhage. *The Lancet Neurology*, *4*(10), 662–672.

McBride, P. E. (1992). The health consequences of smoking: Cardiovascular diseases. *Medical Clinics of North America*, *76*(2), 333–353.

McNeil, J. J., Woods, R. L., Nelson, M. R., Reid, C. M., Kirpach, B., Wolfe, R., Storey, E., Shah, R. C., Lockery, J. E., Tonkin, A. M., et al. (2018). Effect of aspirin on disability-free survival in the healthy elderly. *New England Journal of Medicine*, *379*(16), 1499–1508.

Medicover. (2021). *What is the life expectancy after a brain stroke?* https://www.medicoverhospitals.in/what-is-the-life-expectancy-after-a-brain-stroke/ (accessed: 01.09.2021)

Members, A. F., Camm, A. J., Lip, G. Y., De Caterina, R., Savelieva, I., Atar, D., Hohnloser, S. H., Hindricks, G., Kirchhof, P., for Practice Guidelines (CPG), E. C., et al. (2012). 2012 focused update of the esc guidelines for the management of atrial fibrillation: An update of the 2010 esc guidelines for the management of atrial fibrillation developed with the special contribution of the european heart rhythm association. *European heart journal*, *33*(21), 2719–2747.

Meretoja, A., Kaste, M., Roine, R. O., Juntunen, M., Linna, M., Hillbom, M., Marttila, R., Erilä, T., Rissanen, A., Sivenius, J., et al. (2011). Trends in treatment and outcome of stroke patients in finland from 1999 to 2007. perfect stroke, a nationwide register study. *Annals of medicine*, *43*(sup1), S22–S30.

Millikan, C., CH, M., RB, B., et al. (1975). A classification and outline of cerebrovascular diseases ii.

Modrego, P. J., Mainar, R., & Turull, L. (2004). Recurrence and survival after first-ever stroke in the area of bajo aragon, spain. a prospective cohort study. *Journal of the neurological sciences*, *224*(1), 49–55.

Mogensen, U. B., Olsen, T. S., Andersen, K. K., & 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., & Kenward, M. (2007). *Missing data in clinical studies* (Vol. 61). John Wiley & Sons.

Moran, G. M., Calvert, M., Feltham, M. G., Ryan, R., & Marshall, T. (2015). A retrospective cohort study to investigate fatigue, psychological or cognitive impairment after tia: Protocol paper. *BMJ open*, *5*(4), e008149.

Mosaic. (2014). Mosaic the consumer classification solution for consistent cross-channel marketing, the experian. https://www.experianintact.com/content/uk/documents/productSheets/MosaicConsumerUK.pdf

Mukamal, K. J., Ascherio, A., Mittleman, M. A., Conigrave, K. M., Camargo, C. A., Kawachi, I., Stampfer, M. J., Willett, W. C., & Rimm, E. B. (2005). Alcohol and risk for ischemic stroke

in men: The role of drinking patterns and usual beverage. *Annals of internal medicine*, *142*(1), 11–19.

Mukherjee, D., & Patil, C. G. (2011). Epidemiology and the global burden of stroke. *World neurosurgery*, *76*(6), S85–S90.

Murphy, S. J., & Werring, D. J. (2020). Stroke: Causes and clinical features. *Medicine*.

Murray, C., & Lopez, A. D. (1997). The utility of dalys for public health policy and research: A reply. *Bulletin of the World Health Organization*, *75*(4), 377.

National Collaborating Centre for Chronic Conditions, N., et al. (2006). Atrial fibrillation: National clinical guideline for management in primary and secondary care.

National Institute of Neurological Disorders. (2021). Troke: Hope through research [Accessed: 2020-12-06].

NHS Choices, N. (2017). Stroke-act fast.

NICE. (2008). Stroke: National clinical guideline for diagnosis and initial management of acute stroke and transient ischaemic attack (tia).

NICE. (2010). Clopidogrel and modified-release dipyridamole for the prevention of occlusive vascular events,technology appraisal guidance [ta210] [Accessed: 2021-03-121].

