Prognostic Factors for Epilepsy

Thesis submitted in accordance with the requirements of the University of Liverpool for the degree of Doctor in Philosophy by:

Laura Jayne Bonnett

June 2012

Abstract

Introduction and Aims

Epilepsy is a neurological disorder and is a heterogeneous condition both in terms of cause and prognosis. Prognostic factors identify patients at varying degrees of risk for specific outcomes which facilitates treatment choice and aids patient counselling.

Few prognostic models based on prospective cohorts or randomised controlled trial data have been published in epilepsy. Patients with epilepsy can be loosely categorised as having had a first seizure, being newly diagnosed with epilepsy, having established epilepsy or frequent unremitting seizures despite optimum treatment. This thesis concerns modelling prognostic factors for these patient groups, for outcomes including seizure recurrence, seizure remission and treatment failure.

Methods

Methods for modelling prognostic factors are discussed and applied to several examples including eligibility to drive following a first seizure and following withdrawal of treatment after a period of remission from seizures. Internal and external model validation techniques are reviewed. The latter is investigated further in a simulation study, the results of which are demonstrated in a motivating example. Mixture modelling is introduced and assessed to better predict whether a patient would achieve remission from seizures immediately, at a later time point, or whether they may never achieve remission.

Results

Multivariable models identified a number of significant factors. Future risk of a seizure was therefore obtained for various patient subgroups. The models identified that the chance of a second seizure was below the risk threshold for driving, set by the DVLA, after six months, and the risk of a seizure following treatment withdrawal after a period of remission from seizures was below the risk threshold after three months.

Selected models were found to be internally valid and the simulation study indicated that concordance and a variety of imputation methods for handling covariates missing from the validation dataset were useful approaches for external validation of prognostic models. Assessing these methods for a selected model indicated that the model was valid in independent datasets. Mixture modelling techniques begin to show an improved prognostic model for the frequently reported outcome time to 12-month remission.

Conclusions

The models described within this thesis can be used to predict outcome for patients with first seizures or epilepsy aiding individual patient risk stratification and the design and analysis of future epilepsy trials. Prognostic models are not commonly externally validated. A method of external validation in the presence of a missing covariate has been proposed and may facilitate validation of prognostic models making the evidence base more transparent and reliable and instil confidence in any significant findings.

Acknowledgements

I would like to take the opportunity to thank those who have made this thesis possible. First and foremost I wish to thank Dr Catrin Tudur Smith and Prof Tony Marson for their assistance in the preparation of this thesis, and all their advice throughout my PhD. I cannot thank my supervisors without thanking Prof Jane Hutton, University of Warwick, for all her help in the writing of Chapter 10. She kindly offered theoretical advice and practical solutions to the complexities of mixture modelling.

My appreciation also goes to other members of the Department of Statistics, especially Dr Andrea Jorgensen who has shared an office with me for the duration of my PhD and has therefore borne the brunt of all my questions, whinging and complaining! Mrs Susie Dodd, a fellow PhD student in the department, also deserves a mention for not only providing a listening ear, but also allowing me to play teacher and therefore occasionally feel that I know what I am talking about! Thanks must also go to Mrs Kath Forrest and Miss Karen Barnes for their administrative support and for providing sustenance in the form of biscuits and sweets!

I would also like to thank the National Institute for Health Research for funding this work and Dr Pete Dixon for answering endless data queries.

My personal thanks go to my parents for their constant support and encouragement, and for inspiring my passion for education and hard-work. Special thanks go to my Mum for reading through this thesis in its entirety and providing feedback.

Finally, I want to give my wholehearted thanks to Rob, my partner, for his compassion, understanding and calming influence, and for helping me keep everything in perspective.

Table of Contents

Chapter 1: Introduction to Prognostic Factor Studies 1

1.1 Introduction to Prognostic Factors for Epilepsy 1

1.2 Introduction to Prognostic Factor Studies 3

1.3 Prognostic Factor Studies 3

1.3.1 Approaches to Prognostic Factor Studies 5

1.3.2 Improvements to Prognostic Factor Studies 10

1.4 Modelling Survival Data 11

1.4.1 Introduction 11

1.4.2 Parametric Methods 12

1.4.2.1 Parametric Models 13

1.4.2.2 Accelerated Failure Time 16

1.4.2.3 Cox Proportional Hazards 16

1.4.3 Nonparametric Methods 18

1.4.3.1 Neural Networks 19

1.4.3.2 Hierarchical Trees 20

1.4.4 Comparison of Methods 21

1.5 Thesis Outline 22

1.5.1 Introduction to Epilepsy 23

1.5.2 Methods for Identifying Prognostic Factors for Epilepsy 23

1.5.3 Identifying Prognostic Factors for Epilepsy 24

1.5.4 Implications for Drivers with a First Seizure 24

1.5.5 Implications for Drivers who withdraw AED Treatment 25

1.5.6 Internal Validation of Prognostic Models 25

1.5.7 Externally Validating Prognostic Models – Simulation Study 26

1.5.8 Externally Validating Prognostic Models – Case Study 26

1.5.9 Prognostic Models for Remission – Mixture Modelling 27

1.5.10 Conclusions and Further Work 27

Chapter 2: Introduction to Epilepsy 29

2.1 Introduction 29

2.2 Seizure Types 30

2.2.1 Focal Epilepsies 32
2.2.1.1 Simple or Complex Partial Seizures 33
2.2.1.2 Secondary Generalised Seizures 34
2.2.2 Generalised Epilepsies 34
2.2.2.1 Absence Seizures 35
2.2.2.2 Myoclonic Seizures 35
2.2.2.3 Tonic-Clonic Seizures 35
2.2.2.4 Tonic Seizures 36
2.2.2.5 Atonic Seizures 36
2.2.3 Status Epilepticus 37
2.3 Investigations and Diagnosis 37
2.3.1 Electroencephalogram 38
2.3.2 Magnetic Resonance Imaging 38
2.4 Treatment 39
2.4.1 Non-Drug Treatments 42
2.4.1.1 Resective Surgery 42
2.4.1.2 Vagus Nerve Stimulation 42
2.4.1.3 The Ketogenic Diet 43
2.5 Living with Epilepsy 43
2.5.1 Epilepsy and Driving 44
2.6 Prognostic Factor Studies in Epilepsy 44
2.7 Clinical Motivation 47

Chapter 3: Methods for Identifying Prognostic Factors for Epilepsy 49
3.1 Introduction 49
3.2 Handling Continuous Predictors 49
3.2.1 Cutpoint Model 50
3.2.2 Spline Functions 51
3.2.3 Fractional Polynomials 54
3.2.3.1 Choice of Origin 55
3.2.3.2 Function Selection Procedure 55
3.2.3.3 Multivariable Fractional Polynomial 56
3.2.3.4 Modelling Continuous Variables with a ‘Spike’ at Zero 57
3.2.3.5 Hazard Ratios and Confidence Intervals 59

3.2.4 Comparison of Splines and Fractional Polynomials 60
3.3 Variable Centring 61
3.3.1 Multicollinearity 61
3.3.2 Interpretation 63
3.3.3 Discussion 63
3.4 Model Building and Development 65
3.4.1 Treatment-Covariate Interactions Methods 67
3.4.1.1 Likelihood Ratio Test 68
3.4.1.2 Multivariable Fractional Polynomial Interactions 69
3.5 Testing the Proportional Hazards Assumption 71
3.5.1 Log Cumulative Hazard Plots 71
3.5.2 Time-Dependent Covariate Effects 74
3.5.3 Residuals for Cox regression model 74
3.5.3.1 Schoenfeld Residuals 75
3.5.3.2 Martingale Residuals 75
3.5.3.3 Deviance Residuals 76
3.5.4 Comparison of Deviances 77
3.6 Competing Risks Methods 77
3.6.1 Cumulative Incidence Analysis 78
3.6.2 Fine and Gray 79
3.7 Discussion 80

Chapter 4: Identifying Prognostic Factors for Epilepsy 83
4.1 Introduction 83
4.2 Methods 84
4.2.1 SANAD 84
4.2.2 Prognostic Modelling 86
4.3 Results 89
4.3.1 Time to overall Treatment Failure 91
4.3.2 Time to Treatment Failure due to Competing Risks 96
4.3.2.1 Treatment Failure due to Inadequate Seizure Control 97
4.3.2.2 Treatment Failure due to Unacceptable Adverse Events 98
4.3.3 Time to 12 Month Remission 101
4.4 Discussion 106

Chapter 5: Implications for Drivers with a First Seizure 111
5.1 Introduction 111
5.2 Methods 114
5.2.1 Statistical analysis 118
5.3 Results 121
5.4 Discussion 128

Chapter 6: Implications for Drivers who Withdraw AED Treatment 135
6.1 Introduction 135
6.2 Methods 136
6.2.1 Antiepileptic Drug Withdrawal Study 136
6.2.2 Systematic Review 139
6.2.2.1 Cochrane Collaboration Tool 140
6.2.2.2 Newcastle-Ottawa Scale 141
6.3 Results 141
6.3.1 Antiepileptic Drug Withdrawal Study 141
6.3.2 Recurrence Risk Following Antiepileptic Drug Withdrawal 143
6.3.3 Risk Following Recurrence and Recommencing Treatment 144
6.3.4 Systematic Review 145
6.4 Discussion 151

Chapter 7: Internal Validation of Prognostic Models 157
7.1 Introduction 157
7.2 Evaluating Model Performance 158
7.2.1 Measures of Explained Variation 158
7.2.1.1 Spearman, Kendall and Somer 159
7.2.1.2 Schemper (& Henderson) 161
7.2.1.3 Korn & Simon 162
7.2.1.4 Graf & Schumacher 163
7.2.1.5 Nagelkerke 163
7.2.1.6 Haegerty & Zheng 164
7.2.1.7 O'Quigley 164
7.2.1.8 Kent & O'Quigley 165
7.2.1.9 Maddala 165

7.2.1.10 Conclusion 165
7.2.2 Discrimination 166
7.2.3 Calibration 167
7.3 Internal Validation of Prediction Models 167
7.3.1 Apparent Validation 167
7.3.2 Split-sample validation 168
7.3.3 Cross-validation 168
7.3.4 Bootstrap Validation 169
7.4 Internal Validation of the SANAD Models 170
7.4.1 Time to Treatment Failure 171
7.4.2 Time to 12 Month Remission 173
7.5 Conclusion 175

Chapter 8: Externally Validating Prognostic Models – Simulation Study 179
8.1 Introduction 179
8.2 External Validation 181
8.2.1 Literature Review 183
8.2.2 Predictive Performance 188
8.2.3 Discrimination 189
8.2.4 Calibration 193
8.2.5 Prognostic Separation 194
8.2.6 Royston-Parmar's Alternative to the Cox Model 198
8.2.7 Conclusion 202
8.3 Missing Data 203
8.3.1 Mechanisms 204
8.3.1.1 Missing Completely at Random 204
8.3.1.2 Missing at Random 204
8.3.1.3 Ignorable 204
8.3.1.4 Non-Ignorable 205
8.3.2 Missing Data within Covariates 205
8.3.2.1 Complete Case Analysis 205
8.3.2.2 Available Case Analysis 207
8.3.2.3 Additional Category or Missing Indicator Method 208
8.3.2.4 Imputation 208

8.3.2.5 Maximum Likelihood Algorithms 213
8.3.2.6 The Expectation-Maximisation Algorithm 214
8.3.3 Methods for Handling Missing Covariates 214
8.3.3.1 Remodelling with Variable Matching 216
8.3.3.2 Random Selection with Replacement 216
8.3.3.3 Single Imputation via Estimated Proportions 216
8.3.3.4 Hot Deck Imputation 217
8.3.3.5 Multiple Imputation 217
8.3.3.6 Conclusion 218
8.4 Simulation Study 219
8.4.1 Methods Under Consideration 219
8.4.2 Simulation Procedures – Data Generation 220
8.4.2.1 Survival Data 221
8.4.2.2 Censoring Indicator 223
8.4.2.3 Covariate Data 223
8.4.3 Simulation Procedures - Scenarios to be Investigated 224
8.4.4 Simulation Procedures – Statistical Methods 225
8.4.5 Results 228
8.4.5.1 Concordance 229
8.4.5.2 Deviance 230
8.4.5.3 Prognostic Separation 232
8.4.5.4 Measures of Accuracy 233
8.5 Conclusion 236

Chapter 9: Externally Validating Prognostic Models – Case Study 241
9.1 Introduction 241
9.2 Validation using NGPSE 249
9.2.1 Results 249
9.2.1.1 Sensitivity Analysis 252
9.2.2 Conclusion 254
9.3 Validation using WA 254
9.3.1 Results 254
9.3.2 Conclusion 256
9.4 Validation using FIRST 257

9.4.1 Sensitivity Analysis 257
9.4.1.1 Conditional Seizure Risks 259
9.4.2 Methods for Handling a Missing Covariate 262
9.4.3 Conclusion 264
9.5 Missing Data – Comparing Simulated and Real Results 264
9.5.1 Results 265
9.5.1.1 Missing Binary Covariate 265
9.5.1.2 Missing Categorical Variable 265
9.5.2 Conclusion 269
9.6 Summary 269

Chapter 10: Prognostic Models for Remission – Mixture Modelling 275
10.1 Introduction 275
10.2 Mixture Modelling 276
10.2.1 General Finite Mixture Models 278
10.2.1.1 Estimating the Parameters in Finite Mixture Densities 278
10.2.1.2 Detecting Finite Mixture Densities 281
10.2.1.3 Determining the Number of Components in a Mixture 282
10.2.1.4 Including Covariates in a Finite Mixture Model 284
10.2.2 Finite Mixture Models for Survival 284
10.2.2.1 Long-Term Survivor Model (Proportion Cured Model) 288
10.2.2.2 Fitting by Maximum Likelihood 290
10.2.2.3 Application of the EM Algorithm 291
10.3 Modelling Time to 12 Month Remission 292
10.3.1 Non-Mixture Modelling 292
10.3.1.1 Immediate 12 Month Remission 292
10.3.1.2 Time to Delayed 12 Month Remission 295
10.3.1.3 Discussion 297
10.3.2 Mixture Modelling 298
10.3.2.1 Stage 1: Is there a cure fraction? 298
10.3.2.2 Stage 2: Distribution of Delayed Remission Group 299
10.3.2.3 Step 3: Covariates 303
10.3.2.4 Step 4: Setting Up the Likelihood 306
10.3.2.5 Step 5: Is the assumption valid? 309

10.3.2.6 Discussion 309
10.3.3 Future Work 311
10.4 Conclusion 312

Chapter 11: Conclusions and Further Work 315
11.1 Introduction 315
11.2 Recommendations for Practice 316
11.2.1 Model Development 317
11.2.2 Model Presentation 317
11.2.3 Model Validation 318
11.3 Recommendations for Research 319
11.3.1 Model Development 319
11.3.2 Model Validation 320
11.4 Future Work 321

Bibliography 325

Appendix A – Further Results from Chapter 5 345

Appendix B – Search Strategy used in Chapter 6 351

Appendix C – Further Results from Chapter 8 353

Appendix D – Further Results from Chapter 9 367

Appendix E – Published Papers 371

Abbreviations

AED	Antiepileptic Drug
AIC	Akaike's Information Criterion
CBZ	Carbamazepine
CI	Confidence Interval
CT/MRI	Computerised Tomography or Magnetic Resonance Imaging
EEG	Electroencephalogram
EM	Expectation-maximisation algorithm
FIRST	First tonic-clonic dataset from Italy
FP	Fractional Polynomial
GBP	Gabapentin
HR	Hazard Ratio
IQR	Interquartile Range
LTG	Lamotrigine
MAR	Missing at random
MCAR	Missing completely at random
MESS	Multicentre Study of Early Epilepsy and Single Seizures
MFP	Multivariable Fractional Polynomial
NGPSE	National General Practice Survey of Epilepsy
OXC	Oxcarbazepine
ROC	Receiver Operating Characteristic curve
SANAD	Standard Versus New Antiepileptic Drug trial
TPM	Topiramate
WA	Western Australian dataset of first tonic-clonic seizures

Chapter 1: Introduction to Prognostic Factor Studies

1.1 Introduction to Prognostic Factors for Epilepsy

Epilepsy is a common neurological condition, with an estimated incidence of 50 per 100,000 and prevalence of 5 to 10 per 1000 in the developed world. Between two and three percent of the population will be given a diagnosis of epilepsy at some time in their lives, the majority of whom will go into remission. However, up to 30% of those with epilepsy will fail to respond to monotherapy with standard antiepileptic drugs.

Prognostic factors are defined as demographic, disease-specific, or co-morbid characteristics associated strongly enough with a condition's outcome to predict accurately the eventual development of those outcomes. Prognostic factors can help to identify patients at different degrees of risk for specific outcomes, facilitate treatment choice and aid patient counselling. Few prognostic models based on randomised controlled trial data or prospective cohorts have been published in epilepsy.

Patients with epilepsy can be loosely categorised as having had a first seizure, being newly diagnosed with epilepsy – defined as at least two clinically unprovoked seizures, having established epilepsy, or frequent unremitting seizures despite optimum treatment. Prognostic models for time to treatment failure and time to 12 month remission for newly diagnosed patients are developed in Chapter 4. In Chapter 5 a prognostic model for risk of seizure recurrence following a first ever seizure is developed and in Chapter 6 prognostic models are developed for risk of recurrence for patients who withdraw treatment following a period of remission and for patients who reinstate treatment following seizure recurrence after a period of remission. The models in Chapter 5 and 6 are particularly

focussed on people with epilepsy who drive as the models influence the driving regulations regarding time off driving following a seizure or treatment withdrawal.

Before prognostic models can be accepted into general practice they need to be validated. Internal validation – assessment of model validity for the setting where the development data originated from – is examined in Chapter 7 for the models developed in Chapter 4. External validation – assessment of model validity in other samples which are fully independent from the development data – is examined via a literature review and simulation study in Chapter 8, the results of which are applied to external validation of the model from Chapter 5 in Chapter 9.

The internal validity of the model for remission in Chapter 4 is found to be poor at one year. This may relate to the heterogeneity of outcome for epilepsy – some patients achieve remission immediately at 12 months while others may never achieve remission. Therefore, it may be more appropriate to model the three distinct patient subgroups in one overall model for time to 12 month remission via a three-component mixture model. This is the content of Chapter 10 which precedes an overview of the entire thesis together with suggestions for future research and clinical practice (Chapter 11).

This chapter continues with a summary of prognostic factors studies including approaches to studying prognostic factors in section 1.3.1 and possible improvements that could be made to future prognostic factor studies (section 1.3.2). This is followed by an overview of methods for modelling survival data including parametric methods, such as the accelerated failure time model (section 1.4.2.2) and the Cox proportional hazards model (section 1.4.2.3), and nonparametric methods like neural networks (section 1.4.3.1) and hierarchical trees (section 1.4.3.2). The chapter concludes with an overview of the thesis including a detailed summary of each chapter (section 1.5).

1.2 Introduction to Prognostic Factor Studies

Prognostic factors are defined as demographic, disease-specific, or co-morbid characteristics associated strongly enough with a condition's outcome to predict accurately the eventual development of those outcomes [1]. Prognostic factors can help to identify patients at different degrees of risk for specific outcomes, facilitate treatment choice and aid patient counselling.

When individual prospects for survival, or indeed any outcome, are highly variable, it is natural to look for possible explanations. Knowledge of prognostic variables can aid in the understanding of a disease [2-4] and may lead to variation in treatment according to a patients' predicted probability of survival [5-7]. Prognostic factor studies are described in section 1.3.

In the medical literature, much emphasis has traditionally been given to the identification of predictors. As of January 2007, over 60,000 papers had been published with the term "predictor" or "prognostic factor" [8]. Statistical models for prediction can be discerned into main classes: regression, classification, and neural networks. These models are discussed in section 1.4.

1.3 Prognostic Factor Studies

The purpose of prognostic factor studies is to improve understanding of the disease process, improve the design and analysis of clinical trials and define risk groups based on prognosis. In addition, prognostic factor studies can assist in comparing outcomes between treatment groups in non-randomised studies by allowing adjustment for case mix, predicting disease outcomes more accurately or parsimoniously and guiding clinical decision making, including treatment selection and patient counselling [9, 10].

When conducting and reporting prognostic factor studies, researchers should consider the approach to prognosis, explanatory or outcome prediction, and phase of investigation [11]. Readers of health care research will then be better able to evaluate the goals and interpret, and appropriately use, the results of prognostic factor studies.

Windeler observed that summaries of prognosis are not meaningful unless associated with a particular strategy for treatment and suggested that the greatest importance of prognostic factor studies is to aid decisions about treatment [12]. If the treatment received varies in relation to prognostic variables, then the study cannot deliver an unbiased and meaningful assessment of prognostic ability, unless the different treatments are equally effective - in which case, why vary the treatments? Such variation in treatment may be quite common once there is evidence, usually non-systematic, that a variable is prognostic. Ideally therefore, prognostic variables should be evaluated either in a cohort of patients treated the same way or in a randomised trial [13, 14].