NICE. (2019). Stroke and transient ischaemic attack in over 16s: diagnosis and initial management [[Online; accessed 19-July-2019]].

NICE. (2020a). Scenario of secondary prevention following a stroke or tia [Accessed on 2021-03-04].

NICE. (2020b). Stroke and tia management [Accessed: 2020-12-06].

NICE. (2021). Scenario of secondary prevention of cvd, management of antiplatelet treatment following stroke and tia , clinical knowledge summaries cks [Accessed: 2021-03-121].

NICE. (2022, March). *Hypertension in adults: Diagnosis and management*. https://www.nice.org.uk/guidance/ng136/chapter/Context

Nichols, M., Townsend, N., Scarborough, P., & Rayner, M. (2014). Cardiovascular disease in europe 2014: Epidemiological update. *European heart journal*, *35*(42), 2950–2959.

Nogueira, R. G., Liebeskind, D. S., Sung, G., Duckwiler, G., Smith, W. S., et al. (2009). Predictors of good clinical outcomes, mortality, and successful revascularization in patients with acute ischemic stroke undergoing thrombectomy. *Stroke*, *40*(12), 3777–3783.

Palm, F., Kraus, M., Safer, A., Wolf, J., Becher, H., & Grau, A. J. (2014). Management of oral anticoagulation after cardioembolic stroke and stroke survival data from a population based stroke registry (lusst). *BMC neurology*, *14*(1), 199.

Patel, A., Berdunov, V., Quayyum, Z., King, D., Knapp, M., & Wittenberg, R. (2020). Estimated societal costs of stroke in the uk based on a discrete event simulation. *Age and ageing*, *49*(2), 270–276.

Perry Jr, H. M., Davis, B. R., Price, T. R., Applegate, W. B., Fields, W. S., Guralnik, J. M., Kuller, L., Pressel, S., Stamler, J., Probstfield, J. L., et al. (2000). Effect of treating isolated systolic

hypertension on the risk of developing various types and subtypes of stroke: The systolic hypertension in the elderly program (shep). *Jama*, *284*(4), 465–471.

Phipson, B. (2006). *Analysis of time-to-event data including frailty modeling.* (Doctoral dissertation).

Pohjasvaara, T., Vataja, R., Leppävuori, A., Kaste, M., & Erkinjuntti, T. (2002). Cognitive functions and depression as predictors of poor outcome 15 months after stroke. *Cerebrovascular Diseases*, *14*(3-4), 228–233.

Price, M. H., & Jones, J. H. (2017). A general goodness-of-fit test for survival analysis. *bioRxiv*, 104406.

PROGRESS. (1999). Progress–perindopril protection against recurrent stroke study†: Characteristics of the study population at baseline. *Journal of hypertension*, *17*(11), 1647–1655.

Protty, M. B., Wilkins, S. J., Hoskins, H. C., Dawood, B. B., & Hayes, J. (2018). Prescribing patterns of oral antiplatelets in wales: Evolving trends from 2005 to 2016. *Future cardiology*, *14*(4), 277–282.

Public Health England, P. (2016). *Diabetes prevalence model* [Last accessed 06 August 2021]. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/612306/Diabetesprevalencemodelbriefing.pdf

Public Health England, P. (2017). Longevity risk [Last accessed 09 July 2021]. https://www.gov.uk/government/publications/health-profile-for-england-2018/chapter-1-population-change-and-trends-in-life-expectancy

Public Health England, P. (2018). First incidence of stroke estimates for england 2007 to 2016 [Last accessed 09 July 2021]. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/678444/Stroke_incidence_briefing_document_2018.pdf

Public Health England, P. (2019, March). *Chronic kidney disease (ckd) prevalence model - gov.uk.* https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/612303/ChronickidneydiseaseCKDprevalencemodelbriefing.pdf

Quartagno, M., & Carpenter, J. (2016). Jomo: A package for multilevel joint modelling multiple imputation. *R package version*, *2*(0).