Case-control and cross sectional studies may both be used to examine risk factors but these designs are much weaker [10]. This is because cross sectional studies are primarily used to determine prevalence and are also used to infer causation. They do not provide an explanation for their findings. Case-control studies determine the relative importance of a predictor variable in relation to the presence or absence of the disease. They are retrospective and cannot therefore be used to calculate the relative risk. The major problem with case-control studies are the familiar ones of confounding variables and bias. Case-control designs have, however, been shown to yield optimistic results for evaluations of diagnostic tests, a result that is likely to be relevant to prognostic studies [15]. This is because often, mild cases that are difficult to diagnose are omitted from case-control studies, causing an overestimation of sensitivity as well as specificity. Also, diagnostic accuracy can be overestimated if the test is evaluated in a group of patients already known

to have the disease and a separate group of normal patients, rather than in a relevant clinical population [16]. This is obviously the scenario for a case-control study.

1.3.1 Approaches to Prognostic Factor Studies

Two approaches to prognostic factor studies have been identified by Hayden - explanatory analysis and outcome prediction [11]. Explanatory studies focus on the causal association between prognostic factors and an outcome, whereas outcome prediction studies focus on variables taken together to identify the combination of factors that is most strongly associated with outcome and can be used to stratify patients on an outcome, often to triage them into treatment programs.

Altman and Lyman have also developed criteria for the design of a prognostic factor study [9]. According to this, primary and secondary hypotheses should be clearly stated, including any subgroup analyses planned in advance of the study. Prognostic factors for which there is sufficient evidence to warrant further investigation based on previous studies, biological and clinical plausibility and relevance, and importance to the understanding or treatment of the disease should also be considered.

Additionally, Altman and Lyman suggest that the study population should be defined with specific inclusion and exclusion criteria and methods to judge evaluability [9]. Patient treatment should be either standardised or assigned by randomisation and assays should be reproducible and be performed with knowledge of the clinical data and patient outcome. The planned analysis, including proposed hypothesis testing on subgroups and anticipated interaction in advance of the study should also be specified as well as the key study features including the above information which should be fully detailed in a formal written protocol.

To develop an alternative framework for understanding prognosis, Hayden [11] adapted the work of Altman and Lyman [9] and integrated the discussion of causal understanding published by Goldthorpe [17]. She identified three phases of explanatory prognosis investigation. The purpose of each phase can be described as exploration, confirmation [18], and development of understanding respectively. To support all three phases, additional information can be obtained from clinical observations, basic science, descriptive studies, and other sources that may support the evidence on prognosis.

Phase 1 studies identify associations between a number of potential prognostic factors and a health outcome and are the most common phase of prognostic investigation [11]. Although this type of investigation is necessary to identify new factors, it provides the least conclusive information regarding the independence of a variable as a valid prognostic factor. Phase 1 studies provide hypothesis-generating evidence indicative of a potential association between a prognostic factor and an outcome. They are often recognisable by their objective statements that present broad exploratory aims. Phase 1 studies do not focus on one specific prognostic factor rather the prognostic factor of interest is investigated as one of many factors assessed for their association with the outcome. For this reason, most Phase 1 studies only discuss the findings for factors that are found to be statistically significant, regardless of clinical significance.

Phase 1 investigations are appropriate when it is unclear which variables are important in predicting an outcome for a population [11]. However, it should be recognised that results from multiple studies in this exploratory phase of investigation often have widely varying results as spurious associations are common, and real effects are sometimes missed [9]. Also, some associations are present in one population but not in another. Further testing in a Phase 2 confirmatory investigation is required to confirm the results of a Phase 1 study.

Phase 2 studies test the independence of the association between a prognostic factor and the outcome of interest. These studies aim to measure the independent effect of a prognostic factor while controlling for confounders [11]. Typical Phase 2 studies are recognisable by their objective statements that outline a specific prognostic factor of interest and aim to investigate the relationship between this variable and the outcome. In Phase 2 studies, the importance shifts to assessing whether the factor independently adds to currently known prognostic factors.

For a specific prognostic factor, the finding of a significant association in a study testing an independent association is a stronger finding than an association identified in the previous phase of investigation [11]. However, both Phase 1 and 2 studies are based on the often unrealistic assumption that the effect of the prognostic factor on the outcome is direct and isolated. This may be inadequate to describe complex relationships for most conditions, and where the impact of factors may change over time.

Phase 3 studies of prognosis attempt to describe the complexity of the prognostic pathways or processes. These studies apply knowledge from the previous phases of study on independent associations and incorporate other knowledge from the field of study [11]. Phase 3 explanatory studies start with an explicit theoretical framework that includes the prognostic construct of interest, variables that are thought to influence or modify the effect of that factor, variables that are thought to be intermediate or a mediator in the pathway toward the outcome, potential confounding variables, and the outcome of interest. Phase 3 explanatory studies provide evidence supportive of the mechanism or mechanisms of action of a prognostic factor on the outcome. However, the strength of interpretation of Phase 3 explanatory studies depends on the strength of the theoretical framework and the extent to which existing knowledge about the disease or illness supports the study [11].

Table 1 provides a summary of the phases of an exploratory prognostic factor investigation [11].

Table 1: Phases of an Exploratory Prognostic Factor Investigation

Phase of Investigation	Issue addressed for prognostic factor of interest	Interpretation of results for prognostic factor of interest
Phase 1: Exploration of associations	Measures the presence of a prognostic relationship between factor(s) and the outcome	Provides hypothesis-generating evidence indicative of a potential association between a prognostic factor and outcome
Phase 2: Testing independent associations	Measures the strength of the prognostic relationship between a factor and the outcome while controlling for alternative explanations	Provides evidence supportive of the independent effect of a prognostic factor
Phase 3: Understanding and testing prognostic pathways	Examines the role of the prognostic factor and the process by which it impacts the outcome	Provides evidence supportive of the prognostic factor's mechanism(s) of action on the outcome

This framework differs from that of Altman and Lyman [9] in two ways. Firstly, the two approaches of prognosis studies were distinguished between. Secondly, cross-sectional studies were not included under the umbrella of prognosis which Altman and Lyman [9] described as "hypothesis generating" studies. These studies provide information that may lead to testing in longitudinal studies. However, without a longitudinal component they do not provide information on predictors of outcome over time [11].

Hayden has identified three specific areas where consideration of a framework of explanatory prognosis studies is important: defining the study objectives, presenting the study methods and data, and interpreting and applying the results of the study [11]. In terms of defining study objectives, investigators need to clearly identify the purpose of their study, to specify how their study will be used to improve health and health care, and to outline how their study will add to the existing knowledge in the area. This means they must decide what approach, explanatory or outcome prediction, and what phase of investigation, is needed to advance the knowledge on a specific topic. The extent of the

previous evidence, from prognosis research and other sources of information, will inform the study objectives about the phase of investigation [11]. A systematic and thoughtful review of literature is commonly required before funding for new randomised controlled trials is received, and is equally appropriate before planning and conducting new prognosis studies.

The study design should follow directly from the prognosis study objectives. For Phase 2 and 3 explanatory studies, a clear presentation and rationale should be given to explain which confounders will be adjusted for in the analysis and why [19]. Study reports should give a clear and thoughtful presentation of what the current study adds to the available literature, the limitations of the research, and recommendations for future research. Phase 1 studies provide exploratory evidence regarding potential prognostic factors. Further study, in the form of a Phase 2 study, would be needed to confirm the independent predictive effect of that factor while thoughtful planning and investigation of the prognostic pathway in a Phase 3 study is required to further understand how the prognostic factor affects the outcome.

Interpretation of evidence from Phase 1 or 2 studies, without understanding a potential prognostic factor's causal pathway may lead to inappropriate or harmful treatment recommendations or policy decisions. A review, Phase 2, by Goldman [20] stated that studies had consistently shown that older patients had a higher risk of cardiac complications and in most studies age had been an independent predictive factor even after controlling for the severity of cardiac disease and combined conditions. Two recent studies, Phase 3, of elderly patients suggested that such prevailing attitudes may lead to under-treatment or even inappropriate treatment of the elderly. Maly [21] found a negative association between age and discussion of treatment options including surgery for

breast cancer by physicians while Ellis [22] advocated an aggressive surgical approach in elderly patients for cancer of the oesophagus and cardia.

Interventions that are planned based on an understanding of how factors influence outcome, differences over time, dosage differences, or different contextual settings will allow optimisation of patient management. For example, in patients with laryngeal carcinoma, breast carcinoma and carcinoma of the bladder, prognostic factors are used to select patients who are good candidates for organ preservation [23].

The explanatory framework presented can assist clinicians in their interpretation of the prognosis literature. Researchers can also use the framework to plan innovative research programs based on theoretical models of prognostic pathways.

1.3.2 Improvements to Prognostic Factor Studies

The recently proposed REMARK consensus [24] is a step forward for the standardisation of the reporting of cancer prognostic marker studies. No such guidelines currently exist for prognostic factor studies. In general, authors should avoid highlighting only the significant associations and they should try to present, or at least alert readers to, all covariates that they considered *a priori* versus post hoc. Additionally, journals should encourage the publication of well designed, executed and reported prognostic factor studies, regardless of the significant or non-significant findings.

The development of collaborations and networks between investigators may be beneficial, if such networks focus also on minimising selective reporting and publication bias [25, 26]. Transparency and public availability of protocols, data, analyses and results would also be helpful [25-27]. Funding groups should also realise that investigators should not be supported primarily for their ability to produce statistically significant results, but they

should reward novel ideas, rigorous design, implementation and transparent reporting of results, regardless of their statistical significance.

1.4 Modelling Survival Data

1.4.1 Introduction

When the length of time taken for an event to occur is measured, the resulting data are known as time-to-event, or survival, data. Often the event is death, hence the name of the method - survival analysis where the term survival usually means remaining event free. In the medical setting the event might be discharge from hospital, weaning of a breast-fed infant, recurrence of tumour or remission from a disease and the time starting point might be time of diagnosis, time of surgery or time of entrance into the study.

One problem with time-to-event data is that the event is often not observed on all subjects. This might be because a person dropped out due to death or some other reason, or maybe the study ended before the event had occurred for an individual. Although it is not known whether the event will happen, duration of follow-up for a person without the event being observed is known. Individuals for whom the event is not observed are called censored. Methods of analysis must be able to cope with censored values.

Often a number of variables are observed at the commencement of a trial, and survival is related to the values of these variables; that is, the variables are prognostic. Methods of analysis must be able to take account of the distribution of prognostic variables in the groups under study. There are many ways to analyse time-to-event data, the most popular of which is the Cox proportional hazards model. Alternatives include parametric methods, such as the accelerated failure time model, hierarchical trees and neural networks.

1.4.2 Parametric Methods

A possible approach to the analysis of studies with survival time as the variable of interest is to postulate a distribution for survival time and to estimate the parameters of this distribution from the data. This approach is usually applied by starting with a model for the death rate and determining the form of the resulting survival time distribution.

The death rate will usually vary with time since entry to the study, t, and will be denoted $\lambda(t)$, referred to as the hazard function. Suppose the probability density of survival time is $f(t)$ and the corresponding distribution function is $F(t)$. Then, since death rate is the rate at which deaths occur divided by the proportion of the population surviving,

Equation 1

$$\lambda(t) = \frac{f(t)}{1 - F(t)} = \frac{f(t)}{S(t)}$$

where $S(t) = 1 - F(t)$ is the proportion surviving and is referred to as the survivor function. Therefore $f(t)$ and $S(t)$ can be specified in terms of $\lambda(t)$ and the general solution is obtained by integrating Equation 1 with respect to t and noting that $f(t)$ is the derivative of $F(t)$.

In general, the hazard will depend on time and a family of models and may be written as

Equation 2

$$\lambda(t, \boldsymbol{x}) = \lambda_0(t) \exp(\boldsymbol{\beta^T x})$$

where $\boldsymbol{\beta^T x}$ is the matrix representation of the regression function $\beta_1 x_1 + \cdots + \beta_p x_p$ and $\lambda_0(t)$ is the time-dependent part of the model [28]. The term $\lambda_0(t)$ could represent any parametric function of t such as the exponential, Weibull or Gompertz.

Equation 2 is a regression model in terms of the covariates and is also referred to as a proportional hazards model since the hazards for different sets of covariates remain in the same proportion for all t. Data can be analysed parametrically using Equation 2 provided that some particular form of $\lambda_0(t)$ is assumed. The parameters of $\lambda_0(t)$ and also the regression coefficients, $\boldsymbol{\beta}$, would be estimated.

1.4.2.1 Parametric Models

1.4.2.1.1 Exponential

The simplest model for the hazard function is to assume that it is constant over time. The hazard of death at any time after the time origin of the study is then the same, irrespective of the time elapsed. Under this model the hazard function may be written as $\lambda(t) = \lambda$ for $0 \leq t < \infty$. Thus, the corresponding survivor function is $S(t) = \exp\{-\lambda t\}$ and so the implied probability density function of the survival times is $f(t) = \lambda \exp\{-\lambda t\}$ for $0 \leq t < \infty$.

The exponential model is investigated in Chapter 10. It is examined for its suitability to model time to 12 month remission for patients in The SANAD Trial (Chapter 4) who achieve remission but not immediately at 365 days.

1.4.2.1.2 Weibull

In practice, the assumption of a constant hazard function, or equivalently of exponentially distributed survival times, is rarely tenable. A more general form of hazard function is such that $\lambda(t) = \lambda p(t)^{p-1}$ for $0 \leq t < \infty$, a function that depends on two parameters λ and p, which are both greater than zero. In the particular case that $p = 1$, the hazard function takes a constant value λ, and the survival times have an exponential distribution. For other values of p, the hazard function increases, for $\lambda > 1$, or decreases, for $\lambda < 1$, monotonically. The shape of the hazard function depends on the value of p so p is known as the shape parameter while the parameter λ is a scale parameter.

For this choice of hazard function, the survival function is given by $S(t) = \exp\{-\lambda t^p\}$. The corresponding probability density function is then $f(t) = \lambda p t^{p-1} \exp(-\lambda t^p)$ for $0 \leq t < \infty$.

The Weibull model is investigated in Chapter 10. It is examined for its suitability to model time to 12 month remission for patients in The SANAD Trial (Chapter 4) who achieve remission but not immediately at 365 days.

1.4.2.1.3 Log-logistic

One limitation of the Weibull hazard function is that it is a monotonic function of time. However, situations in which the hazard function changes direction can arise. For example, following a heart transplantation, a patient faces an increasing hazard of death over the first ten days after transplant while the body adapts to the new organ. The hazard then decreases with time as the patient recovers. In situations such as this, a unimodal hazard function may be appropriate.

A particular form of unimodal hazard is the function

$$h(t) = \frac{e^{\theta} \kappa t^{\kappa-1}}{1+e^{\theta} t^{\kappa}},$$

for $0 \leq t < \infty, \kappa > 0$. This hazard function decreases monotonically if $\kappa \leq 1$, but if $\kappa > 1$, the hazard has a single mode. The survivor function corresponding to the hazard function is given by

$$S(t) = \left\{1 + e^{\theta} t^{\kappa}\right\}^{-1},$$

and the probability density function is

$$f(t) = \frac{e^{\theta} \kappa t^{\kappa-1}}{\left(1+e^{\theta} t^{\kappa}\right)^2}.$$

This is the density of a random variable T that has a log-logistic distribution with parameters θ and κ. The distribution is so called because the variable $\log T$ has a logistic distribution, a symmetric distribution whose probability density function is very similar to that of the normal distribution.

The log-logistic model is investigated in Chapter 10. It is examined for its suitability to model time to 12 month remission for patients in The SANAD Trial (Chapter 4) who achieve remission but not immediately at 365 days.

1.4.2.1.4 Gompertz

Although the Weibull model is the most widely used parametric model, the Gompertz has found application in demography and biological sciences. Indeed the distribution was introduced by Gompertz in 1825 as a model for human mortality [29].

The hazard function of the Gompertz distribution is given by $\lambda(t) = \lambda \exp(\theta t)$ for $0 \leq t < \infty$ and $\lambda > 0$. In the particular case where $\theta = 0$, the hazard function has a constant value λ and the survival times then have an exponential distribution. The parameter θ determines the shape of the hazard function, positive values leading to a hazard function that increases with time. Like the Weibull hazard function, the Gompertz hazard increases or decreases monotonically.

The survivor function of the Gompertz distribution is given by $S(t) = \exp\left\{\frac{\lambda}{\theta}(1 - \exp(\theta t)\right\}$ and the corresponding density function is $f(t) = \lambda \exp(\theta t) \exp\left\{\frac{\lambda}{\theta}(1 - \exp(\theta t)\right\}$.

The Gompertz distribution is included for completeness only. It is not used within this thesis.

1.4.2.2 Accelerated Failure Time

Another way of representing the effect of the covariates is to suppose that the distribution of survival time is changed by multiplying the time-scale by $\exp(\boldsymbol{\beta}_a^T\boldsymbol{x})$, that is, that the logarithm of survival time is increased by $\boldsymbol{\beta}_a^T\boldsymbol{x}$. The effect of a predictor is to alter the rate at which a subject proceeds along the time axis (i.e. to accelerate the time to failure [30]).

The hazard can then be written

Equation 3

$$\lambda(t,\boldsymbol{x}) = \lambda_0[t\exp(-\boldsymbol{\beta}_a^T\boldsymbol{x})]\exp(-\boldsymbol{\beta}_a^T\boldsymbol{x})$$

This is referred to as an accelerated failure time model. For the exponential distribution, $\lambda_0(t) = \lambda$, Equation 2 and Equation 3 are equivalent, with $\boldsymbol{\beta}_a = -\boldsymbol{\beta}$, so the accelerated failure time model is also a proportional hazards model. The same is true for the Weibull and Gompertz but in general the accelerated failure time model would not be a proportional hazards model [28].

The accelerated failure time model would have been used in Chapters 4, 5 and 6 if the proportional hazards assumption of the Cox model was found not to be valid for any of the developed prognostic models. In this case the results would be compared to those from the accelerated failure time model to determine if the violation of the assumption affected the conclusion. The developed models all satisfied the proportional hazards assumption so the accelerated failure time model was not required.

1.4.2.3 Cox Proportional Hazards

Since often an appropriate parametric form of $\lambda_0(t)$ is unknown and, in any case, not of primary interest, it would be more convenient if it were unnecessary to substitute any particular form for $\lambda_0(t)$ in Equation 2. This was the approach introduced by Cox [31]. The model is then non-parametric with respect to time but parametric in terms of the

covariates. Estimation of $\boldsymbol{\beta}$ and inferences are developed by considering the information supplied at each time that a death occurred.

Consider a death occurring at time t_j and suppose that there were n_j' subjects alive just before t_j, that the values of $\boldsymbol{x}$ for these subjects are $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n_j'}$ and that the subject that died is denoted, with no loss of generality, by the subscript 1. The set of n_j' subjects at risk is referred to as the risk set. The risk of death at time t_j for each subject in the risk set is given by Equation 2.

This does not supply absolute measures of risk but does supply the relative risks for each subject since, although $\lambda_0(t)$ is unknown, it is the same for each subject [28]. Thus, the probability that the death observed at t_j was of the subject who did die at that time is

$$p_j = \frac{\exp(\boldsymbol{\beta}^T \boldsymbol{x}_1)}{\sum \exp(\boldsymbol{\beta}^T \boldsymbol{x}_i)}$$

where summation is over all members of the risk set. Similar terms are derived for each time that a death occurred and are combined to form a likelihood.

The Cox regression or proportional hazards model [31, 32] is the standard statistical tool used to analyse multiple prognostic factors. The risk function for a patient with values $(Z_1, Z_2, \ldots, Z_k)$ of the prognostic factors can be written as $\lambda(t|Z_1, Z_2, \ldots, Z_k) = \lambda_0(t)\exp(\beta_1 Z_1 + \beta_2 Z_2 + \cdots + \beta_k Z_k)$ where $\lambda(t|\cdot)$ denotes the hazard function of the event-free or overall survival time random variable T and $\lambda_0(t)$ is the baseline hazard function representing the instantaneous risk of dying at time t given survival up to time t for a patient with $Z_1 = Z_2 = \cdots = Z_p = 0$. Since this model can be re-expressed in the form $\log\left\{\frac{\lambda_i(t)}{\lambda_0(t)}\right\} = \beta_1 Z_{1i} + \cdots + \beta_k Z_{ki}$, the proportional hazards model may also be regarded as a linear model for the logarithm of the hazard ratio. The estimated log-hazard ratios $\hat{\beta}_j$ can then be interpreted as estimated effects of the factors Z_j $(j = 1, \ldots, k)$.

If Z_j is measured on a quantitative scale then $\exp(\hat{\beta}_j)$ represents the increase or decrease in risk if Z_j is increased by one unit. If Z_j is a binary covariate then $\exp(\hat{\beta}_j)$ is the hazard ratio of the alternative category to the baseline category, which is assumed to be constant over the time range considered.

Although the Cox model is non-parametric to the extent that no assumptions are made about the form of the baseline hazard, there are still a number of important assumptions that need to be assessed after a model has been fitted to an observed set of survival data. The first such assumption is non-informative censoring. To satisfy this assumption the design of the underlying study must ensure that the mechanisms giving rise to censoring of individual subjects are not related to the probability of an event occurring. For example, in clinical studies, care must be taken that continuation of follow-up does not depend on a participants' medical condition.

The second assumption of the Cox model is that of proportional hazards. In a regression setting this means that the survival curves for two strata, determined by the particular choices of values for the variables, must have hazard functions that are proportional over time, i.e. constant relative hazard. There are tests that can be applied to test proportionality.

The Cox proportional hazard model is used in Chapters 4, 5, 6, 8, 9 and 10 to model time to event for various outcomes and various types of patients.

1.4.3 Nonparametric Methods

The nonparametric methods described in sections 1.4.3.1 and 1.4.3.2 are included for completeness only and are not used within this thesis.

1.4.3.1 Neural Networks

Faraggi and Simon [33], and others [34-37], have proposed a neural network generalisation of the Cox regression model defined by $\lambda(t|Z_1, Z_2, \dots, Z_k) = \lambda_0(t)\exp(f_{FS}(Z, w))$ where $f_{FS}(Z, w) = \sum_{j=1}^{r} W_j \Lambda\left(w_{0j} + \sum_{i=1}^{k} w_{ij} Z_i\right)$ and $Z_1, Z_2, \dots, Z_k$ are the values of the k prognostic factors. The weights $w = (W_0, \dots, W_r, w_{01}, \dots, w_{kr})$ can be estimated from the data via maximisation of the partial likelihood although other optimisation procedures are often used. Although the problem of censoring is satisfactorily solved in this approach, there remain problems with potentially serious over-fitting of the data, especially if the number, r, of hidden units is large [38].

In general, feed-forward neural networks with one hidden layer are universal approximators [39] and can consequently approximate any function defined by the conditional probability that Y is equal to one given Z with arbitrary precision by increasing the number of hidden units. This flexibility can however lead to serious over-fitting. This can be compensated for by introducing some weight decay [40, 41], for example by adding a penalty term $-\lambda\left(\sum_{j=1}^{r} W_j^2 + \sum_{j=1}^{r}\sum_{i=1}^{k} w_{ij}^2\right)$ to the log-likelihood [38]. The smoothness of the resulting function is then controlled by the decay parameter λ.

Another form of neural networks that have been applied to survival data are the so-called single time point models [42]. As they are identical to a logistic perceptron, otherwise known as a common logistic regression model, or a feed-forward neural network with a hidden layer, they correspond to fits of logistic regression models or their generalisations to survival data. In practice, a single time point t^* is fixed and the network is trained to predict the survival probability. The corresponding model is given by $P(T < t^*|Z) = \Lambda\left(w_0 + \sum_{i=1}^{k} w_i Z_i\right)$, or its generalisation when introducing a hidden layer, where $\Lambda(\cdot)$

denotes the logistic function, $\Lambda(u) = (1 + \exp(-u))^{-1}$, and is called the activation function [43].

A common drawback of these naïve approaches is that they do not enable incorporation of censored observations in a straightforward manner, which is closely related to the fact that they are based on unconditional survival probabilities instead of conditional survival probabilities. Neither omission of the censored observations, as suggested by Burke [44], nor treating censored observations as uncensored [43] are valid approaches but both instead are a serious source of bias. De Laurentiis and Ravdin [42] and Ripley [41] propose to impute estimated conditional survival probabilities for the censored cases from a Cox regression model. Further work is needed in this area.

1.4.3.2 Hierarchical Trees

Hierarchical trees are an approach for nonparametric modelling of the relationship between a response variable and several potential prognostic factors [45-49]. The idea of Classification And Regression Trees (CART), a synonym for different types of tree based analyses, is to construct subgroups that are internally as homogeneous as possible with regard to the outcome and externally as separated as possible. Hence the method leads directly to prognostic subgroups defined by the potential prognostic factors and is achieved by a recursive tree building algorithm.

The tree building algorithm produces a binary tree with a set of patients, a splitting rule, and the minimal p value, at each interior node. For patients in the resulting final nodes, various statistics can be computed such as Kaplan-Meier estimates of event-free survival or hazard ratios with respect to specific references or combined groups.

Unfortunately prognostic factors are usually measured on different scales meaning the number of possible partitions will be different [38]. Correction of p values and/or

restriction to a set of pre-specified cutpoints may be useful to overcome the problem that factors allowing more splits have a higher chance of being selected by the tree building algorithm. Due to multiple testing, the algorithm may also be biased in favour of these factors over binary factors with prognostic relevance [38].

To improve the predictive ability of trees, stabilising methods based on resampling have been proposed [50-54]. However, the results are difficult to interpret which reduces their value for practical applications.

1.4.4 Comparison of Methods

Although traditional statistical methods such as Cox proportional hazards or logistic regression are easy to perform and routinely available in standard software packages, machine learning methods such as hierarchical trees and neural networks are thought to predict more accurately because of greater model-fitting flexibility [55].

Artificial neural networks are popularly used as universal non-linear inference models. However, they suffer from two major drawbacks. The way they work is hidden because of the distributed nature of the representations they form [56], and this makes it difficult to interpret what they do. Worse still, there are no clearly accepted models of generality which makes it difficult to demonstrate reliability when applied to future data.

Cox proportional hazards models are well suited for regression modelling of survival data. They are simple to fit, can deal with time-varying regression coefficients as well as time-dependent covariates and no assumption is made on the distribution of the lifetimes of the baseline population. However, they are not flexible enough to deal with time-varying dynamics of covariate effects [57]. Additionally, the Cox model has the advantage over neural networks of providing some insight into which variables are most influential for

prognosis. Nevertheless, it is likely that the assumptions required by the Cox model may not be satisfied in all datasets, justifying the use of neural networks in certain cases.

Parametric regression models, such as the exponential, Weibull and Gompertz, may involve stronger distributional assumptions than it is suitable to make and inference procedures may not be sufficiently robust to departures from these assumptions [30]. This seems particularly to be the case in medical applications in which only limited experimentation in similar situations may have preceded the study in question or in which data are recorded by a number of individuals. Parametric models are also less flexible than proportional hazard models [38].

Although rank tests, encompassing accelerated failure time models, are derived with certain alternatives in mind for which optimum parametric procedures exist, they generally possess greater robustness than the corresponding parametric tests and are generally less sensitive to outliers [30]. In addition, for testing the null hypothesis, these tests generally involve only a small loss in efficiency compared to the parametric procedure when such a procedure is appropriate. Unfortunately though, accelerated failure time models are difficult to extend to handle time-varying effects.

1.5 Thesis Outline

This chapter has considered prognostic factors studies and statistical models for prediction. The thesis continues with chapters on epilepsy, methods for identifying prognostic factors for epilepsy, prognostic modelling of time to treatment failure and time to 12 month remission for newly diagnosed patients, prognostic modelling of risk of recurrence for patients with a first seizure only, prognostic modelling of risk of recurrence for patient who withdraw their medication, validation methods for prognostic models and more

sophisticated mixture modelling methods. Further descriptions of each chapter can be found in sections 1.5.1 to 1.5.10.

1.5.1 Introduction to Epilepsy

In Chapter 2, the condition of epilepsy will be summarised including descriptions of seizure types, such as simple partial seizures and absence seizures, and classifications, such as focal epilepsy or generalised epilepsy. Methods of identification and diagnosis such as electroencephalogram (EEG) and magnetic resonance imaging (MRI) will also be outlined together with potential treatments including resective surgery, antiepileptic drug treatment and the ketogenic diet.

The chapter will conclude with a literature review of prognostic factors studies in epilepsy which will highlight the dearth of such studies and provide justification for further research in this area. The clinical background of epilepsy, described in this chapter, will inform the terminology and medical concepts used throughout my thesis (Chapters 4 to 11).

1.5.2 Methods for Identifying Prognostic Factors for Epilepsy

Chapter 3 will focus on generic methods of modelling prognostic factors. Concepts such as handling continuous predictors by splines or fractional polynomial transformations will be considered as well as whether variable centring is appropriate, and if so, how it should be done. Following on from this, model building and development will be considered including variable selection via sequential or all-subset strategies. Methods for testing the proportional hazards assumption, required for the Cox model to be valid, will be outlined with particular interest in residuals. Finally, a description of the use of competing risks analysis will be provided as such an analysis is required if patients can fail treatment for multiple reasons. The chapter concludes with a section describing the chosen methods of analysis used in Chapters 4, 5 and 6.

1.5.3 Identifying Prognostic Factors for Epilepsy

The work in Chapter 4 has been published in Lancet Neurology and I am first author.

Using data from arm A of The SANAD Trial, multivariable models are produced for time to treatment failure, both overall and in the competing risks setting, and for time to 12 months of remission. In addition to presenting the models, risk estimates are calculated for various combinations of risk factors. To aid the interpretation of these risk estimates by non-statisticians, forest-style plots are used to illustrate the heterogeneity of outcome for differing patient groups.

Validation of these models is required and is the subject of subsequent chapters. Internal validation is undertaken in Chapter 7 while the concept of external validation is examined in Chapter 8.

1.5.4 Implications for Drivers with a First Seizure

The work in Chapter 5 has been published in the British Medical Journal, together with an editorial, and I am again first author.

Using data from the Multicentre study of Early epilepsy and Single Seizures (MESS), a prognostic model for the risk of a second seizure following a first ever seizure is developed. In an extension to the model, the time until the recurrence risk is below 20% is calculated for various combinations of risk factors to determine how long people with a first seizure should refrain from driving before their risk of a second seizure falls before the recommended risk level set by the U.K. based Driving and Vehicle Licensing Agency (DVLA). As a result of this work, the DVLA changed their guidelines concerning time off driving for patients with a first seizure.

This model again needs to be validated. External validation of this model is performed in Chapter 9.

1.5.5 Implications for Drivers who withdraw AED Treatment

The work in Chapter 6 has been published in the Journal of Neurology, Neurosurgery and Psychiatry, and I am again first author.

Using data from the Medical Research Council Antiepileptic Drug Withdrawal Study (MRC AEDWS) two prognostic models are developed. The first considers time to seizure recurrence following treatment withdrawal while the second considers time to seizure recurrence following treatment reinstatement after seizure recurrence for patients withdrawing their antiepileptic drug medication. In addition to the modelling presented, a literature review is described which identified three studies that provide additional information on risk of seizure recurrence following treatment withdrawal for patients in remission from seizures.

1.5.6 Internal Validation of Prognostic Models

Internal validation is necessary to assess the ability of models to predict outcome in settings where the development data originated from. Methods of assessing model performance are outlined in Chapter 7, such as the R^2 measure, together with methods of discrimination, via use of the c-statistic, and calibration, via plots. Methods of internal validation are also considered including bootstrap resampling.

The models derived in Chapter 4 using The SANAD Trial for time to treatment failure and time to 12 months of remission are validated internally in Chapter 7. Both models show fairly good internal validity although there are signs of poor validity at one year for the remission model. Reasons for this poor validity, together with an alternative way of modelling time to 12 month remission for patients with epilepsy, are described in Chapter 10.

1.5.7 Externally Validating Prognostic Models – Simulation Study

Chapter 8 begins with a literature review considering methods of external validation of prognostic models and methods used to handle covariates missing from the independent dataset used for validation - the validation dataset. Discovering that very few prognostic models are externally validated, possibly due to a lack of guidelines on appropriate methods, Chapter 8 describes a detailed simulation study aimed at identifying methods of external validation and methods of handling covariates missing from the validation dataset. The so-called 'imputation' methods are derived from methods used to handle missing data within a covariate. Three methods of external validation are examined – discrimination via the c-statistic, calibration via deviance, and prognostic separation via Royston and Sauerbrei's method.

The parameters used for the simulation study are based upon those derived in Chapter 4, from The SANAD Trial.

1.5.8 Externally Validating Prognostic Models – Case Study

Having examined methods of external validation, and methods for handling a covariate missing from the validation dataset in Chapter 8, Chapter 9 applies the methods to real data. The model derived in Chapter 5 looking at risk of a second seizure following a first ever seizure, derived using the MESS data, is used as the development dataset. Three independent datasets are used as validation data. One of the validation datasets was collected in the U.K., one in Italy and one in Western Australia.

The three methods of external validation considered in Chapter 8 are examined. Additionally, the Italian dataset is missing a covariate so the five imputation methods also described in Chapter 8 are tested on this dataset. Using all these methods it is possible to decide if the MESS model from Chapter 5 is a clinically useful tool.

1.5.9 Prognostic Models for Remission – Mixture Modelling

In Chapter 7 the internal validity of the model for time to 12 month remission, derived using The SANAD Trial in Chapter 4, was found to be fairly poor at one year. This may be because of the heterogeneity of the remission outcome for patients with epilepsy – some patients will achieve 12 month remission immediately, some will achieve 12 month remission but not immediately, and some may never achieve remission.

In Chapter 10, two types of mixture modelling are examined to handle the varying types of remission that a patient can achieve. First, logistic regression is used to model immediate remission and survival analysis is used to examine time to 12 month remission conditional on not achieving it immediately. Second, a three component mixture model is derived to model all three types of remission simultaneously.

1.5.10 Conclusions and Further Work

Chapter 11 uses the work of Chapters 2 to 10 to suggest recommendations for research and recommendations for practice.

Recommendations for research can be split into model development and model validation. The main suggestions for model development include interpreting hazard ratios and confidence intervals for large numbers of interactions (Chapter 4), deriving a likelihood function for a mixture model and using it to obtain risk estimates (Chapter 10), and assessing model validation of a mixture model (Chapters 7-10). The key ideas for model validation include assessing alternative scenario for external validation and handling a missing covariate (Chapter 8), assessing alternative methods of external validation (Chapter 8), and producing a definition of plausibly related datasets (Chapter 8).

Further research needs to include external validation of the models derived in Chapter 4 using The SANAD Trial, and the combining of all four datasets examined in Chapter 9, to

give a more accurate estimate of recurrence risk as well as suggesting improved driving guidelines.

Chapter 2: Introduction to Epilepsy

2.1 Introduction

This chapter begins with an overview of epilepsy (section 2.1). Various seizure types are then described (section 2.2) including those associated with focal epilepsies (section 2.2.1) and those with generalised epilepsies (section 2.2.2). Section 2.3 then summarises frequent methods of investigation and diagnosis before moving on to look at treatment options in section 2.4.

The impact of epilepsy on a person's life is considered in section 2.5, especially with regards to driving (section 2.5.1) – following a seizure a person must refrain from driving. This is true for all types of seizures, whether or not the person has a diagnosis of epilepsy or not. The issue of seizures and driving is considered in detail in Chapters 5 and 6. The chapter concludes with a literature review aimed at identifying prognostic factor studies in epilepsy (section 2.6) before outlining the clinical motivation for the thesis (section 2.7).

Epilepsy is one of the most common disorders of the brain [58]. It accounts for 1% of the global burden of disease, a figure equivalent to breast cancer in women and lung cancer in men [59]. Worldwide, epilepsy affects 50 million people [58], 80% of whom are in developing countries. Within the United Kingdom, one in 131 people, approximately 456,000, have epilepsy, making it the most common serious neurological condition [60]. Every day in the UK, 75 people are diagnosed with epilepsy.

The definition of epilepsy requires the occurrence of two unprovoked seizures [61] and is the tendency to have recurrent seizures due to abnormal electrical discharge from central nervous system neurons [62]. The estimated prevalence of epilepsy is 0.5% to 1% in the developed world [63] and the overall incidence of epilepsy is 40 to 70 cases per 100,000

per year in industrialised countries and 100 to 190 cases per 100,000 per year in resource-poor countries [64].

Although most people refer to 'epilepsy', it is more accurate to talk about 'the epilepsies' because epilepsy is a heterogeneous condition [58]; for example some types of epilepsy are associated with childhood onset while some will only start in later life. Anyone can develop epilepsy and although males are slightly more likely to develop it than females it is not known why [65, 66]. Children and adults over the age of 60 are also more likely to develop epilepsy than young adults [67].

Approximately three percent of the population will have epilepsy at some point in their lives [65] while 5% will have a single seizure at some time [68]. The majority of people with epilepsy will go into remission, however up to 30% will develop refractory, or drug-resistant, epilepsy and continue to have seizures despite treatment with combinations of antiepileptic drugs [69, 70]. Only 52 per cent of people with epilepsy in the UK are seizure free [71] although it is estimated that 70 per cent could be seizure free with the right treatment [72].

Sometimes the reason epilepsy develops is clear; it can be related to brain damage caused by surgery, a severe head injury, a stroke, or an infection of the brain, such as meningitis or encephalitis. Very occasionally, the cause is a brain tumour. For six out of ten people, there is no known structural cause [73], and for many of these the cause is thought to be genetic.

2.2 Seizure Types

Neurons, or specialised nerve cells, normally generate and transmit electrochemical impulses that act on other neurons, glands, and muscles to produce human thoughts, feelings, and actions. In epilepsy, the normal pattern of neuronal activity becomes

disturbed, causing strange sensations, emotions, and behaviour, or sometimes convulsions, muscle spasms, and loss of consciousness. During a seizure, neurons may fire as many as 500 times a second, much faster than normal. In some people, this happens only occasionally; for others, it may happen hundreds of times a day.

Seizures can arise in any part of the cerebral cortex and patient experiences during a seizure will depend on where in the brain the seizure is happening. There are around 40 different types of seizure and a person may have more than one type [74].

In 1981, the International League Against Epilepsy established standardised classifications and terminology for epileptic seizures [75] which provided a universal vocabulary that not only facilitated communication among clinicians, but also established a foundation for the performance of quantitative clinical and basic research on epilepsy. There have been subsequent iterations of these classifications the last of which was published in 2010 [76]. According to this, Figure 1 summarises possible symptoms according to seizure type.

Seizures can be categorised according to which part, or parts, of the brain the epileptic activity starts in, and how widely and rapidly it spreads. Partial seizures, also known as focal seizures, involve epileptic activity that starts in just one part of the brain. Sometimes, epileptic activity that starts as a partial seizure can spread to the rest of the brain. When this happens, the seizure is known as secondary generalised. Conversely, in generalised seizures, the seizures involve both cerebral hemispheres from the onset.

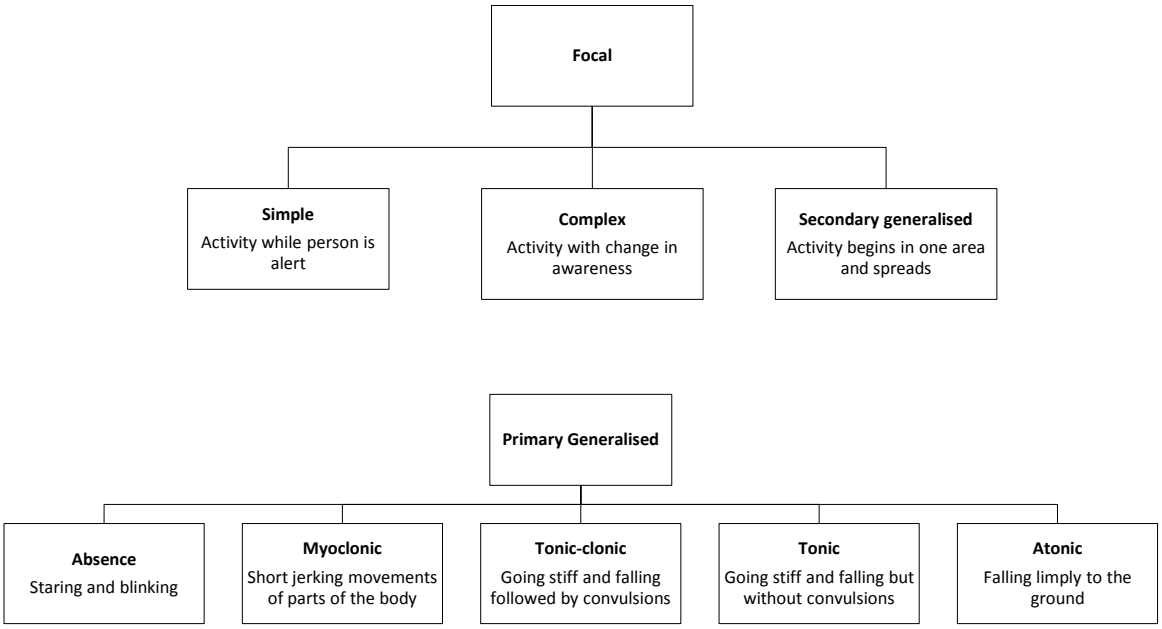


Figure 1: Seizure map and brief summary of possible symptoms

2.2.1 Focal Epilepsies

Within the focal epilepsies, the epilepsy is classified according to the site of onset and the aetiology. Regarding the site of onset, this is usually classified according to which lobe the seizure starts in. The brain has four lobes - temporal, frontal, occipital or parietal, as shown in Figure 2. Therefore, it may be possible to classify a patient as having complex partial seizures as part of a temporal lobe epilepsy.