Quartagno, M., Grund, S., & Carpenter, J. (2019). Jomo: A flexible package for two-level joint modelling multiple imputation. *R Journal*, *9*(1).

Rahman, F., Kwan, G. F., & Benjamin, E. J. (2014). Global epidemiology of atrial fibrillation. *Nature Reviews Cardiology*, *11*(11), 639–654.

Robert, S., Jordan, B., David, S., & Amytis, T. (2021). *Long-term survival prognosis after stroke a practical guide for clinicians.* https://practicalneurology.com/articles/2020-feb/long-term-survival-prognosis-after-stroke (accessed: 01.09.2021)

Rodrıguez, G. (2005). Non-parametric estimation in survival models. https://data.princeton.edu/pop509/NonParametricSurvival.pdf

Rønning, O. M., & Guldvog, B. (1998). Stroke unit versus general medical wards, ii: Neurological deficits and activities of daily living. *Stroke*, *29*(3), 586–590.

Roth, G. A., Mensah, G. A., Johnson, C. O., Addolorato, G., Ammirati, E., Baddour, L. M., Barengo, N. C., Beaton, A. Z., Benjamin, E. J., Benziger, C. P., et al. (2020). Global burden of cardiovascular diseases and risk factors, 1990–2019: Update from the gbd 2019 study. *Journal of the American College of Cardiology*, *76*(25), 2982–3021.

Rothwell, P. M., Coull, A., Giles, M., Howard, S., Silver, L., Bull, L., Gutnikov, S., Edwards, P., Mant, D., Sackley, C., et al. (2004). Change in stroke incidence, mortality, case-fatality, severity, and risk factors in oxfordshire, uk from 1981 to 2004 (oxford vascular study). *The Lancet*, *363*(9425), 1925–1933.

Rothwell, P. M., Giles, M. F., Chandratheva, A., Marquardt, L., Geraghty, O., Redgrave, J. N., Lovelock, C. E., Binney, L. E., Bull, L. M., Cuthbertson, F. C., et al. (2007). Effect of urgent treatment of transient ischaemic attack and minor stroke on early recurrent stroke (express study): A prospective population-based sequential comparison. *The Lancet*, *370*(9596), 1432–1442.

Royston, P. (2004). Multiple imputation of missing values. *The Stata Journal*, *4*(3), 227–241.

Rubin, D. B. (1987). Multiple imputation for nonresponse in surveys. hoboken.

Rutten-Jacobs, L. C., Arntz, R. M., Maaijwee, N. A., Schoonderwaldt, H. C., Dorresteijn, L. D., van Dijk, E. J., & de Leeuw, F.-E. (2013). Long-term mortality after stroke among adults aged 18 to 50 years. *Jama*, *309*(11), 1136–1144.

Sacco, R. L., Benjamin, E. J., Broderick, J. P., Dyken, M., Easton, J. D., Feinberg, W. M., Goldstein, L. B., Gorelick, P. B., Howard, G., Kittner, S. J., et al. (1997). Risk factors. *Stroke*, *28*(7), 1507–1517.

Sacco, R. L., Shi, T., Zamanillo, M., & Kargman, D. (1994). Predictors of mortality and recurrence after hospitalized cerebral infarction in an urban community: The northern manhattan stroke study. *Neurology*, *44*(4), 626–626.

Sacco, R. L., Wolf, P. A., Kannel, W., & McNamara, P. (1982). Survival and recurrence following stroke. the framingham study. *Stroke*, *13*(3), 290–295.

Saka, Ö., McGuire, A., & Wolfe, C. (2009). Cost of stroke in the united kingdom. *Age and ageing*, *38*(1), 27–32.

Saltman, A. P., Silver, F. L., Fang, J., Stamplecoski, M., & Kapral, M. K. (2015). Care and outcomes of patients with in-hospital stroke. *JAMA neurology*, *72*(7), 749–755.

Sarti, C., Rastenyte, D., Cepaitis, Z., & Tuomilehto, J. (2000). International trends in mortality from stroke, 1968 to 1994. *Stroke*, *31*(7), 1588–1601.

Saxena, S., Car, J., Eldred, D., Soljak, M., & Majeed, A. (2007). Practice size, caseload, deprivation and quality of care of patients with coronary heart disease, hypertension and stroke in primary care: National cross-sectional study. *BMC health services research*, *7*(1), 1–9.

Schafer, J., Ezzati-Rice, T., Johnson, W., Khare, M., Little, R., & Rubin, D. (1996). The nhanes iii multiple imputation project. *Race/ethnicity*, *60*(21.2), 15–5.

Schafer, J. L. (1999). Multiple imputation: A primer. *Statistical methods in medical research*, *8*(1), 3–15.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological methods*, *7*(2), 147.

Schafer, J. L., & Olsen, M. K. (1998). Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate behavioral research*, *33*(4), 545–571.

Scheike, T. H., & Martinussen, T. (2004). On estimation and tests of time-varying effects in the proportional hazards model. *Scandinavian Journal of Statistics*, *31*(1), 51–62.

Scheike, T. H., & Zhang, M.-J. (2008). Flexible competing risks regression modeling and goodness-of-fit. *Lifetime data analysis*, *14*(4), 464–483.

Scheike, T. H., & Zhang, M.-J. (2011). Analyzing competing risk data using the r timereg package. *Journal of statistical software*, *38*(2).

Schoenfeld, D. (1982). Partial residuals for the proportional hazards regression model. *Biometrika*, *69*(1), 239–241.

Sharma, A., Lewis, S., & Szatkowski, L. (2010). Insights into social disparities in smoking prevalence using mosaic, a novel measure of socioeconomic status: An analysis using a large primary care dataset. *BMC Public Health*, *10*(1), 1–7.

Shavelle, R. M., Brooks, J. C., Strauss, D. J., & Turner-Stokes, L. (2019). Life expectancy after stroke based on age, sex, and rankin grade of disability: A synthesis. *Journal of Stroke and Cerebrovascular Diseases*, *28*(12), 104450.

Siegler, J. E., & Martin-Schild, S. (2011). Early neurological deterioration (end) after stroke: The end depends on the definition. *International Journal of Stroke*, *6*(3), 211–212.

Slot, K. B., Berge, E., Dorman, P., Lewis, S., Dennis, M., & Sandercock, P. (2008). Impact of functional status at six months on long term survival in patients with ischaemic stroke: Prospective cohort studies. *Bmj*, *336*(7640), 376–379.

Sommerfeld, D. K., Eek, E. U.-B., Svensson, A.-K., Holmqvist, L. W., & von Arbin, M. H. (2004). Spasticity after stroke. *Stroke*, *35*(1), 134–139.

SPARCL et al. (2006). Stroke prevention by aggressive reduction in cholesterol levels :high-dose atorvastatin after stroke or transient ischemic attack. *N engl J med*, *2006*(355), 549–559.

Stroke Association, S. O. (2016). *State of the nation stroke statistics*. https://www.stroke.org.uk/sites/default/files/stroke_statistics_2015.pdf

Taggar, J. S., Coleman, T., Lewis, S., & Szatkowski, L. (2012). The impact of the quality and outcomes framework (qof) on the recording of smoking targets in primary care medical records: Cross-sectional analyses from the health improvement network (thin) database. *BMC public health*, *12*(1), 1–11.

Tanny, S. P. T., Busija, L., Liew, D., Teo, S., Davis, S. M., & Yan, B. (2013). Cost-effectiveness of thrombolysis within 4.5 hours of acute ischemic stroke. *Stroke*, *44*(8), 2269–2274.

Taylor, F., Ward, K., Moore, T. H., Burke, M., Smith, G. D., Casas, J. P., & Ebrahim, S. (2011). Statins for the primary prevention of cardiovascular disease. *Cochrane database of systematic reviews*, (1).

Tei, H., Uchiyama, S., Ohara, K., Kobayashi, M., Uchiyama, Y., & Fukuzawa, M. (2000). Deteriorating ischemic stroke in 4 clinical categories classified by the oxfordshire community stroke project. *Stroke*, *31*(9), 2049–2054.