Regarding aetiology, there are three main categories; symptomatic where there is a known cause for a person's epilepsy, such as a head injury; idiopathic epilepsy where, despite investigation, no apparent cause for the epilepsy can be found and a genetic aetiology is assumed; and cryptogenic epilepsy for which, like idiopathic epilepsy, no apparent cause can be found however, there is strong evidence to suggest that this type of epilepsy may be the result of brain damage. For example, a patient with an epilepsy arising from a temporal lobe tumour would be considered as having a symptomatic temporal lobe epilepsy.

Partial (focal) seizures can happen as a result of epileptic activity in any area of the brain, although the most common areas are the temporal lobes [73]. This is often as a result of mesial temporal sclerosis, the aetiology of which is poorly understood, or other pathologies such as head injury, or infection such as meningitis.

Partial seizures are the most common type of seizures experienced by people with epilepsy. Virtually any movement, sensory, or emotional symptom can occur as part of a partial seizure, including complex visual or auditory hallucinations. There are three main categories of partial seizures: simple, complex and secondary generalised.

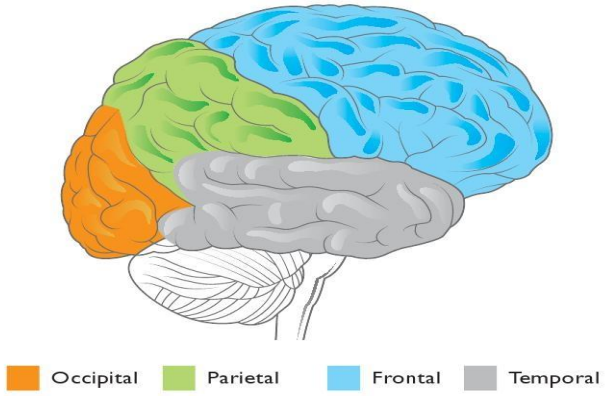


Figure 2: The areas of the brain

2.2.1.1 Simple or Complex Partial Seizures

Partial seizures are subdivided into simple partial seizures and complex partial seizures. People who have simple partial seizures do not lose consciousness during the seizure. However, some people, although fully aware of what is going on, find they cannot speak or move until the seizure is over. They remain awake and aware throughout. Sometimes they can talk quite normally to other people during the seizure and they can usually remember exactly what happened to them while it is going on. However, simple partial seizures can affect movement, emotion, sensations, and feelings in unusual and sometimes even frightening ways.

Complex partial seizures affect a larger area of the brain than simple partial seizures and they affect consciousness. During a complex partial seizure, a person cannot interact normally with other people, is not in control of his or her movements, speech or actions, does not know what he or she is doing and cannot remember afterwards what happened during the seizure. Although complex partial seizures can affect any area of the brain, they often take place in one of the brain's two temporal lobes.

Although someone may appear to be conscious because he or she remains standing with eyes open and moving about, it will be an altered consciousness - a dreamlike, almost trancelike state. A person may even be able to speak, but the words are unlikely to make sense and he or she will not be able to respond to others in an appropriate way.

All seizures have a tendency to cluster. This will, however, be more obvious for seizures that occur more frequently such as simple partial seizures and complex partial seizures.

2.2.1.2 Secondary Generalised Seizures

Partial seizures may be followed by a generalised seizure, called secondary generalisation, which causes loss of consciousness. Secondary generalisation occurs when a partial seizure spreads and activates the entire cerebrum bilaterally. Activation may occur so rapidly that the initial partial seizure is not clinically apparent or is very brief.

2.2.2 Generalised Epilepsies

A number of generalised epilepsies are recognised, and they are classified according to the seizure types the patient experiences, EEG changes and other characteristics such as learning disability. The majority of the generalised epilepsies are classified as idiopathic and are believed to have a genetic aetiology. Examples include childhood absence epilepsy, juvenile absence epilepsy, juvenile myoclonic epilepsy and generalised epilepsy

with tonic-clonic seizures on awaking. The main primary generalised seizure types are absence, myoclonic, tonic-clonic, tonic and atonic.

2.2.2.1 Absence Seizures

Absence seizures mainly happen in childhood [77]. During an absence seizure, the child appears to be daydreaming or switching off. However, in an absence seizure, the child cannot be woken up. They are very briefly unconscious and unaware of what is happening around them. Because most children tend to daydream at times, absences can be very hard to spot. In some cases, children can be having hundreds of absence seizures a day. This prevents them from learning and fully taking part in school or family activities. These children are missing out on tiny pieces of information. For example, they might hear the first part of a sentence but not the end. They may hear the instruction to go out and play, but not when to be back. Because they might then not do what is expected of them, this may be mistaken for poor behaviour.

2.2.2.2 Myoclonic Seizures

Most people experience a sudden jerk as they are falling asleep. Although this jerk is not epilepsy, it is similar to the jerks experienced by someone who has myoclonic seizures. The term myoclonic comes from 'myo' meaning muscle, and 'clonus' meaning jerk. Myoclonic seizures can affect the whole body, but are usually restricted to one or both arms and sometimes the head. During the seizure, the person is not conscious, but the seizure is so brief that the person appears to remain fully conscious. Myoclonic seizures may occur as a single seizure or a cluster of seizures.

2.2.2.3 Tonic-Clonic Seizures

This is the most common and widely recognised generalised seizure. A tonic-clonic seizure affects both cerebral hemispheres from the onset. There are two phases to this type of seizure: the 'tonic' phase followed by the 'clonic' phase. In the tonic phase the person goes

stiff because all their body's muscles contract, the person loses consciousness and falls to the floor and breathing patterns change with the result that there is less oxygen than normal in the lungs. Because of this, the blood circulating in the body is less red than usual, causing the skin (particularly around the mouth and under the finger nails) to appear blue in colour. This is called 'cyanosis'. The person may bite their tongue and the inside of their cheeks.

After the tonic phase has passed, the clonic phase of the seizure begins. The limbs jerk because now the muscles tighten and relax in turn. After a further minute or so, the muscles relax and the person's body goes limp. Slowly they will regain consciousness, but may well be groggy or confused. They will gradually return to normal but may not be able to remember anything for a while. Very often, the person remains sleepy and may have a headache and aching limbs. Recovery times differ from person to person.

2.2.2.4 Tonic Seizures

Tonic seizures result in all the muscles tightening. The body stiffens and the person will fall over unless they are supported. Tonic seizures usually last less than 20 seconds and most often happen during sleep.

2.2.2.5 Atonic Seizures

Atonic seizures are also called drop attacks or akinetic seizures. In an atonic seizure, all muscle tone is lost and the person drops heavily to the ground. The seizure is very brief and the person is not confused afterwards and is usually able to get up again straight away. However, because the person's body usually falls forward in an atonic seizure, they are at risk of banging their head on furniture or other hard objects.

2.2.3 Status Epilepticus

Status epilepticus is defined as a condition in which seizures occur so frequently that the patient does not fully recover from one seizure before having another. Status epilepticus is also defined as a single prolonged seizure. A seizure lasting more than five minutes – especially a tonic-clonic seizure in an older child or adult – should be considered status epilepticus.

Status epilepticus is a medical emergency because the longer the seizure lasts the less likely it is to stop on its own. It is also less likely to stop with emergency medicines. Status epilepticus can lead to brain damage, or even death.

Several types of status epilepticus exist depending on seizure type: tonic-clonic, simple partial, complex partial, and absence. Tonic-clonic status epilepticus is the most common and most life-threatening type.

2.3 Investigations and Diagnosis

Diagnosis is mainly based on the description given of the seizure(s) and, if possible, an eye witness account. There is no single test to determine if someone does or does not have epilepsy and frequently, when epilepsy begins, the tests are normal. However, there are various hospital tests that can help clinicians make a diagnosis, although not everyone needs to have every test.

In the first case a clinician will want to determine if a patient has had a seizure. Answering this is almost entirely dependent upon getting a good description of the seizure(s). The next question might involve determining if the seizures were focal or generalised. An electroencephalogram (EEG) may be used to answer this. Finally, the clinician will probably want to investigate why the person had a seizure. The EEG can again help with this if it shows features of an idiopathic generalised epilepsy. Magnetic Resonance Image (MRI)

brain scans can also help answer this question as they identify abnormalities that have caused the epilepsy. These methods will be explained further in sections 2.3.1 and 2.3.2.

2.3.1 Electroencephalogram

The electroencephalogram, or EEG, gives clinicians information about the activity of the brain. However an EEG only shows what is happening in the brain when the test is being done - it is not able to show what has already happened, or what is going to happen in the future.

Although an EEG is localised, it integrates neuronal activity over a very large number of neurons which can make EEGs difficult to interpret. Additionally, EEGs are reported in a qualitative way which makes it challenging to standardise reporting. Typical examples of EEGs can be seen in Chapter 5.

During an EEG, a technician places electrodes on the scalp which are then connected to the EEG machine that records the electrical signals in the brain. During the EEG test, a person is asked to do things such as open and close their eyes, breathe deeply and look at flashing lights. If seizures appear to be connected to sleep patterns, clinicians may suggest doing an EEG during sleep or after a period of sleep deprivation.

Some people have an abnormal EEG but it does not necessarily mean they have epilepsy. Similarly, a normal EEG does not necessarily mean that someone does not have epilepsy.

2.3.2 Magnetic Resonance Imaging

The magnetic resonance imaging, or MRI, uses radio frequency waves and a magnetic field. It can show if there is a structural cause for the epilepsy.

Most of the human body is comprised of water molecules comprising hydrogen and oxygen atoms. Within the centre of each hydrogen atom is a particle called a proton. These

protons are very sensitive to magnetic fields. Therefore, when the magnets within the MRI scanner are switched on, the protons in the patient's body are pulled towards the same direction. Short bursts of radio waves to certain areas of the body pull the protons out of position. As this happens, each proton transmits a radio signal that provides its exact location in the body. The radio signals of millions of protons can be collected together and combined to create a detailed image of the inside of the body [78].

An MRI is not always used for people who have generalised seizures. This is because generalised epilepsy is unlikely to be caused by damage to a part of the brain's structure. In this situation the scan is unlikely to add anything to the diagnosis.

2.4 Treatment

The most common way epilepsy is treated is with antiepileptic drugs (AEDs). AEDs do not cure epilepsy, but are designed to try and stop seizures happening. AEDs make the brain more resistant to the spread of the abnormal electrical activity [79].

There are many different AEDs available and the most appropriate drug is chosen based on the type of seizures and the person. It is common practice to start the drug at a low dosage, referred to as a loading dose, with a gradual increase, to keep side effects to a minimum. However, everyone is different, and some people find that they are particularly sensitive to a drug, even at low dosages.

Wherever possible, clinicians treat epilepsy with one drug. Where seizures prove difficult to control with one AED, more than one drug may be considered. Different AEDs have different mechanisms of action and as a result of this, combining drugs, so called polytherapy or adjunct therapy, can often lead to improved seizure control.

Each AED has a recommended maintenance, or daily, dosage range that is very different, depending on the drug. For example, 500 milligrams (mg) of lamotrigine is the recommended maximum dose [80], whereas for sodium valproate it is 2500 mg [81]. However, this is only a guide for clinicians when prescribing, because some people may need dosages outside the usual range. Common antiepileptic drugs can be seen in Table 2 together with their abbreviation, year of introduction and primary indication [82]. The drugs listed here do not represent a comprehensive list of all antiepileptic drugs. However, these are the drugs used in The SANAD Trial (Chapter 4) and the MESS study (Chapter 5).

The length of time taken to gain control of the seizures varies from person to person. The majority of people find the right AED and dosage very quickly. However, for some people it can take a lot longer to find the right drug or drugs at the right dosage(s). About 30% of people with epilepsy have epilepsy that is difficult to control with AEDs [69, 70].

Table 2: Common Antiepileptic Drugs as used in SANAD and MESS

Generic Drug Name	Abbreviation	Date of U.K. Licence	Primary Indication
Carbamazepine	CBZ	1965	Focal and secondary generalised tonic-clonic seizures, primary generalised tonic-clonic seizures
Gabapentin	GBP	1993	Monotherapy and adjunctive treatment of focal seizures with or without secondary generalisation
Lamotrigine	LTG	1991	Monotherapy and adjunctive treatment of focal seizures and generalised seizures including tonic-clonic seizures; seizures associated with Lennox-Gastaut syndrome; monotherapy of typical absence seizures in children
Oxcarbazepine	OXC	2000	Monotherapy or adjunctive treatment of focal seizures with or without secondary generalised tonic-clonic seizures
Phenytoin	PHT	1938	All forms of epilepsy except absence seizures
Topiramate	TPM	1995	Monotherapy and adjunctive treatment of generalised tonic-clonic seizures or focal seizures with or without secondary generalisation; adjunctive treatment of seizures associated with Lennox-Gastaut syndrome
Valproate	VPS	1993	All forms of epilepsy

2.4.1 Non-Drug Treatments

Most people with epilepsy take AEDs for their epilepsy and these can be very effective. Other treatments that are scientifically proven to work are surgery, vagus nerve stimulation and the ketogenic diet [79].

2.4.1.1 Resective Surgery

Only a small number of people with epilepsy are suitable for resective surgery. Clinicians will usually consider surgery for epilepsy if a person is finding it difficult to get control of their seizures with AEDs. This is most likely to be if a person has tried lots of different AEDs and taken a combination of more than one type.

Surgery is not an option for everyone with epilepsy. This may be because there is no obvious structural cause for the epilepsy, because the exact site of seizures cannot be identified, or because to operate would be too great a risk.

2.4.1.2 Vagus Nerve Stimulation

Vagus Nerve Stimulation, VNS, is a treatment for epilepsy where a small generator is implanted under the skin below the left collar bone. This generator is connected to a lead with three coils at one end, which are wrapped around the vagus nerve in the left side of the neck [79]. The VNS stimulates the vagus nerve at intervals to reduce the frequency and intensity of seizures. VNS doesn't necessarily work straight away, taking from a couple of months to two years to see a noticeable reduction.

Although it is rare for a person to become seizure free as a result of VNS, many people have fewer or less severe seizures and therefore report a better quality of life overall [83]. Some people experience no change in seizure control.

2.4.1.3 The Ketogenic Diet

Since it was first used in the 1920s, this diet has been a controversial therapy for children with difficult to control epilepsy. It can, however, be very good at controlling seizures [84, 85]. Interest in the diet has grown again in recent years and it is achieving increasingly widespread use. Although some children find the diet unpleasant and difficult to follow, other children manage it very well. The ketogenic diet is high in fats, low in carbohydrates and adequate in protein.

In the small number of studies done so far, it seems that the ketogenic diet could also be good at controlling seizures in adults [86]. The difficulty is that most adults find the diet unpleasant and too strict to keep to.

2.5 Living with Epilepsy

Having seizures, or being diagnosed with epilepsy, can affect people in different ways. Some people feel relieved to be given a name and treatment for their condition. However, one of the biggest concerns for people with epilepsy is how it affects day to day life.

Many people with epilepsy do lead full lives. However, it is fairly common for people with epilepsy to find it hard to remember things as epileptic seizures can affect memory. If there is damage to the brain that causes the epilepsy, this can also cause memory problems as may the antiepileptic medication itself. Depression is also common which may be because of the epilepsy medication or the epilepsy itself. However, someone with epilepsy may have to change things about their life that they did not want to and other people may treat a person differently because of their epilepsy.

With the right support and relevant safety precautions, there is little that someone with epilepsy should need to avoid. Many people with epilepsy have their seizures completely

controlled by antiepileptic medication and do not need to take any greater safety precautions than anyone else be it at home, or out and about.

Most people with epilepsy can participate fully in school but at times, seizures or side effects of seizure medicine may interfere with schoolwork. Additionally, having seizures may make it more difficult to find a job or to work at certain occupations.

With the exception of the armed forces it is against the law for employers to discriminate against people with epilepsy. However, if the epilepsy is not well controlled there will be some careers such as working at heights, becoming an airline pilot, train driver, policeman etc that are not appropriate.

2.5.1 Epilepsy and Driving

For people who have epilepsy or seizures, there are specific regulations about holding a driving license. The regulations depend on the type of license held (Group 1, car drivers, or Group 2, HGV drivers), and when the person has seizures (as some people have seizures only while they are asleep). There are additional regulations for individuals who have a single (one-off) seizure.

For people who drive, one immediate effect of having a seizure is that they have to stop driving. This is true for all types of seizures, whether or not the person has a diagnosis of epilepsy or not. For many people this can have a big impact on their life, and be very difficult or upsetting.

The effect of seizures on driving is further considered in Chapters 5 and 6.

2.6 Prognostic Factor Studies in Epilepsy

The choice of a first antiepileptic drug is arguably the most important decision that a patient with epilepsy will make as the majority of patients achieve seizure control on that

drug, possibly at the consequence of adverse effects. It is thus important to have reliable evidence about the likely risk-benefit ratio of potential treatments to inform treatment choice. For chronic conditions such as epilepsy this evidence will largely come from head to head randomised controlled trials in which long term outcomes have been assessed. Relatively few head to head trials have been undertaken in epilepsy with the majority of those that have been carried out being industry sponsored studies that have tended to recruit selected populations, make treatment comparisons that do not reflect every day treatment choices, and have patient follow-up that is typically less than 12 months such that long term seizure control outcomes cannot be assessed.

The epilepsies are a heterogeneous group of disorders with varied outcome and a number of clinical factors have been shown to influence outcome. In a search of the literature we found 54 studies that identified prognostic factors for epilepsy. Studies of interest considered seizure outcomes for patients with either a single seizure only, with at least two seizures but newly diagnosed epilepsy, or patients with refractory epilepsy. Refractory epilepsy was defined as per the study, or at least two years without seizure control despite at least two antiepileptic drugs at optimal dose either as monotherapy or polytherapy.

28 studies considered prognostic factors for children with epilepsy. Nine of these studies looked at prognostic factors for the risk of seizure recurrence after withdrawal of antiepileptic drugs. The most frequently reported prognostic factors in these studies were aetiology, seizure type, age and EEG result [87-94]. Another 12 of the 28 studies considered prognostic factors for seizure outcomes, three of which focussed on particular seizure types with the most common factors being aetiology, seizure frequency, seizure type and neurological/mental retardation [95-103]. Three of the remaining studies in children considered remission in newly referred patients with the prognostic factor aetiology being recorded for all studies [104-106]. The final four studies in children

considered, in one case, remission after a second drug failure with prognostic factors seizure type and frequency [107], and three considered seizure outcome for medically intractable epilepsy with the common prognostic factor being history of status epilepticus [108, 109].

Eight studies focussed on seizure outcomes for either a specific seizure type or specific type of epilepsy. Of these, three studies considered prognostic factors for patients with absence seizures with the most common factor being IQ level, also described as mental retardation [110-112], while two studies considered prognostic factors for patients with juvenile myoclonic seizures. No prognostic factor was common to both studies [113, 114]. The other three studies focussed on either a specific epilepsy type or seizure type. For patients with mesial temporal lobe epilepsy prognostic factors for febrile seizure history, mesial temporal sclerosis, age at onset and EEG result [115] were identified. For symptomatic focal epilepsy prognostic factors seizure type, nocturnal seizures, history of status epilepticus and EEG result [116] were identified. Finally, for complex partial epilepsy prognostic factors history of neonatal seizures, time between seizures, EEG result, and style of therapy (monotherapy or polytherapy) [117] were found to be significant.

Six studies considered patients who were newly diagnosed with epilepsy. In most cases there was little overlap in the identified prognostic factors, however number of seizures was identified by three studies [118-120]. Five studies looked at seizure outcomes for patients withdrawing their antiepileptic drug therapy. Of these the most commonly reported prognostic factors were duration of remission and seizure type [121-123].

Seven other studies identified prognostic factors for epilepsy, one of which was concerned with prognostic factors for intractability - aetiology, neurological development, seizure frequency, history of status epilepticus and effect of AED therapy [124]. Two other studies considered prognostic factors for long-term retention of treatment on either lamotrigine or

topiramate. The factors for lamotrigine were seizure type and treatment [125] while for topiramate they were learning disability, age at onset, treatment history, style of treatment and dosage [126]. Seizure type was found to be a prognostic factor for seizure outcome in a study comparing phenytoin and valproate in adults [127]. The final three studies all looked at partial epilepsy and common prognostic factors for seizure outcomes in these were EEG result and seizure type [128-130].

2.7 Clinical Motivation

Few prognostic models based on prospective cohorts or randomised controlled trial data have been published in epilepsy. Those that have, include models from the National General Practice Survey of Epilepsy (NGPSE) [131], the MRC antiepileptic drug withdrawal study [119, 132] and the Multicentre study of early Epilepsy and Single Seizures (MESS) [133]. The NGPSE models identify patient characteristics that modify seizure recurrence while the MRC antiepileptic withdrawal study provides important predictors of risk of seizure recurrence with and without AED withdrawal for patients in remission from seizures. However, many patients in NGPSE were not prescribed antiepileptic drugs. For patients with single seizures and early epilepsy the MESS models predict the risk of seizure recurrence for the policies of immediate and delayed AED treatment.

Although a number of clinical and population based studies have been undertaken to assess outcome and predictors of outcome in epilepsy, as shown in section 2.6, few predictive models have been published that can be used to predict an individual patient's outcome in every day practice. This may be due to the sample of patients studied, the specifics of the outcome considered or the size or design of the study. Additionally no prognostic models represent an epilepsy cohort accrued at the start of treatment.

There is clearly a need for prognostic models in epilepsy not only for patients who have had a first seizure but also for patients who are attempting to achieve remission from seizures and for those wishing to withdraw from epilepsy treatment following a period of remission.

Two prognostic models for patients with newly diagnosed epilepsy are presented in Chapter 4. A prognostic model for the risk of seizure recurrence following a first ever seizure in the context of driving is developed in Chapter 5 while in Chapter 6, a model is developed for patients wishing to withdraw from treatment following a period of remission, again in the context of driving.

Any prognostic model, which is to be implemented in general practice, needs to be validated both internally and externally. Internal validation of models is addressed in Chapter 7 while external validation is considered via a simulation study and case study in Chapters 8 and 9 respectively. Finally, more complex prognostic models, which combine patients at all stages of the condition simultaneously, are considered via mixture modelling in Chapter 10.

Chapter 3: Methods for Identifying Prognostic Factors for Epilepsy

3.1 Introduction

In order to fully investigate potential prognostic factors for patients with epilepsy, and thus develop an appropriate prognostic model, various techniques and methods need to be considered such as curve fitting, variable centring, model selection, assumption testing, treatment-covariate interactions and competing risks. This chapter will provide an overview of this theory. The methods described in this chapter will be used in subsequent chapters.

Handling continuous predictors via splines and fractional polynomials will be examined in section 3.2. In section 3.3 variable centring will be discussed while methods for model building and development will be presented in section 3.4. Methods for testing the proportional hazards assumption for the Cox model will be addressed in section 3.5 before competing risks methods are described in section 3.6. The chapter will conclude (section 3.7) with a description of the methods used in Chapters 4, 5 and 6.

3.2 Handling Continuous Predictors

In the past, a linear function has been an acceptable model for data-dependent model building with continuous covariates. This may not be appropriate in applications where the data contain strong evidence against linearity. In these situations a cutpoint model may be used. Since then, further guidance has been provided suggesting spline model fits or fractional polynomial transformations.

The cutpoint model (section 3.2.1) is included for completeness only. Continuous predictors are investigated in Chapter 4 via fractional polynomial (section 3.2.3) transformations. The knot positions from a potential spline model fit (section 3.2.2) are used to categorise the continuous predictors for presentation only, post analysis.

3.2.1 Cutpoint Model

In the cutpoint model, a change point is selected and then the dichotomised or categorical variable is used in a univariate analysis. More specifically, denote the covariate of interest by Z. If this covariate has been measured on a quantitative scale, the proportional hazards [31] cutpoint model is defined as

$$\lambda(t|Z > \mu) = \exp(\beta)\,\lambda(t|Z \leq \mu),\ t > 0$$

where

$$\lambda(t|\cdot) = \lim_{h\to 0}\frac{1}{h}\Pr(t \leq T < t + h|T \geq t,\cdot)$$

denotes the hazard function of the event-free survival time random variable T [43]. The parameter $\theta = \exp(\beta)$ is referred to as the hazard ratio of observations and is estimated through $\hat{\theta} = \exp(\hat{\beta})$ by maximising the corresponding partial likelihood [31] with given cutpoint μ. μ is usually unknown and hence a problem of model selection where the cutpoint μ also has to be estimated from the data.

A popular approach for this type of data-dependent categorization is the 'minimum p value' method where, within a certain range of the distribution of Z, called the selection interval, the cutpoint $\hat{\mu}$ is taken such that the p value for the comparison of observations below and above the cutpoint is a minimum [43]. This method, also referred to as the optimal cutpoint method, may lead to a severe overestimation of the absolute value of the log-hazard ratio, which is equivalent to an inflation of the type 1 error rate [134]. In order

to correct for the overestimation of the absolute value of the log-hazard ratio, Verweij [135] proposes shrinking the parameter estimates by a shrinkage factor, c. In the cutpoint model the log-hazard ratio should then be estimated by $\hat{\beta}_{cor} = \hat{c} \cdot \hat{\beta}$ where $\hat{\beta}$ is based on the minimum p value method and $\hat{c}$ is the estimated shrinkage factor. In a recent paper by Schumacher [136] several methods to estimate $\hat{c}$ were compared.

As a somewhat *ad hoc* approach, the shrinkage factor can be obtained using the corrected P-value. A second approach is based on cross-validation calibration with leave-one-out, 10-fold or bootstrap cross-validation methods. The final approach examined by Schumacher used a heuristic estimator described by van Houwelingen and Le Cessie [137].

Even with the correction via a shrinkage factor, the optimal cutpoint approach still has disadvantages; in particular, different studies will most likely yield different cutpoints, making comparisons across studies extremely difficult or even impossible. Thus other approaches should be preferred.

3.2.2 Spline Functions

Using spline functions in regression models enables modelling of continuous covariates. For fitting functions of a single variable, the general principle is that the domain of a covariate is covered by knots and local polynomial pieces are anchored at these knots. Available techniques differ in the number of knots used, the approach for determining knot positions and in the way the parameters for the polynomial pieces are estimated [138].

The simplest spline function is a linear spline function, a piecewise linear function. Suppose that the x-axis is divided into intervals with endpoints at a, b, and c called knots. The linear spline function is given by $f(X) = \beta_0 + \beta_1 X + \beta_2 (X - a)_+ + \beta_3 (X - b)_+ + \beta_4 (X - c)_+$ where $(u)_+ = \begin{cases} u, & u > 0 \\ 0, & u \leq 0 \end{cases}$ [139].

Although the linear spline is simple and can approximate many common relationships, it is not smooth and will not fit highly curved functions well. Alternatives include smoothing splines which place one knot at each unique covariate value and use a roughness penalty for parameter estimation. With regression splines, only a small number of knots are used, with the advantage that parameter estimation can be performed by standard regression methods. Regression splines can be extended by using a large number of knots combined with penalised parameter estimation [140] to make the exact knot position less important. Such procedures are closely related to smoothing splines and share the advantage that only a single smoothing parameter must be selected per covariate.

Cubic polynomials have been found to have good properties with an ability to fit sharply curving shapes. This ensures linearity in the tails of the distribution and thus avoids unrealistic end effects of the fitted functions. However, Stone and Koo [141] have found that cubic splines do have a drawback in that they can be poorly behaved in the tails, that is before the first knot and after the last knot. They cite advantages of constraining the function to be linear in the tails. Their restricted cubic spline function has the additional advantage that only $k-1$ parameters must be estimated, besides the intercept, as opposed to $k+3$ parameters with the unrestricted cubic spline.

The restricted cubic spline function with k knots $t_1, \dots, t_k$ is given by [142]

$$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots \beta_{k-1} X_{k-1}$$

where $X_1 = X$ and for $j = 1, \dots, k-2$,

$$X_{j+1} = (X - t_j)_+^3 - \frac{(X-t_{k-1})_+^3 (t_k - t_j)}{t_k - t_{k-1}} + \frac{(X-t_k)_+^3 (t_{k-1} - t_j)}{t_k - t_{k-1}}.$$

Stone [143] found that the location of knots in a restricted cubic spline model is not very crucial in most situations. The fit depends much more on the choice of k, the number of

knots. Placing knots at fixed quantities, percentiles, of a predictor's marginal distribution is a good approach in most datasets. This ensures that enough points are available in each interval and also guards against letting outliers overly influence knot placement. Recommended equally spaced quantiles are shown in Table 3.

Table 3: Recommended equally spaced quantiles for placement of between three and seven knots (k) with restricted cubic spline models

k	Quantiles						
3	0.10	0.50	0.90				
4	0.05	0.35	0.65	0.95			
5	0.05	0.28	0.50	0.73	0.95		
6	0.05	0.23	0.41	0.59	0.77	0.95	
7	0.03	0.18	0.34	0.50	0.66	0.82	0.98

Stone has found that more than five knots are seldom required in a restricted cubic spline model [143]. The principle decision is then between $k = 3, 4$ or 5. For many datasets, $k = 4$ offers an adequate fit of the model and is a good compromise between flexibility and loss of precision caused by overfitting a small sample. When the sample size is large (i.e. $n > 100$ with a continuous uncensored response variable), $k = 5$ is a good choice. A linear spline and a cubic spline, each with five knots, can be seen in Figure 3.

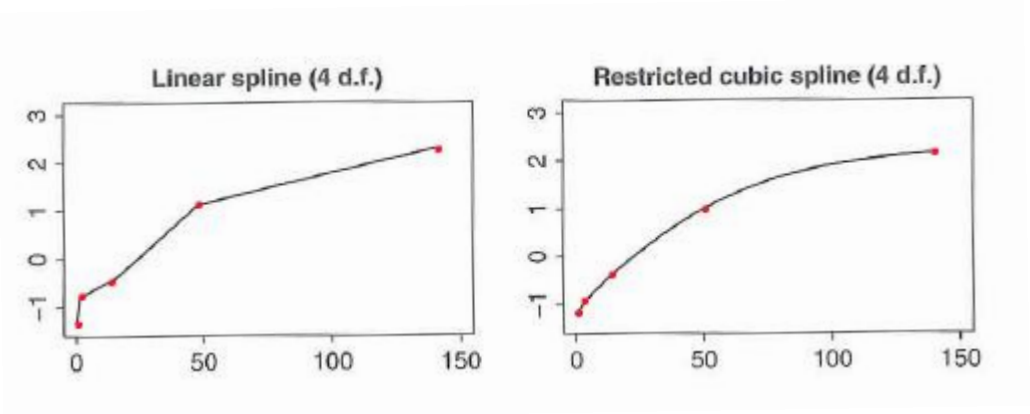


Figure 3: Linear spline and cubic spline with five knots (4 degrees of freedom)

3.2.3 Fractional Polynomials

An approach to prognostic modelling of survival data that provides more flexibility than polynomials [144] while preserving the simplicity of the final model to an acceptable degree was originally developed by Royston and Altman [144] and is termed the fractional polynomial (FP) approach. One, two or more power transformations of the form x^p are fitted, the exponent(s), p, being chosen from a small, preselected set $S = \{-2, -1, -0.5, 0, 0.5, 1, 2, 3\}$ where x^0 is defined as $\log x$. The set S includes the most commonly used power transformations and has the advantages of stabilising a function and enabling transportability to other settings.

An integer suffix to FP denotes the degree of the fractional polynomial. An FP1 transformation of a positive argument $x > 0$ with power p is defined as x^p where $p \in S$ while an FP1 function or model is defined as

$$\varphi_1^*(x; p) = \beta_0 + \beta_1 x^p = \beta_0 + \varphi_1(x; p)$$

An FP2 transformation of x with powers $\mathbf{p} = (p_1, p_2)$, or for $p_1 = p_2$ ('repeated powers') is the vector $x^{\mathbf{p}}$ with

$$x^{\mathbf{p}} = x^{(p_1, p_2)} = \begin{cases} (x^{p_1}, x^{p_2}), & p_1 \neq p_2 \\ (x^{p_1}, x^{p_1} \log x), & p_1 = p_2 \end{cases}$$

while an FP2 function or model with parameter vector $\beta = (\beta_1, \beta_2)^T$ and powers $\mathbf{p}$ is

$$\varphi_2^*(x; \mathbf{p}) = \beta_0 + \beta_1 x^{p_1} + \beta_2 x^{p_2} = \beta_0 + x^{\mathbf{p}}\beta = \beta_0 + \varphi_2(x; p)$$

The constant, β_0, is optional and depends on the context. For example, it is usually included in a normal-errors model, but not in the Cox regression model for survival data. There are eight FP1 functions and 36 FP2 functions comprising eight 'repeated powers

functions' where $p_1 = p_2$ and 28 combinations of each of the eight members of S where $p_1 \neq p_2$. More than two terms are rarely required in practical applications.

The general definition of an FPm function with powers $\mathbf{p} = (p_1 \leq \cdots \leq p_m)$ is most easily written as a recurrence relation. Let $h_0(x) = 1$ and $p_0 = 0$. Then

$$\varphi_m^*(x;\mathbf{p}) = \beta_0 + \varphi_m(x;\mathbf{p}) = \sum_{j=0}^{m} \beta_j h_j(x)$$

where $h_j(x) = \begin{cases} x^{p_j} & \text{if } p_j \neq p_{j-1} \\ h_{j-1}(x)\log x & \text{if } p_j = p_{j-1} \end{cases}$ for $j = 1, \ldots, m$.

This extension of standard polynomials generates a considerable range of curve shapes while still preserving simplicity when compared to smoothing splines or other nonparametric techniques for example.

3.2.3.1 Choice of Origin

If non-positive values of x can occur, then a preliminary transformation of x to ensure positivity is needed. Common cases are where x is a count, when $\log(x+1)$ is traditionally used; where x is a positive random variable such as a physical quantity for which recorded values can be zero, due to imprecise measurement and/or rounding of observations; or where x is a difference or log ratio between two quantities. A simple solution is to choose a non-zero origin γ and work with $x' = x - x_{min} + \gamma$ where x_{min} is the smallest observed, or smallest possible, value of x. A possible choice of γ is the rounding interval of the sample values of x, or the minimum increment between successive ordered sample values of x [144]. If x is a count, then a natural choice is $\gamma = 1$ [145].

3.2.3.2 Function Selection Procedure

An obvious question to ask with FP modelling concerns how to select a suitable function for the data. The approach of Royston and Sauerbrei [145] has been developed to answer the

question in terms of a closed test procedure. In the following description of the functional selection procedure (FSP), a linear function is assumed as the default when x is either selected because of statistical significance or included in the model on *a priori* reasoning. The procedure runs as follows:

1. Test the best FP2 model for x at the α significance level against the null model using four degrees of freedom. If the test is not significant, stop, concluding that the effect of x is not significant at the α level. Otherwise continue.
2. Test the best FP2 for x against a straight line at the α level using three degrees of freedom. If the test is not significant, stop, the final model being a straight line. Otherwise continue.
3. Test the best FP2 for x against the best FP1 at the α level using two degrees of freedom. If the test is not significant, the final model is FP1; otherwise the final model is FP2. End of procedure.

The test at step 1 is of overall association of the outcome with x. The test at step 2 examines the evidence for nonlinearity. The test at step 3 chooses between a simpler or more complex nonlinear model. Before applying the procedure, the user must decide on the nominal P-value, α, and on the degree m of the most complex FP model allowed.

Holländer and Schumacher [146] showed, in a simulation study, that FSP preserves the family-wise type 1 error probability and generally ends up with a log-linear relationship if it is present.

3.2.3.3 Multivariable Fractional Polynomial

Sauerbrei and Royston [147] have extended the FP approach, proposing a model building strategy consisting of FP-transformations and selection of variables by backward elimination, called the multivariable FP approach (MFP). It combines variable selection by

backward elimination with determination of functional form for continuous predictors by the FSP [10, 11].

The MFP algorithm investigates, in a systematic way, whether the effect of a continuous covariate is better modelled by a non-linear function from the class of fractional polynomials (FP) or by a linear function.

3.2.3.4 Modelling Continuous Variables with a 'Spike' at Zero

Royston *et al.* [148] describe a procedure to deal with the 'spike' at zero problem which commonly arises in epidemiology and other fields. For example, a variable for number of cigarettes smoked per day will have a high number of zero responses. General fractional polynomial (FP) methodology does not address the question of how to model the factor when there is such a spike at zero. It is not always the case that the spike will represent genuine zero responses. For example, in laboratory assays some samples will fall below the limit of detection. In these cases zero-inflated mixtures are used such as the zero-inflated Poisson model [149].

The definition of epilepsy requires two unprovoked seizures. Therefore, in trials of newly diagnosed patients, there will be a higher number of patients with only two seizures. This is particularly relevant to the SANAD dataset and in Chapter 4, FPs and FPs with a spike (at two) will be considered when modelling total number of seizures before randomisation.

In the logistic regression setting, theoretical results [150, 151] suggest that a binary variable, say z, should be included in the model to represent exposure/non-exposure. However, the dose-response function for the exposed individuals still needs to be determined. For univariate dose-response modelling, Royston and Sauerbrei [145] suggested a new approach to the problem by extending FP modelling to include z. Their procedure comprises of two stages [148]: first, determine the best FP function when z is

included in the model; second, assess whether z or the FP component can be eliminated without harming the model fit.

3.2.3.4.1 FSP-Spike Procedure

Consider a model whose linear predictor, η, is given by

$$\eta = \begin{cases} \beta, & x = 0 \\ \beta_0 + \mathrm{FP2}(x + c; p_1, p_2), & x > 0 \end{cases}$$

where p_1 and p_2 are powers from the standard set S of FP transformations. The linear predictor η is an FP function of $x + c$ where $x > 0$ and a constant (β) when $x = 0$. Thus η is a discontinuous function of x with a possible jump at $x = 0$. The expression for η is equivalent to

$$\eta = \beta_0 + (\beta - \beta_0)z + (1 - z)\mathrm{FP2}^+(x + c; p_1, p_2)$$

where $z = \begin{cases} 1, & x = 0 \\ 0, & x > 0 \end{cases}$, and

$$\mathrm{FP2}^+(x + c; p_1, p_2) = \begin{cases} 0, & x = 0 \\ \mathrm{FP2}(x + c; p_1, p_2), & x > 0 \end{cases}.$$

The function selection procedure (FSP)-spike procedure for selecting a model has two stages. In the first stage, the most complex model comprising z and $\mathrm{FP2}^+(x + c; p_1, p_2)$ is compared with the null model on 5 degrees of freedom. If the test is significant, the steps of the FSP for selecting an FP function are followed, but with z always included in the model. In the second stage, which is performed separately, z, and the remaining FP or linear component are each tested for removal from the model. If both parts are significant, the final model includes both; if one or both parts are non-significant the one with the smaller deviance difference is removed. In the latter case, the final model comprises either the binary dummy variable or the selected FP function. If only the FP function is selected, the spike at zero plays no further part. Since the selection of an FP function may be

affected by the presence of the binary dummy variable, the resulting model may differ from a standard FP analysis [148].

The decision to use a model including z as just described, or to work with the standard FP class, is best made on subject-matter grounds rather than by considering the fit of functions with or without z.

3.2.3.4.2 Multivariable Case

In reality it may be unrealistic to consider assessing the effect of an exposure with a spike at zero in univariate models; possible confounders may need to be considered. The confounder model can easily be determined by using a variation - the multivariable FP (MFP) procedure [138, 145]. For variables with a spike at zero within MFP the standard FSP is replaced with the first stage of FSP-spike. The second stage must be done separately for all spike variables after MFP has been completed.

3.2.3.5 Hazard Ratios and Confidence Intervals

To aid interpretation, it can be helpful to present continuous variables as categorised variables. Having modelled the variable using an appropriate continuous model fit, hazard ratios and associated 95% confidence intervals may be calculated *post hoc* from the fitted FP model.

Denote the centre of the baseline, reference, category as x_{ref} where x is the variable of interest and x_1 and x_2 are the 1st and 2nd FP2 transformations of the variable with coefficients β_1 and β_2 respectively. Then,

$$\begin{aligned}\log HR &= \log\left\{\frac{h_i(t)}{h_0(t)}\right\} = \log\big(h_i(t)\big) - \log\big(h_0(t)\big)\\ &= \big(\hat{\beta}_1 x_1 + \hat{\beta}_2 x_2\big) - \big(\hat{\beta}_1 x_1^{ref} + \hat{\beta}_2 x_2^{ref}\big)\\ &= \hat{\beta}_1\big(x_1 - x_1^{ref}\big) + \hat{\beta}_2\big(x_2 - x_2^{ref}\big).\end{aligned}$$

The variance can be calculated as

$$var(\log HR) = var(\hat{\beta}_1)(x_1 - x_1^{ref})^2 + var(\hat{\beta}_2)(x_2 - x_2^{ref})^2 + 2cov(\hat{\beta}_1, \hat{\beta}_2)(x_1 - x_1^{ref})(x_2 - x_2^{ref}),$$

and thus a 95% confidence interval for the *post hoc* hazard ratios is $\exp(\log HR \pm 1.96\sqrt{var(\log HR)})$.