Therneau, T. M., & Grambsch, P. M. (2000). Testing proportional hazards. In *Modeling survival data: Extending the cox model* (pp. 127–152). Springer.

TNIoNDa, S. (1995). Tissue plasminogen activator for acute ischemic stroke. the national institute of neurological disorders and stroke rt-pa stroke study group. *The New England Journal of Medicine*, *333*(24), 1581–1587.

Townsend. (1979). *Poverty in the united kingdom: A survey of household resources and standards of living*. Univ of California Press.

Townsend. (2017). *Uk data service | census data 2011 uk townsend deprivation scores*. https://statistics.ukdataservice.ac.uk/dataset/2011-uk-townsend-deprivation-scores

Townsend, Wickramasinghe, K., Bhatnagar, P., Smolina, K., Nichols, M., Leal, J., Luengo-Fernandez, R., & Rayner, M. (2012). Coronary heart disease statistics 2012.

Truelsen, T., Piechowski-Jóźwiak, B., Bonita, R., Mathers, C., Bogousslavsky, J., & Boysen, G. (2006). Stroke incidence and prevalence in europe: A review of available data. *European journal of neurology*, *13*(6), 581–598.

Truelsen, T., Mähönen, M., Tolonen, H., Asplund, K., Bonita, R., & Vanuzzo, D. (2003). Trends in stroke and coronary heart disease in the who monica project. *Stroke*, *34*(6), 1346–1352.

Tseng, M.-C., & Chang, K.-C. (2006). Stroke severity and early recovery after first-ever ischemic stroke: Results of a hospital-based study in taiwan. *Health Policy*, *79*(1), 73–78.

Turner, G. M. (2016). *Missed opportunities for primary prevention of stroke and transient ischaemic attack (tia) and residual impairments after tia* (Doctoral dissertation). University of Birmingham.

UK, H. (2020, August). *Cholesterol facts and figures*. https://www.heartuk.org.uk/about-us/press-office

Uretsky, S., Messerli, F. H., Bangalore, S., Champion, A., Cooper-DeHoff, R. M., Zhou, Q., & Pepine, C. J. (2007). Obesity paradox in patients with hypertension and coronary artery disease. *The American journal of medicine*, *120*(10), 863–870.

Van Buuren, S. (2018). *Flexible imputation of missing data*. CRC press.

Van den Berg, L. A., Dijkgraaf, M. G., Berkhemer, O. A., Fransen, P. S., Beumer, D., Lingsma, H. F., Majoie, C. B., Dippel, D. W., van der Lugt, A., van Oostenbrugge, R. J., et al. (2017).

Two-year outcome after endovascular treatment for acute ischemic stroke. *New England Journal of Medicine*, *376*(14), 1341–1349.

Van Noorden, R., Maher, B., & Nuzzo, R. (2014). The top 100 papers. *Nature News*, *514*(7524), 550.

Vanderpump, M. P. (2011). The epidemiology of thyroid disease. *British medical bulletin*, *99*(1).

van Wijk, I., Kappelle, L., van Gijn, J., Koudstaal, P., Franke, C., Vermeulen, M., Gorter, J. W., Algra, A., Group, L. S., et al. (2005). Long-term survival and vascular event risk after transient ischaemic attack or minor ischaemic stroke: A cohort study. *The Lancet*, *365*(9477), 2098–2104.

Vasudevan, A. R., Burns, A., & Fonseca, V. A. (2006). The effectiveness of intensive glycemic control for the prevention of vascular complications in diabetes mellitus. *Treatments in endocrinology*, *5*(5), 273–286.

Vaupel, J. W., Manton, K. G., & Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, *16*(3), 439–454.

Vidyanti, A. N., Chan, L., Lin, C.-L., Muo, C.-H., Hsu, C. Y., Chen, Y.-C., Wu, D., & Hu, C.-J. (2019). Aspirin better than clopidogrel on major adverse cardiovascular events reduction after ischemic stroke: A retrospective nationwide cohort study. *PloS one*, *14*(8), e0221750.