In the case of FP1 transformations, the above equations are much simplified. In addition, the log hazard ratios can obviously be varied to odds ratios.

3.2.4 Comparison of Splines and Fractional Polynomials

MFP and spline model-building procedures give roughly comparable multivariable models. Some of the weaker predictors may be included by one procedure but not by another, and estimated functions often show some differences. In regions with sparse data the differences can be major, but where there is enough data the functions and their interpretation are similar. If interest lies mainly in the simpler task of deriving a good predictor, the differences between the selection procedures become smaller.

Results from MFP are more stable and easier to interpret and transfer to other environments while spline functions with large degrees of freedom produce many artefacts and spline functions cannot easily be written down. However, if sample size is adequate, MFP and splines with small degrees of freedom give similar results. Predictors are similar but there can be substantial differences when interest is in individual variables. Differences may appear with weakly influential predictors and in regions with sparse data. In general, transportability and general usefulness are more important than minor improvements in goodness of fit [145].

3.3 Variable Centring

Centring involves shifting the scale of a variable by subtracting a single value from all of the data points. It is called centring because people often use the mean as the value they subtract, so the new mean is now at zero, but it does not have to be the mean. In fact, there are many situations when a value other than the mean is most meaningful.

There are mixed opinions on the value of centring. Cronbach [152] suggests, 'in regression analysis, always centre' stating reasons such as increased relevance of the estimated regression coefficients and diminished multicollinearity. If centring is done unnecessarily, the cost is minor. Aiken and West [153] and Cohen *et al* [154] have described centring and the consequences of non-centred data while Glantz and Slinker [155] and Kromrey and Foster-Johnson [156] take the stand that centring does not usually change the statistical results, is necessary only in certain circumstances and can thus easily be avoided.

All variables analysed in the thesis are centred. This is particularly of relevance to the work in Chapters 4, 5, 6 and 10.

3.3.1 Multicollinearity

There are two sources of correlation between a predictor and an even power of the predictor, say between X and X^2 [154]. The first is non-essential multicollinearity that exists merely due to the scaling, non-zero mean, of X. The second is essential multicollinearity, correlation that exists because of any non-symmetry in the distribution of the original X variable. Marquardt [157] refers to the problems of multicollinearity produced by non-centred variables as non-essential ill-conditioning, whereas those that exist because of actual relationships between variables in the population are referred to as essential ill-conditioning.

Problems with multicollinearity in least squares regression are well documented, particularly with multiple regression models containing both main effects and interaction terms [154] – in general for two factors A and B, if the effect of variable A on the outcome varies according to the level of variable B, there is said to be an interaction between A and B.

Although the least squares estimates of the regression coefficients remain unbiased, as multicollinearity increases, the determinants of the independent variables covariance and correlation matrices approach zero and the standard errors of the coefficients increase. The resulting ill-conditioning yields coefficients, and an associated variance-covariance matrix, that are unstable. Small changes due to measurement or rounding error may be magnified resulting in large changes in the coefficients and associated variance-covariance matrix. In addition, when multicollinearity is present, slight sampling fluctuations in the estimates of the covariances can result in great variability in the values and signs of least squares estimates of the coefficients. Finally, as a result of the increase in the expected distance between the vector of the least squares coefficient and the vector of true regression coefficients, estimates with excessively large values or unreasonable signs may result when extreme collinearity is present [156].

The problems of collinearity may be overcome in several ways. In some situations the collinearity will have arisen purely as a computational problem and may be solved by alternative definitions of some of the variables. For example, if both X and X^2 are included as explanatory variables and all the values of X are positive, then X and X^2 are likely to be highly correlated. This can be overcome by redefining the quadratic term as $(X - \bar{X})^2$, which will reduce the correlation whilst leading to an equivalent regression [28]. If the multicollinearity is structural, it can often be dealt with by centring the measured

independent variables on their mean values before computing the power, e.g. squared, and interaction, cross-product, terms specified by the regression equation.

3.3.2 Interpretation

Lower order coefficients in higher order regression equations, regression equations containing terms of higher than order unity, only have meaningful interpretation if the variable has a meaningful zero. For example, if some behaviour were predicted from a measure of motivation, X, and a seven point attitude scale, Z, ranging from one to seven, the regression coefficient for Y on X would be the slope of Y on X at the value $Z = 0$, a value not even defined on the scale. Similarly, if strength of athletes were produced from their height and weight, the regression coefficient predicting strength from height would represent the regression of strength on height for athletes weighing 0 pounds.

There is a simple solution to making the value, zero, meaningful on any quantitative scale; centre the linear predictor. Thus the regression of Y on X at $X = 0$ becomes meaningful; it is the linear regression of Y on Z at the mean of the variable X. To gain the benefits of interpretation of lower order terms, it is unnecessary to centre the criterion Y. This can be left in raw score form so that predicted scores will be in the metric of the observed criterion [154].

3.3.3 Discussion

Cohen *et al* [154] strongly recommend the use and reporting of centred polynomial equations. They suggest that doing so has no effect on the estimate of the highest order interaction in the regression equation and also yields two straightforward, meaningful interpretations of each first-order regression coefficient of predictors entered into the regression equation. Firstly, the effects of the individual predictors at the mean of the sample and secondly, the average effects of each individual predictors across the range of the other variables. Aiken and West [153] also recommend centring, this time for

computational reasons because the centred overall regression analysis provides regression coefficients for primary terms that may be informative.

The main disadvantage of centring, however, is that the variables are no longer the natural variables of the problem. If a predictor has a meaningful zero point, then one may wish to keep the predictor in non-centred form. Centring also produces a puzzling effect. When predictors are centred and entered into regression equations containing interactions, the regression coefficients for the first order effects are different numerically from those obtained by performing a regression analysis on the same data in raw score or non-centred form. The regression coefficients do not change when predictors are centred in regression equations containing no interactions [154]. Differences between the non-centred equation, and the centred one get absorbed into the intercept [155], therefore, according to Glantz and Slinker [155], centring will only be beneficial if an intercept term is included in the model.

Studies performed by Kromrey and Foster-Johnson [156] showed that regression equations obtained with centred and raw data were equivalent, results of hypothesis testing with either type of data were exactly the same and neither approach provided a viable vehicle for the interpretation of main effects in regression. They therefore suggest 'one might just as well not bother.' There is, though, very little cost to unnecessary centring but the costs of not centring when it is necessary can be major [158] as using non-centred data in regression analysis, often leads to inconsistent and misleading results. Thus it would always be better to centre in regression analyses.

It can be argued that not centring represents a *de facto* decision that all ordinal variables be centred at zero, that all binary and categorical independent variables be coded somewhat arbitrarily, 1 and 0, and that one category, also often arbitrarily chosen, be used as the reference category. This can lead to serious errors of statistical inference [158].

Kraemer and Blasey [158] recommend the following default approach to protect against most errors in statistical inference. Each binary independent variable should be coded $+1/2$ and $-1/2$ while each ordinal independent variable should be centred with the median response. Categorical independent variables should be 'dummy coded' as usual, but instead of coding each response as 1 and 0, the values $1 - 1/m$ and $-1/m$ should be used where m is the number of categories. As in the usual situation, one categorical 'dummy' is omitted, but with the proposed centring it does not matter which one. For example, for a three level variable, the traditional dummy coding may be as in the left hand side of Table 4 while the recommended coding with centring is as per the right hand side.

Table 4: Alternatives for dummy coding of a three level categorical variable

	Traditional Coding		Coding with Centring	
	x_1	x_2	x_1	x_2
Baseline	0	0	$-1/3$	$-1/3$
Level 1	1	0	$2/3$	$-1/3$
Level 2	0	1	$-1/3$	$2/3$

Requiring that centring always be done merely asks that what is done implicitly anyway be done explicitly and thoughtfully, which promotes better application and understanding of the results of regression analysis.

3.4 Model Building and Development

When developing regression models, data analysts are often faced with many predictor variables that may influence an outcome variable. Often a practitioner's task is to build a reliable regression model that fits the data well, is plausible, simple enough to be interpretable and useful in practice. Model selection reduces to a decision to include or exclude each variable. Many procedures for selecting variables have been proposed although often they do not lead to the same solution when applied to the same problem

and the 'best' way of selecting a multivariable model is still unresolved despite more than half a century of research. It is generally agreed that subject matter knowledge, when available, should guide model building, however, such knowledge is often limited and data-dependent model building is required [159].

If the number of independent variables is large, a parsimonious model is preferred, i.e. a subset of 'important' predictors whose regression coefficients β_j differ from 0. There are two main types of strategies for variable selection to achieve such a model – sequential strategies and all-subsets strategies. Sequential strategies, such as forward selection (FS), stepwise selection (StS) or backwards elimination (BE) procedures, are based on a sequence of tests of whether a given variable should be added to the current model or removed from it, or whether selection should stop. A nominal significance level, α, for each of these tests is chosen in advance and largely determines how many variables will end up in the model.

Mantel [160] and others [161, 162] argue strongly in favour of backward elimination over forward selection, especially when collinearity is present. Forward selection starts by considering k univariate models, all of which underfit the data and have a large residual variance. In contrast, backward elimination starts with the full model. It is likely that some of the variables in the full model have no effect, but it is at least a reasonable starting point. If collinearity between predictors is low, then forward selection and backward elimination frequently select the same model.

Some computer programs use a stepwise strategy and combine it with Akaike's Information Criterion (AIC) [163] or the Bayesian Information Criterion (BIC) [164] as selection criteria. In all-subsets strategies, a model is selected by minimising an information criterion over the 2^k available models with up to k variables. For a given model, M, the information criterion, IC, is of the form $\mathrm{IC} = -2l + a\,\dim(M)$ where l is the

maximised log likelihood of M, $\dim(M)$ is the number of estimated parameters and a is a penalty constant. The penalty constant $a = 2$ gives the AIC, independent of sample size.

AIC is an index used in a number of areas as an aid to choosing between competing models. It is defined as

$$-2L_m + 2m$$

where L_m is the maximized log-likelihood and m is the number of parameters in the model [165]. The index takes into account both the statistical goodness of fit and the number of parameters that have to be estimated to achieve this particular degree of fit, by imposing a penalty for increasing the number of parameters.

BIC has a penalty constant $a = \log n$ and therefore depends on the sample size. For censored survival data, the sample size n may be replaced by the number of events [166]. The penalty constant of $\log n$ for BIC is larger than that of 2 for AIC, generally resulting in models with a smaller number of predictors. This approach however, has major drawbacks, including the possibility of selecting models which omit important predictors [163].

Variable selection via AIC is implemented in Chapters 4, 5 and 6 to develop parsimonious multivariable models.

3.4.1 Treatment-Covariate Interactions Methods

Multivariable models need to be checked for interactions, including the relevant main effects [145]. Testing for binary by binary interactions can be done by standard methods such as the likelihood ratio test which may also be suitable for continuous by binary and continuous by continuous interactions. Alternatives for these latter two combinations are available in the form of Multivariable Fractional Polynomial Interactions (MFPI).

Treatment-covariate interactions are considered in Chapter 4 via the likelihood ratio test (section 3.4.1.1) and multivariable fractional polynomial interactions (section 3.4.1.2).

3.4.1.1 Likelihood Ratio Test

In order to compare alternatives models, fitted to an observed set of survival data, a statistic that measures the extent to which the data are fitted by a particular model is required. As the likelihood function summarises the information that the data contain about the unknown parameters in a given model, a suitable summary statistic is the value of the likelihood function when the parameters are replaced by their maximum likelihood estimates. This is the maximised likelihood under an assumed model and can be computed from Equation 4 by replacing the β's by their maximum likelihood estimates under the model [167]. $x_{(j)}$ is the vector of covariates for the individual who dies at the jth ordered death time, $t_{(j)}$. For a given set of data, the larger the value of the maximum likelihood, the better the agreement between the model and the observed data.

Equation 4

$$L(\beta) = \prod_{j=1}^{r} \frac{\exp(\beta' x_{(j)})}{\sum_{l \in R(t_{(j)})} \exp(\beta' x_l)}$$

It is more convenient to use minus twice the logarithm of the maximised likelihood in comparing alternative models. If the maximised likelihood for a given model is denoted by $\hat{L}$, the summary measure of agreement between the model and the data is $-2\log\hat{L}$. $\hat{L}$ is the product of a series of conditional probabilities, so this statistics will be less than unity [167]. In consequence, $-2\log\hat{L}$ will always be positive, and for a given dataset, the smaller the value of $-2\log\hat{L}$, the better the model.

The statistic $-2\log\hat{L}$ cannot be used on its own as a measure of model adequacy as the value of $\hat{L}$, and hence of $-2\log\hat{L}$, is dependent upon the number of observations in the

dataset. Thus if, after fitting a model to a set of data, additional data become available to which the fit of the model was the same as that to the original data, the value of $-2\log\hat{L}$ for the enlarged dataset would be different from that of the original data [167]. Accordingly the value of $-2\log\hat{L}$ is only useful when making comparisons between models fitted to the same data.

3.4.1.2 Multivariable Fractional Polynomial Interactions

Royston and Sauerbrei [168] proposed an extension of the MFP algorithm to investigate interactions between categorical and continuous covariates. It is an extension of the MFP procedure for simultaneous selection of influential prognostic variables and the selection of the function form for a continuous covariate. The multivariable fractional polynomial interaction, MFPI, algorithm models the prognostic effect of a continuous variable by FP2 transformations within treatment groups, but under the constraint of the same powers. It can be done in a univariate setting or by adjusting the model for other covariates. The type 1 error probability of MFPI is found to be close to its nominal value.

Assume a covariate T has two levels coded 1 and 2. The influence of the continuous covariate Z on the estimated treatment effect is determined by $t(Z) = \hat{f}_2(Z) - \hat{f}_1(Z)$ where $\hat{f}_i(Z), i = 1,2$ are the estimated functions for the prognostic effect of Z in treatment group i. A plot of $t(Z)$ together with a point-wise confidence band is called a treatment-effect plot.

Let z_2 be a continuous covariate and z_1 be, for simplicity, a binary treatment variable, coded $\{0,1\}$ and let x be a vector of potential prognostic factors. With a pre-specified hypothesis, z_2 is the only covariate to be investigated for interaction with z_1. The relationship between the outcome and z_2 is modelled by an FP with the same powers but different regression coefficients at each level of z_1. A standard test of interaction is performed on regression coefficients at the final step. To reduce possible confounding,

adjustment for $\boldsymbol{x}$ may be made. Since z_2 cannot belong to the adjustment model, a different model may be chosen for each z_2 that is investigated. The complete procedure is as follows:

1. Apply the MFP algorithm to $\boldsymbol{x}$, possibly including z_2, with a P-value threshold of α^* for selecting variables and FP transformations. Let $\boldsymbol{x}^*$ be the resulting covariate vector, called the adjustment model. $\boldsymbol{x}^*$ may include transformed variables in $\boldsymbol{x}$ selected by the MFP algorithm. If all variables in $\boldsymbol{x}$ are uninfluential, then $\boldsymbol{x}^*$ may even by empty.
2. Find by maximum likelihood the best-fitting FP2 powers, $\boldsymbol{p} = (p_1, p_2)$, for which z_2, with $p_1, p_2 \in S$, always adjusting for z_1 and $\boldsymbol{x}^*$. Denote the FP2 transformations $z_2^{\boldsymbol{p}} = \left(z_2^{p_1}, z_2^{p_2}\right)$.
3. For groups $j = 0{,}1$ and powers p_i for $i = 1{,}2$, define new predictors $z_{ji} = z_2^{p_i}$ if $z_1 = j$, and $z_{ji} = 0$ otherwise.
4. The test of $z_1 \times z_2$ interaction is a likelihood ratio test between the nested models $z_1, z_{01}, z_{02}, z_{11}, z_{12}, \boldsymbol{x}^*$ and $z_1, z_2^{p_1}, z_2^{p_2}, \boldsymbol{x}^*$. The difference in deviance is compared with χ^2 on two degrees of freedom.
5. If an interaction is not found, then z_2 is regarded as a potential prognostic factor only. To investigate if an FP2 function is still needed for z_2, the final model is chosen by repeating step 1, but including z_2 as a potential prognostic factor.

The reason why FP2 functions are fitted to z_2, rather than simpler functions, is to find the best-fitting specification from a flexible class. In terms of bias or variance trade-off, increased variance incurred through the use of FP2 powers for z_2 may be tolerated at the price of low bias. To avoid excessive overfitting, leading to serious artefacts in the fitted functions, estimation of different powers in each treatment group is not considered. An FP2 function with the same powers in each treatment group is already a flexible

specification. When z_2 is binary or categorical, the approach reduces to the usual approach of estimating and testing multiplicative interaction term(s), adjusting for $\boldsymbol{x}^*$.

Further refinements of the procedure are possible. A less flexible option would involve use of linear functions. Alternatively, FP1 functions could be used. For greater flexibility it may be possible to allow the powers to be different in the treatment groups. However, except in very large samples, the advantage of increased flexibility would probably be outweighed by the increased instability [145].

3.5 Testing the Proportional Hazards Assumption

As mentioned previously, the Cox model relies on the proportional hazards assumption. Therefore the proportional hazards assumption will be investigated using log cumulative hazard plots and Schoenfeld residual plots. The assumption will also be tested to see whether incorporating a time-dependent covariate effect indicates significant violations.

If the proportional hazards assumption is found not to be valid, the results will be compared to those from the accelerated failure time model. This will determine if the violation of the assumption affects the conclusion.

The proportional hazards assumption is tested in Chapters 4, 5 and 6 via log cumulative hazard plots (section 3.5.1), the inclusion of time-dependent covariate effects (section 3.5.2), and examination of Schoenfeld residuals (section 3.5.3.1). Other types of residuals are included in this chapter for completeness only.

3.5.1 Log Cumulative Hazard Plots

A graphical approach for checking the proportional hazards assumption involves comparing log–log survival curves over different combinations of categories of variables being investigated. Parallel curves indicate that the proportional hazards assumption is satisfied.

A log–log survival curve is a transformation of an estimated survival curve that results from taking the natural log of an estimated survival probability twice. Mathematically, a log–log curve is written as $-ln(-ln\hat{S})$.

Recall that the formula for the survival curve of the Cox model is

$$S(t,X) = S_0(t)^{\exp\left(\sum_{j=1}^{p}\beta_j X_j\right)}.$$

The first time logs are taken the resulting expression is

$$\log(S(t,X)) = \exp\left(\sum_{j=1}^{p}\beta_j X_j\right) \times \log(S_0(t)).$$

Since $S(t,X)$ denotes a survival probability, its value for any t and any specification of the vector X will be some number between 0 and 1. It follows that the natural log of any number between 0 and 1 is a negative number, so that the log of $S(t,X)$ as well as the log of $S_0(t)$ are both negative numbers. Thus, when taking the second log, the log of $-\log(S(t,X))$ must be calculated.

After some algebra, this expression can be rewritten as the sum of two terms, one of which is the linear sum of the $\beta_j X_j$ and the other is the log of the negative log of the baseline survival function. Suppose two different specifications of the $\boldsymbol{X}$ vector are considered, corresponding to two different individuals, $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$. Then the corresponding log–log curves for these individuals are given in Equation 5 [169].

Equation 5

$$-\ln[-\ln S(t,\boldsymbol{X}_1)] = -\sum_{j=1}^{p}\beta_i X_{1j} - \ln[-\ln S_0(t)]$$

$$-\ln[-\ln S(t,\boldsymbol{X}_2)] = -\sum_{j=1}^{p}\beta_i X_{2j} - \ln[-\ln S_0(t)]$$

Subtracting the second log–log curve from the first yields the expression shown in Equation 6. This expression is a linear sum of the differences in corresponding predictor values for the two individuals. Note that the baseline survival function has dropped out, so that the difference in log–log curves involves an expression that does not involve time *t*.

Equation 6

$$-\ln[-\ln S(t,\boldsymbol{X}_1)] - \ln[-\ln S(t,\boldsymbol{X}_2)] = \sum_{j=1}^{p} \beta_j (X_{2i} - X_{1i})$$

Alternatively, using algebra, Equation 6 can be re-written by expressing the log–log survival curve for individual $\boldsymbol{X}_1$ as the log–log curve for individual $\boldsymbol{X}_2$ plus a linear sum term that is independent of t, as shown in Equation 7. This says that if a Cox model is used and the estimated log–log survival curves for individuals are plotted on the same graph, the two plots will be approximately parallel. The distance between the two curves is the linear expression involving the differences in predictor values, which does not involve time. In general, if the vertical distance between two curves is constant, then the curves are considered parallel.

Equation 7

$$-\ln[-\ln S(t,\boldsymbol{X}_1)] = -\ln[-\ln S(t,\boldsymbol{X}_2)] + \sum_{j=1}^{p} \beta_j \left(X_{2j} - X_{1j}\right)$$

The parallelism of log–log survival plots for the Cox proportional hazards model provides a graphical approach for assessing the proportional hazards assumption. That is, if a Cox model is appropriate for a given set of predictors, one should expect that empirical plots of log–log survival curves for different individuals will be approximately parallel.

3.5.2 Time-Dependent Covariate Effects

When time-dependent variables are used to assess the proportional hazards assumption for a time-independent variable, the Cox model is extended to contain interaction terms involving the time independent variable being assessed and some function of time. When assessing predictors one-at-a-time, the extended Cox model takes the general form shown in Equation 8 for the predictor X.

Equation 8

$$H(t, \boldsymbol{X}) = h_0(t) \exp[\beta X + \delta X g(t)]$$

One choice for the function $g(t)$ is simply $g(t)$ equal to t, so that the product term takes the form Xt. Other choices for $g(t)$ are also possible, for example, log t. Using the above one-at-a-time model, the proportional hazards assumption is assessed by testing for the significance of the product term. The null hypothesis is therefore $\delta = 0$. The test can be carried out using either a Wald statistic or a likelihood ratio statistic. In either case, the test statistic has a chi-square distribution with one degree of freedom under the null hypothesis.

3.5.3 Residuals for Cox regression model

A goodness of fit testing approach is appealing because it provides a test statistic and p-value for assessing the proportional hazards assumption for a given predictor of interest. Thus, the researcher can make a more objective decision using a statistical test than is typically possible when using a graphical approach. Many goodness of fit testing procedures are based on residuals which are values that can be calculated for each individual in the study, and have known behaviour when the fitted model is satisfactory. Schoenfeld residuals, Martingale residuals and deviance residuals are the most popular.

3.5.3.1 Schoenfeld Residuals

According to Collett [167], the Schoenfeld residuals are particularly useful in evaluating the assumption of the proportional hazards after fitting a Cox regression model. Grambsch and Therneau [170] have shown that the expected value of the i^{th} scaled Schoenfeld residuals, for the j^{th} explanatory variable, X_j, in the model, $\boldsymbol{r}^*_{p_{ji}}$, is given by $E\left(\boldsymbol{r}^*_{p_{ji}}\right) \approx \beta_j(t_i) - \hat{\beta}_j$ where $\beta_j(t)$ is taken to be a time-varying coefficient of X_j, $\beta_j(t_i)$ is the value of the coefficient at the i^{th} death time, t_i and $\hat{\beta}_j$ is the estimated value of β_j in the fitted Cox regression model. Consequently, a plot of the values of $\boldsymbol{r}^*_{p_{ji}} + \hat{\beta}_j$ against the death times should give information about the form of the time-dependent coefficient of X_j, $\beta_j(t)$. In particular, a horizontal line will suggest that the coefficient of X_j is constant and the proportional hazards assumption is satisfied. A smoothed curve can be superimposed onto this plot to aid interpretation. This plot can also be supplemented by fitting a straight line, and formally testing if the slope of this line is zero. However, this procedure has its limitations, since a slope that is not significantly different from zero may be found when there is, in fact, a non-linear relationship between the coefficient and time.

3.5.3.2 Martingale Residuals

Suppose that the i^{th} survival time is a censored observation, t_i^*, and let t_i be the actual, but unknown, survival time, so that $t_i > t_i^*$. The Cox-Snell [171] residual for this individual, evaluated at the censored survival time, is then given by $r_{Ci} = \hat{H}_i(t_i^*) = -\log \hat{S}_i(t_i^*)$, where $\hat{H}_i(t_i^*)$ and $\hat{S}_i(t_i^*)$ are the estimated cumulative hazard and survivor functions respectively for the i^{th} individual at the censored survival time. Also, suppose that δ_i is a censoring indicator which takes the value zero if the observed survival time of the i^{th} individual is censored and unity if it is uncensored.

Residuals, $r_{Mi} = \delta_i - r_{Ci}$, are known as Martingale residuals since they can be derived using Martingale methods. They take values between $-\infty$ and unity with the residuals for

censored observations being negative. In large samples, the Martingale residuals are uncorrelated with one another and have an expected value of zero. They are not symmetrically distributed about zero which therefore makes plots based on the residuals difficult to interpret.

3.5.3.3 Deviance Residuals

The deviance residuals, which were introduced by Therneau *et al* [172], are more symmetrically distributed around zero than Martingale residuals. They are defined by [167]

$$r_{Di} = \text{sgn}(r_{Mi})[-2\{r_{Mi} + \delta_i \log(\delta_i - r_{Mi})\}]^{1/2}$$

where r_{Mi} is the Martingale residual for the i^{th} individual, δ_i is the event indicator and the function $\text{sgn}(\cdot)$ is the sign function - the function that takes the value $+1$ if its argument is positive and -1 if negative. Thus $\text{sgn}(r_{Mi})$ ensures that the deviance residuals have the same sign as the Martingale residuals.

The deviance is a statistic that is used to summarise the extent to which the fit of a model of current interest deviates from that of a model which is a perfect fit to the data. The statistic is given by $D = -2\{\log \hat{L}_c - \log \hat{L}_f\}$ where $\hat{L}_c$ is the maximised partial likelihood under the current model and $\hat{L}_f$ is the maximised partial likelihood under the full model. The deviance residuals are then such that $D = \sum r_{Di}^2$, so that observations that correspond to relatively large deviance residuals are those that are not well fitted by the model.

An alternative way of viewing the deviance residuals is to consider them as Martingale residuals which have been transformed to produce values that are symmetric about zero when the fitted model is appropriate. Although these residuals can be expected to be symmetrically distributed about zero when an appropriate model has been fitted, they do not necessarily sum to zero.

3.5.4 Comparison of Deviances

Two disadvantages of the Martingale residuals are that they depend heavily on the observed survival time and require an estimate of the cumulative hazard function. Both of these disadvantages are overcome in a residual proposed by Schoenfeld [173]. This residual differs from the Martingale residuals in one other important respect - there is not a single value of the residual for each individual, but a set of values, one for each explanatory variable included in the fitted Cox regression model. While Martingale residuals are useful for assessing the correct functional form for the effect of a covariate, almost all of the available literature on testing the proportional hazards assumption uses Schoenfeld residuals.

3.6 Competing Risks Methods

When only one of several different types of event can occur, the probability of these events is referred to as competing risks. In the epilepsy setting, treatment failure may arise as a result of either adverse events or inadequate seizure control. Cumulative incidence analysis will be used to show the contribution to the treatment failure outcome across drugs while covariates will be tested by Gray's method [174].

Many sources [169, 175] suggest that modelling competing risks survival data can be carried out using a Cox model. The drawback of performing a survival analysis for each event type separately, where the other, competing, event types are treated as censored categories is the assumption that the competing risks are independent. Cumulative incidence analysis makes no such assumption and allows the assessment of cause-specific withdrawal in the presence of other competing risks.

Competing risks analyses are undertaken in Chapter 4 via cumulative incidence analysis (section 3.6.1) with examination of covariates by Fine and Gray's method (section 3.6.2).

3.6.1 Cumulative Incidence Analysis

The cumulative incidence curve estimates the marginal probability of an event and is therefore not estimated using a product-limit formulation. The independence of competing risks is still required, when a proportional hazard model is used to obtain hazard ratio estimates for individual competing risks, as an intermediate step in the computation of a cumulative incidence curve. This assumption will be satisfied whenever competing risks are mutually exclusive and events are non-recurrent; that is, one and only one event can occur at any one time and only once over time. Irrespectively, the cumulative incidence curve has a meaningful interpretation in terms of treatment utility regardless of whether competing risks are independent [30, 169].

To construct a cumulative incidence curve, first estimate the hazard at ordered time points t_j when the event of interest occurs. This hazard estimate is simply the number of events that occur at t_j divided by the number at risk at t_j. This can be written as

$$\hat{h}_{\boldsymbol{c}}(t_j) = {}^{m_{\boldsymbol{c}j}}/_{n_j}$$

where the $m_{\boldsymbol{c}j}$ denotes the number of events for risk $\boldsymbol{c}$ at time t_j and n_j is the number of subjects at that time. Thus, at any particular time, ${}^{m_{\boldsymbol{c}j}}/_{n_j}$ is the estimated proportion of subjects failing from risk $\boldsymbol{c}$.

To be able to fail at time t_j, the subject must have survived the previous time when a failure occurred. The probability of surviving the previous time t_{j-1} is denoted $S(t_{j-1})$, where $S(t)$ denotes the overall survival curve rather than the cause-specific survival curve $S_{\boldsymbol{c}}(t)$. Overall survival must be considered here because the subject must have survived all other competing events. The incidence of failing from event-type $\boldsymbol{c}$ at time t_j is then simply the probability of surviving the previous time period multiplied by $\hat{h}_{\boldsymbol{c}}(t_j)$. The cumulative

incidence at time t_j is then the cumulative sum up to time t_j of these incidence values over all event-type $\boldsymbol{c}$ failure times.

3.6.2 Fine and Gray

Fine and Gray [176] provide methodology for modelling the cumulative incidence curve with covariates using a proportional hazards assumption. They refer to the cumulative incidence curves as sub-distribution functions. The cumulative incidence curve models developed by Fine and Gray are analogous to the Cox model but, for any failure type, they model a cumulative incidence curve.

Fine and Gray's model is based on $\gamma(t, \boldsymbol{x}) = \gamma_0(t)\exp(\boldsymbol{\beta x})$ where γ is the hazard of the subdistribution, γ_0 is the baseline hazard of the subdistribution, $\boldsymbol{x}$ is the vector of the covariates and $\boldsymbol{\beta}$ is the vector of the coefficients. The form of the partial likelihood is similar to that used in the Cox model. Written for just one covariate, x, the partial likelihood is given by Equation 9 [177]

Equation 9

$$L(\beta) = \prod_{j=1}^{r} \frac{\exp(\beta x_i)}{\sum_{i \in R_j} w_{ij} \exp(\beta x_i)}$$

where the product is taken over all r time points $(t_1 < t_2 < \cdots < t_r)$ where an event of interest was observed.

There are two main differences between the partial likelihood for Fine and Gray's model and that of the Cox model; the risk set R_j is defined in a different way and weights w_{ij} are added. The risk set is formed of those who did not experience an event by time t and of those who experienced a competing risk event by time t,

$$R_j(t) = \{i;\ T_i \leq t \text{ and the subject experienced a competing risks event})\}.$$

Thus those who experienced other types of events remain in the risk set all the time. The weights are defined as

$$w_{ij} = \frac{\hat{G}(t_j)}{\hat{G}(\min(t_i, t_j))}$$

where $\hat{G}$ is the Kaplan-Meier estimate of the survivor function of the censored distribution. The censoring distribution is defined by (T_i, C_i) where T_i is the time to the first event and C_i is 1 if no event is observed and 0 if any kind of event was observed. At each time point where an event of interest occurred, indexed by j, the set of individuals at risk, indexed by i, includes those who did not have any type of event by time t_j as well as those who had a competing risks event before t_j. The weight is 1 for the former and less than or equal to 1 for the latter. In this way those individuals experiencing a competing risk event do not participate fully in the partial likelihood: the further the time point, t_j, is from the time when the competing risk event occurred, the smaller the weight.

To investigate the proportionality assumption for the competing risks regression, $\log(-\log(1 - \text{CIF}))$ can be plotted against $\log(\text{time})$. Additionally, a time-dependent covariate may be included in the model, which tests whether the hazard ratio changes with time [177].

3.7 Discussion

Time to event outcomes will be summarised by Kaplan-Meier curves for each level of a categorical prognostic factor and compared overall using log-rank tests. A Cox model for the prognostic factor effect will then be fitted using indicator variables for the levels of the categorical variable. Univariate hazard ratios together with associated 95% confidence intervals for levels of each categorical prognostic factor will also be calculated. Hazard

ratios together with their associated 95% confidence intervals will be calculated for any continuous variables.

The hazard of each event in the multivariate setting will be assessed using a Cox proportional hazards regression model. Best-fitting, parsimonious, multivariable models will be produced with backwards elimination by Akaike's Information Criterion (AIC) [178]– the smaller the value of this statistic, the better the model [167]. Schoenfeld residual plots will be used to test the proportional hazards assumption.

It is necessary to consider the effect of treatment in each case; to do this treatment will be forced to stay in the model. I.e. it will be included as a variable and, irrespective of the backward elimination process, it will remain in the model.

Potential prognostic factors may be investigated via implementation of methods for curve fitting, variable centring, model selection, assumption testing, treatment-covariate interactions and competing risks. These methods will be employed in Chapter 4 to develop prognostic models for time to treatment failure and time to 12 months of remission for patients with epilepsy.

Chapter 4: Identifying Prognostic Factors for Epilepsy

4.1 Introduction

This chapter begins with an explanation of how it is possible to use The SANAD Trial data to investigate more than overall estimates of treatment effect (section 4.1). In section 4.2 The SANAD Trial is described in detail (section 4.2.1) together with the prognostic modelling methods used (section 4.2.2). A prognostic model for time to treatment failure can be seen in section 4.3.1 while a model for time to 12 month remission can be seen in section 4.3.3.

The SANAD dataset will be further examined in Chapter 7 where the internal validation of the models presented in this chapter will be considered. Additionally, a simulation study into methods of external validation will be considered in Chapter 8 and the SANAD dataset will be used to inform the simulations. Finally, the SANAD dataset will also be considered in Chapter 10 where a more complex method of modelling time to 12 month remission will be considered.

The work presented in this chapter has been published in Lancet Neurology and I am first author. This is important research and has been discussed at various epilepsy meetings and conferences since its publication.

The epilepsies represent a heterogeneous group of disorders and we recognise a number of differing seizure types [76] and epilepsy syndromes [179] with differing aetiologies, severity and age of onset. While results from the Standard and New Antiepileptic Drug trial (SANAD) published so far [180, 181] provide overall estimates of treatment effect, SANAD recruited a large heterogeneous group of patients providing an opportunity to use prognostic modelling to investigate which clinical factors might influence outcome. The

results will inform patient counselling about likely outcomes, treatment choices and the design of future monotherapy epilepsy trials.

4.2 Methods

4.2.1 SANAD

SANAD [180] had two arms, A and B. In the case of arm A, the recruited patients were those for whom clinicians considered carbamazepine the first line standard treatment, 89% of whom were classified with a focal epilepsy. Patients were randomly allocated to treatment with carbamazepine, gabapentin, lamotrigine, oxcarbazepine or topiramate. Overall results indicated that lamotrigine was a potential first line treatment as it was significantly superior to carbamazepine for the outcome time to treatment failure (lamotrigine vs. carbamazepine: HR 0.78 95% CI (0.63 to 0.97)), but similar to carbamazepine for time to 12 month remission. Gabapentin and topiramate were identified as poor first line treatments, gabapentin due to lack of efficacy and topiramate due to both lack of efficacy and poor tolerability.

The methods for the SANAD study have been published elsewhere [180]. In summary, patients were eligible for inclusion in the study if, in the previous year, they had a history of at least two clinically definite unprovoked epileptic seizures, they were at least five years old, and if the recruiting clinician deemed carbamazepine as opposed to valproate, to be the optimal standard treatment option in the case of arm A and valproate rather than carbamazepine to be the optimal in the case of arm B. Patients generally had newly diagnosed and untreated epilepsy, but could also be recruited if they were taking a monotherapy drug that was not being investigated in SANAD (e.g. phenytoin) with a change in drug indicated, or if they had previously been treated with an antiepileptic drug,

had had a seizure remission, but had a recurrence of seizures following antiepileptic drug withdrawal.

Exclusion criteria included treatment contraindication, a history of progressive neurological disease or a history of only acute symptomatic seizures such as febrile seizures. Patients were recruited to the trial between December 1st 1999 and August 31st 2004 and were followed-up, if possible, until between May 1st 2005 and August 31st 2005 although some follow-up data were obtained up to January 13th 2006. Following enrolment by the recruiting physician and obtainment of informed consent, patients were allocated randomly to treatment. Between December 1st 1999 and June 1st 2001 patients were allocated in a ratio of 1:1:1:1 to carbamazepine, gabapentin, lamotrigine, and topiramate in arm A, and 1:1:1 to valproate, lamotrigine and topiramate in arm B.

From 1st June 2001 to 31st August 2004 an oxcarbazepine group was added to arm A of the trial and patients were randomly allocated in a ratio of 1:1:1:1:1 to carbamazepine, gabapentin, lamotrigine, oxcarbazepine or topiramate. Randomisation was by a central telephone service via the minimisation method which was balanced across centre, gender and treatment history. While the drug allocation was randomised, the dose and preparation was based on the clinicians' usual practice aided by guidelines in the trial protocol.

Baseline clinical information and demographic data such as neurological history, EEG, brain imaging results, seizure classification and epilepsy type were obtained for all randomised patients. Where there was uncertainty between focal onset and generalised onset seizures, patients were recorded as having unclassified convulsive, or other unclassified, seizures. Follow-up occurred at three, six and 12 months and successive yearly intervals from randomisation with additional visits as clinically indicated. At each visit the

occurrence of seizures, adverse events, hospital admissions and antiepileptic drug treatment were recorded.

There were two primary outcomes - time to treatment failure from randomisation and time to 12 month remission from randomisation. Treatment failure can be split into two main categories; inadequate seizure control for which the randomised drug might be withdrawn or a second treatment added, or due to unacceptable adverse events. Patients were categorised into these two main failure reasons as in the original SANAD analyses [180, 181]. Patients categorised as failing due to inadequate seizure control included those citing inadequate seizure control as the only reason for treatment failure, those with both inadequate seizure control and unacceptable side effects, and those who died where the cause of death was related to epilepsy. Patients were categorised as failing due to adverse events if they recorded unacceptable adverse events, withdrew treatment following a period of remission of less than 12 months where adverse events likely influenced the decision to withdraw treatment, and those withdrawing treatment due to pregnancy or planned pregnancy due to concerns about teratogenic effects. Time to 12 month remission was defined as time from randomisation to 365 days of seizure freedom.

4.2.2 Prognostic Modelling

Analyses were undertaken using *R* version 2.9.2 [182] and SAS version 9 [183].

The aim of the prognostic modelling was to identify two sets of factors – one set that predict time to 12 month remission, and one set that predict time to treatment failure. Clinical consensus and knowledge from previous prognostic studies in epilepsy [119, 120] led to the following list of potential prognostic factors: gender, febrile seizure history, first degree relative with epilepsy, CT/MR scan result, treatment history, age, time from first seizure to randomisation, neurological insult (e.g. hemiparesis), total number of seizures ever before randomisation, EEG result, seizure type and epilepsy type. For the CT/MR

results, scans were classified as normal, abnormal and not done. Patients were classified as having neurological insult if they had learning disabilities or neurological deficit. EEGs were classified as normal, not done, non-specific abnormality and epileptiform abnormality (focal or generalised spikes or spike and slow wave activity).

Seizure types were classified according to the International League Against Epilepsy seizure classification [76]. Epilepsy type was first classified as focal, generalised or unclassified. In arm A, patients with a focal epilepsy were further classified as temporal lobe, frontal lobe, parietal lobe, occipital lobe, benign focal epilepsy, or focal epilepsy not localised. For the regression modelling, due to small numbers of participants, occipital lobe, parietal lobe and benign focal epilepsy were combined in a group referred to as other.

Modelling methods used are as outlined in detail in Chapter 3 but in summary variables associated with a higher chance of treatment failure and a higher chance of achieving 12 month remission were determined after adjusting for multiple variables using Cox proportional hazards modelling methods. As oxcarbazepine was only included in the randomisation after June 1st 2001 all analyses were stratified by randomisation period to account for the reduced patient numbers for this drug.

Variables were centred to diminish multicollinearity [158] and tested via the likelihood ratio test [167]. Best-fitting, parsimonious, multivariable models were produced with backwards elimination with AIC – selection starts with the full model and eliminates predictors one at a time, at each step considering whether the criterion will be improved, smaller AIC, by adding back in a variable removed at a previous step [178]. From the multivariable model the probability of the event was calculated for combinations of risk factors. The proportional hazards assumption was investigated using Schoenfeld residual plots [173] and incorporation of time-dependent covariate effects while the internal

validity was assessed by the concordance statistic which evaluates the discriminatory power and the predictive accuracy of nonlinear statistical models [184].