Von Hippel, P. T. (2020). How many imputations do you need? a two-stage calculation using a quadratic rule. *Sociological Methods & Research*, *49*(3), 699–718.

Wang, Y., Lim, L. L.-Y., Heller, R. F., Fisher, J., & Levi, C. R. (2003). A prediction model of 1-year mortality for acute ischemic stroke patients 1, 2. *Archives of physical medicine and rehabilitation*, *84*(7), 1006–1011.

Wang, Y., Rudd, A. G., & Wolfe, C. D. (2013). Trends and survival between ethnic groups after stroke: The south london stroke register. *Stroke*, *44*(2), 380–387.

Wang, Y., Pan, Y., Zhao, X., Li, H., Wang, D., Johnston, S. C., Liu, L., Meng, X., Wang, A., Wang, C., et al. (2015). Clopidogrel with aspirin in acute minor stroke or transient ischemic attack (chance) trial: One-year outcomes. *Circulation*, *132*(1), 40–46.

Waziry, R., Heshmatollah, A., Bos, D., Chibnik, L. B., Ikram, M. A., Hofman, A., & Ikram, M. K. (2020). Time trends in survival following first hemorrhagic or ischemic stroke between 1991 and 2015 in the rotterdam study. *Stroke*, *51*(3), 824–829.

Weimar, C., Ziegler, A., König, I. R., Diener, H.-C., et al. (2002). Predicting functional outcome and survival after acute ischemic stroke. *Journal of neurology*, *249*(7), 888–895.

Weir, C. J., Murray, G. D., Dyker, A. G., & Lees, K. R. (1997). Is hyperglycaemia an independent predictor of poor outcome after acute stroke? results of a long term follow up study. *Bmj*, *314*(7090), 1303.

WHO, W. M. (1988). The world health organization monica project (monitoring trends and determinants in cardiovascular disease): A major international collaboration. *Journal of clinical epidemiology*, *41*(2), 105–114.

Wolfe, C. D., Smeeton, N. C., Coshall, C., Tilling, K., & Rudd, A. G. (2005). Survival differences after stroke in a multiethnic population: Follow-up study with the south london stroke register. *Bmj*, *331*(7514), 431.

Wood, A. M., White, I. R., & Royston, P. (2008). How should variable selection be performed with multiply imputed data? *Statistics in medicine*, *27*(17), 3227–3246.

World health Organisation, W. (2020). *Body mass index - bmi*. https://www.euro.who.int/en/health-topics/disease-prevention/nutrition/a-healthy-lifestyle/body-mass-index-bmi

Yang, X., Belin, T. R., & Boscardin, W. J. (2005). Imputation and variable selection in linear regression models with missing covariates. *Biometrics*, *61*(2), 498–506.

Yao, X., Abraham, N. S., Alexander, G. C., Crown, W., Montori, V. M., Sangaralingham, L. R., Gersh, B. J., Shah, N. D., & Noseworthy, P. A. (2016). Effect of adherence to oral anticoagulants on risk of stroke and major bleeding among patients with atrial fibrillation. *Journal of the American Heart Association*, *5*(2), e003074.

Yeatts, S. D., & Martin, R. H. (2015). What is missing from my missing data plan? *Stroke*, *46*(6), e130–e132.

Zhang, M., Peng, P., Gu, K., Cai, H., Qin, G., Shu, X. O., & Bao, P. (2018). Time-varying effects of prognostic factors associated with long-term survival in breast cancer. *Endocrine-related cancer*, *25*(5), 509–521.

Zhang, Q., Wang, C., Zheng, M., Li, Y., Li, J., Zhang, L., Shang, X., & Yan, C. (2015). Aspirin plus clopidogrel as secondary prevention after stroke or transient ischemic attack: A systematic review and meta-analysis. *Cerebrovascular Diseases*, *39*(1), 13–22.

Zhou, X., & Reiter, J. P. (2010). A note on bayesian inference after multiple imputation. *The American Statistician*, *64*(2), 159–163.