Assessment of the different reasons for treatment withdrawal requires a competing risks analysis which considers the probability of one of several different events occurring. Cumulative incidence analyses were undertaken to assess the probability of one of the two treatment failure events occurring (inadequate seizure control and unacceptable adverse events), with covariates tested by Gray's method [174].

Continuous variables were investigated using log, fractional polynomial and fractional polynomial with spike transformations [144, 145, 185, 186]. The results for the continuous variables are presented as post-hoc defined categorical variables with categories chosen according to knot positions for a spline model fit to the data [143] as described in Chapter 3. In the final model the variables were retained as continuous and presented as categorical variables on for ease of interpretation by non-statisticians.

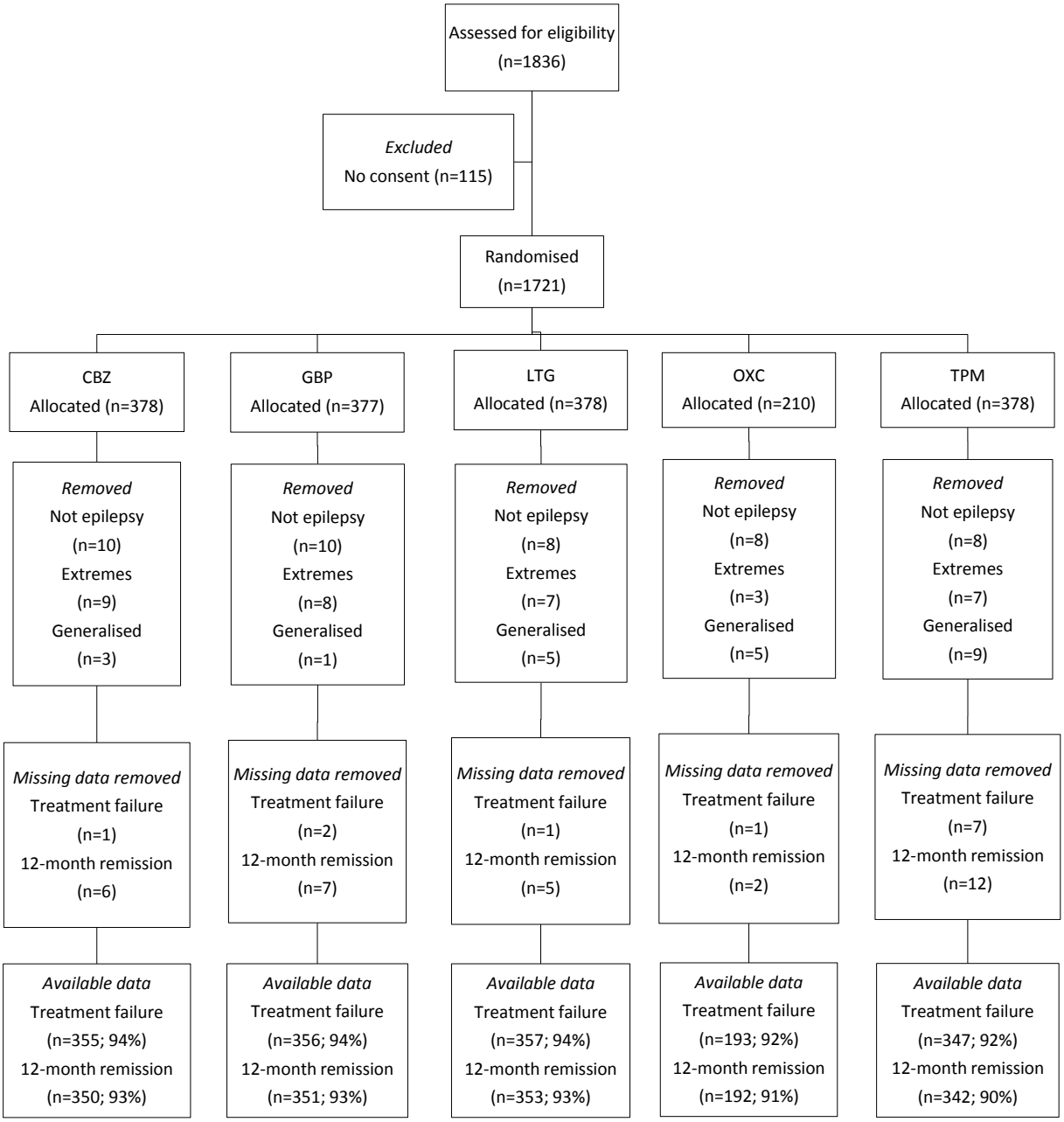
The variable for time from first seizure to randomisation includes extreme values. A sensitivity analysis was performed which examined time to treatment failure and time to 12 month remission in two scenarios. First, time from first seizure to randomisation was included unchanged, and second the variable was reduced by excluding patients with extreme values (1% at either end of the range). Patients with a seizure 18 days or less before randomisation and with a seizure over 41 years before randomisation were consequently examined to determine their impact on the analysis.

An alternative scenario would involve regarding the extreme values as missing data and re-estimating them via imputation. However, recommended practice is to remove the 1% extreme values as this smoothly draws in the extreme tails [187]. Following this sensitivity analysis the 1% extreme values were removed therefore people with a time from first

seizure to randomisation in the first or last one percent of the variable were removed from the dataset – this applied to 34 patients (labelled 'extremes' in Figure 4).

4.3 Results

Figure 4 illustrates patient disposition for patients in arm A of SANAD. 1721 patients were recruited into the study, 44 of whom were excluded from the analyses as they were subsequently found not to have epilepsy. Outcome data were available for 1664 patients who contributed to the analysis of time to treatment failure and 1644 who contributed to the analysis of time to 12 month remission. Table 5 summarises the demographic details of the included patients; 55% of the patients were male, the median age was about 37 years, 5% had a history of febrile seizures, 10% had a first degree relative with epilepsy and 82% were treatment naive. The median time from first seizure to randomisation was 1.4 years, 11% had history of neurological insult, the median number of seizures ever before randomisation was 12 and approximately 46% had abnormal EEG results. An abnormal CT/MR result was seen in 26% while 32% had simple or complex partial seizures only, 58% had secondary generalised seizures and 10% uncertain seizures. 90% of patients were classified with focal epilepsy and 10% unclassified epilepsy. Data were similar for time to 12 month remission.



'Extremes removed' relates to the removal of patients with a time from first seizure to randomisation within the first or last one percent of the variable while 'Generalised removed' relates to removal of patients with generalised epilepsy type. Percentages in final row are percentage of allocated patients with available data – 93% of all randomised patients were available for analysis of time to treatment failure; 92% of all randomised patients were available for analysis of time to 12 month remission

Figure 4: The SANAD Trial – arm A: Patient Disposition

Table 5: The SANAD trial – arm A: Baseline patient demographics

	Treatment Failure Outcome					
	CBZ (n=355)	GBP (n=356)	LTG (n=357)	OXC (n=193)	TPM (n=347)	Total (n=1608)
Gender: Male	196 (55)	197 (55)	199 (56)	102 (53)	193 (56)	887 (55)
History of febrile seizures	24 (7)	15 (4)	22 (6)	7 (4)	13 (4)	81 (5)
First degree relative with epilepsy	36 (10)	41 (12)	33 (9)	21 (11)	34 (10)	165 (10)
Treatment history						
Treatment naive	290 (82)	289 (81)	292 (82)	167 (87)	282 (81)	1320 (82)
Taking non-SANAD AED	57 (16)	56 (16)	58 (16)	23 (12)	57 (16)	251 (16)
Seizures after remission	8 (2)	11 (3)	7 (2)	2 (1)	9 (3)	37 (2)
Age at randomisation in years, Median (IQR)	38 (25 to 53)	36 (24 to 50)	34 (22 to 51)	40 (27 to 56)	37 (25 to 52)	38 (24 to 52)
Years from 1st seizure to randomisation, Median (IQR)	1.4 (0.5 to 4.9)	1.3 (0.6 to 6.0)	1.4 (0.5 to 4.7)	1.3 (0.5 to 4.0)	1.4 (0.5 to 5.3)	1.4 (0.5 to 5.1)
Neurological insult	42 (12)	41 (12)	41 (11)	16 (8)	43 (12)	113 (7)
Number seizures ever before randomisation, Median (IQR)	12 (4 to 75)	14 (4 to 70)	12 (4 to 60)	11 (4 to 53)	12 (4 to 100)	12 (4 to 70)
Seizure type						
Simple or complex partial only	122 (34)	112 (31)	105 (29)	55 (28)	116 (33)	510 (32)
2° Generalised tonic-clonic	202 (57)	209 (59)	216 (61)	117 (61)	193 (56)	937 (58)
Uncertain	31 (9)	35 (10)	36 (10)	21 (11)	38 (11)	161 (10)
Epilepsy type						
Focal	325 (92)	323 (91)	321 (90)	173 (90)	311 (90)	1453 (90)
Temporal	*150 (46)*	*120 (37)*	*107 (33)*	*56 (32)*	*114 (37)*	*547 (38)*
Frontal	*20 (6)*	*17 (5)*	*34 (11)*	*5 (3)*	*31 (10)*	*107 (7)*
Other	*20 (6)*	*28 (9)*	*17 (5)*	*13(8)*	*21 (6)*	*99 (7)*
Not specified	*135 (42)*	*158 (49)*	*163 (51)*	*99 (57)*	*145 (47)*	*700 (48)*
Unclassified	30 (8)	33 (9)	36 (10)	20 (10)	36 (10)	155 (10)
EEG results						
Normal	150 (42)	177 (50)	169 (47)	84 (44)	144 (41)	724 (45)
Non-specific abnormality	59 (17)	48 (13)	56 (16)	32 (17)	59 (17)	254 (16)
Epileptiform abnormality	113 (32)	100 (28)	100 (28)	54 (28)	104 (30)	471 (29)
Not done	33 (9)	31 (9)	32 (9)	23 (12)	40 (12)	159 (10)
CT/MR results						
Normal	205 (58)	225 (63)	207 (58)	109 (56)	188 (54)	934 (58)
Abnormal	99 (28)	82 (23)	83 (23)	53 (27)	103 (30)	420 (26)
Not done	51 (14)	49 (14)	67 (19)	31 (16)	56 (16)	254 (16)

Values in the table are actual number with percentages in brackets unless otherwise stated.

4.3.1 Time to overall Treatment Failure

Figure 5 shows the unadjusted time to treatment failure for the 1608 patients that contributed to the analysis. Results for both the univariate and multivariable modelling are presented in Table 6. In the univariate model the following factors were significantly associated (at the 5% level) with higher risk of treatment failure: gender, having an epileptiform EEG abnormality, being randomised to carbamazepine rather than to lamotrigine, treatment history, age, time from first seizure to randomisation, total number of seizures ever before randomisation, epilepsy type and seizure type.

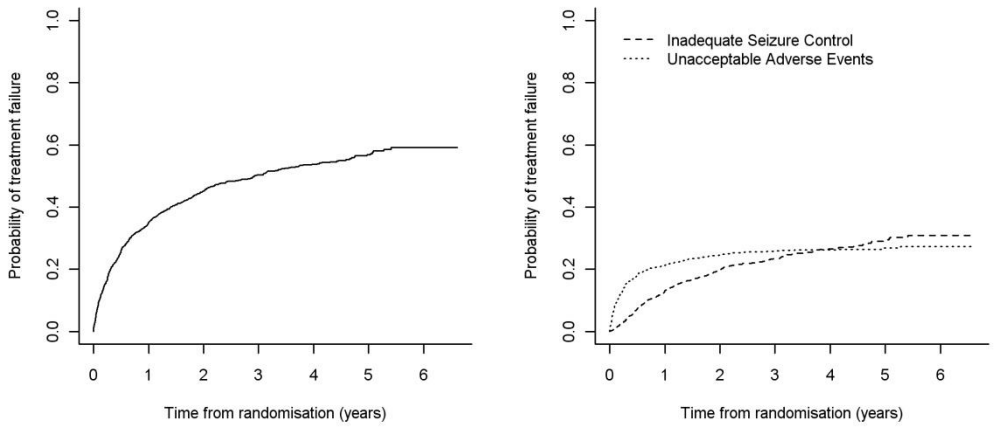


Figure 5: The SANAD Trial – arm A: Unadjusted Kaplan-Meier curve for overall time to treatment failure (left) and cumulative incidence curves for time to treatment failure according to competing risks (right)

For the treatment failure outcome, a HR > 1 indicates that treatment failure is more likely. The resulting parsimonious multivariable model included variables for gender, treatment history, age, total number of seizures ever before randomisation, EEG result, seizure type, focal epilepsy site of onset and treatment which was forced into the model.

Treatment failure rates reduced as age increased and patients aged 10 years or under were significantly more likely to experience treatment failure. Treatment failure rates increased as total number of seizures before randomisation increased and patients with less than or equal to two seizures before randomisation were significantly less likely to experience treatment failure. Treatment failure was significantly (5% level) more likely in women than men (male vs. female: HR 0.86 95% CI (0.75 to 0.99)), patients taking a non-SANAD antiepileptic drug immediately prior to randomisation rather than being treatment naive (non-SANAD vs. naive: HR 1.27 95% CI (1.05 to 1.53)), patients with an epileptiform EEG abnormality rather than a normal EEG (abnormal vs. normal: 1.26 95% CI (1.07 to 1.50)), patients with simple or complex partial seizures without secondary generalisation rather

than secondary generalised tonic-clonic seizures (secondary generalised vs. simple or complex: HR 0.78 95% CI (0.66 to 0.91)), epilepsy that is not localised rather than temporal lobe epilepsy (not localised vs. temporal: HR 1.25 95% CI (1.06 to 1.47)) and patients taking carbamazepine rather than lamotrigine (lamotrigine vs. carbamazepine: HR 0.76 95% CI (0.61 to 0.95)).

In order to illustrate the range of treatment failure rates predicted by the multivariable model, Figure 6 gives estimates of the proportion of patients having treatment failure events at 1 and 3 years after randomisation for patients treated with either carbamazepine or lamotrigine together with the associated 95% confidence intervals. Patients were assumed to be treatment naive, have normal EEG results, simple or complex partial seizures and temporal lobe epilepsy.

In general, based on the forest plots in Figure 6 and consequently a subset of the SANAD data, overall treatment failure rates are lowest in patients treated with lamotrigine. There is a slight decrease in the risk of failure with an increase in age and men have a slightly lower chance of treatment failure than women. There is also a very slight increase in the chance of treatment failure with an increase in total number of seizures ever before randomisation.

Table 6: The SANAD Trial – arm A: Univariate and Multivariable Hazard Ratios and Confidence Intervals for Overall Time to Treatment Failure

			Hazard Ratio (95% CI)	
Prognostic Factor	Comparison	Univariate p-value	Univariate	Multivariable
Gender	Female Male	0.01	1.00 0.82 (0.72 to 0.95)	1.00 0.86 (0.75 to 0.99)
Febrile seizure history	Absent Present	0.78	1.00 1.05 (0.77 to 1.43)	N/A
First degree relative with epilepsy	Absent Present	0.15	1.00 1.18 (0.95 to 1.48)	NA
CT/MR scan result	Normal Abnormal Not Done	0.79	1.00 1.06 (0.90 to 1.24) 1.01 (0.82 to 1.24)	N/A
Treatment history	Treatment naive Seizures after remission Taking non-SANAD AEDs	0.01	1.00 1.44 (0.93 to 2.24) 1.32 (1.10 to 1.59)	1.00 1.35 (0.87 to 2.10) 1.27 (1.05 to 1.53)
Age (years) [Linear]	≤10 (10 to 24) (25 to 36) (37 to 49) (50 to 70) >71	0.02	1.00 0.95 (0.92 to 0.99) 0.89 (0.82 to 0.98) 0.84 (0.73 to 0.97) 0.78 (0.63 to 0.96) 0.71 (0.54 to 0.94)	1.00 0.95 (0.91 to 0.99) 0.88 (0.80 to 0.97) 0.82 (0.71 to 0.96) 0.76 (0.61 to 0.94) 0.68 (0.51 to 0.91)
Time from 1st seizure (months) [Log]	≤2 (2 to 5) (6 to 17) (18 to 59) (60 to 239) >240	0.00	1.00 1.04 (1.01 to 1.06) 1.07 (1.02 to 1.12) 1.10 (1.03 to 1.18) 1.11 (1.05 to 1.19) 1.17 (1.05 to 1.30)	N/A
Neurological insult	Absent Present	0.67	1.00 1.05 (0.85 to 1.30)	N/A
Total number of seizures [Scale: 100, power: 0]	≤2 (2 to 3) (4 to 11) (12 to 50) (51 to 299) >300	0.00	1.00 1.03 (1.02 to 1.03) 1.09 (1.07 to 1.12) 1.19 (1.13 to 1.25) 1.33 (1.23 to 1.44) 1.61 (1.41 to 1.84)	1.00 1.02 (1.01 to 1.03) 1.08 (1.05 to 1.11) 1.17 (1.10 to 1.23) 1.28 (1.17 to 1.40) 1.52 (1.31 to 1.76)
EEG results	Normal Not done Non-specific abnormality Epileptiform abnormality	0.02	1.00 1.22 (0.95 to 1.57) 1.22 (0.99 to 1.49) 1.27 (1.08 to 1.50)	1.00 1.25 (0.96 to 1.61) 1.20 (0.98 to 1.47) 1.26 (1.07 to 1.50)
Epilepsy type	Focal Unclassified	0.02	1.00 0.74 (0.57 to 0.96)	N/A
Seizure type	Simple or complex partial only 2° Generalised tonic-clonic Uncertain	0.00	1.00 0.77 (0.66 to 0.89) 0.62 (0.47 to 0.81)	1.00 0.78 (0.66 to 0.91) 0.33 (0.05 to 2.37)
Focal epilepsy site of onset	Temporal Not localised Frontal Other Unclassified	0.13	1.00 1.04 (0.89 to 1.22) 1.08 (0.81 to 1.45) 0.87 (0.63 to 1.20) 0.75 (0.57 to 0.99)	1.00 1.25 (1.06 to 1.47) 1.18 (0.88 to 1.58) 0.92 (0.66 to 1.28) 2.69 (0.37 to 19.74)
Treatment	Carbamazepine Gabapentin Lamotrigine Oxcarbazepine Topiramate	0.00	1.00 1.18 (0.96 to 1.45) 0.75 (0.60 to 0.94) 0.93 (0.71 to 1.22) 1.21 (0.99 to 1.49)	1.00 1.23 (1.00 to 1.51) 0.76 (0.61 to 0.95) 0.94 (0.72 to 1.23) 1.23 (1.00 to 1.52)

Hazard ratio greater than 1 indicates that treatment failure is more likely in the alternative group than in the baseline group.

Risk Factors				Patients with treatment failure at 1 year % (95% CI)	Patients with treatment failure at 3 years % (95% CI)
Age	**Gender**	**Seizures**	**Treatment**	**Total**	**Total**
10	M	2	CBZ	27 (21 to 35)	42 (33 to 52)
10	F	2	CBZ	31 (24 to 39)	47 (37 to 58)
10	M	2	LTG	21 (16 to 28)	34 (26 to 43)
10	F	2	LTG	24 (18 to 31)	38 (30 to 48)
10	M	10	CBZ	32 (25 to 40)	49 (40 to 59)
10	F	10	CBZ	36 (29 to 45)	54 (45 to 64)
10	M	10	LTG	25 (20 to 32)	40 (32 to 49)
10	F	10	LTG	29 (22 to 36)	45 (36 to 54)
40	M	2	CBZ	23 (18 to 30)	37 (30 to 46)
40	F	2	CBZ	27 (21 to 34)	42 (34 to 51)
40	M	2	LTG	18 (14 to 24)	30 (23 to 38)
40	F	2	LTG	21 (16 to 27)	34 (26 to 42)
40	M	10	CBZ	28 (23 to 34)	44 (36 to 52)
40	F	10	CBZ	32 (26 to 39)	49 (41 to 57)
40	M	10	LTG	22 (17 to 28)	35 (28 to 43)
40	F	10	LTG	25 (20 to 31)	40 (32 to 48)
75	M	2	CBZ	20 (15 to 26)	32 (25 to 41)
75	F	2	CBZ	23 (17 to 30)	36 (28 to 45)
75	M	2	LTG	15 (11 to 21)	25 (19 to 33)
75	F	2	LTG	18 (13 to 24)	29 (22 to 37)
75	M	10	CBZ	24 (19 to 30)	38 (30 to 47)
75	F	10	CBZ	27 (21 to 34)	42 (34 to 52)
75	M	10	LTG	19 (14 to 24)	30 (23 to 38)
75	F	10	LTG	21 (16 to 28)	34 (27 to 43)

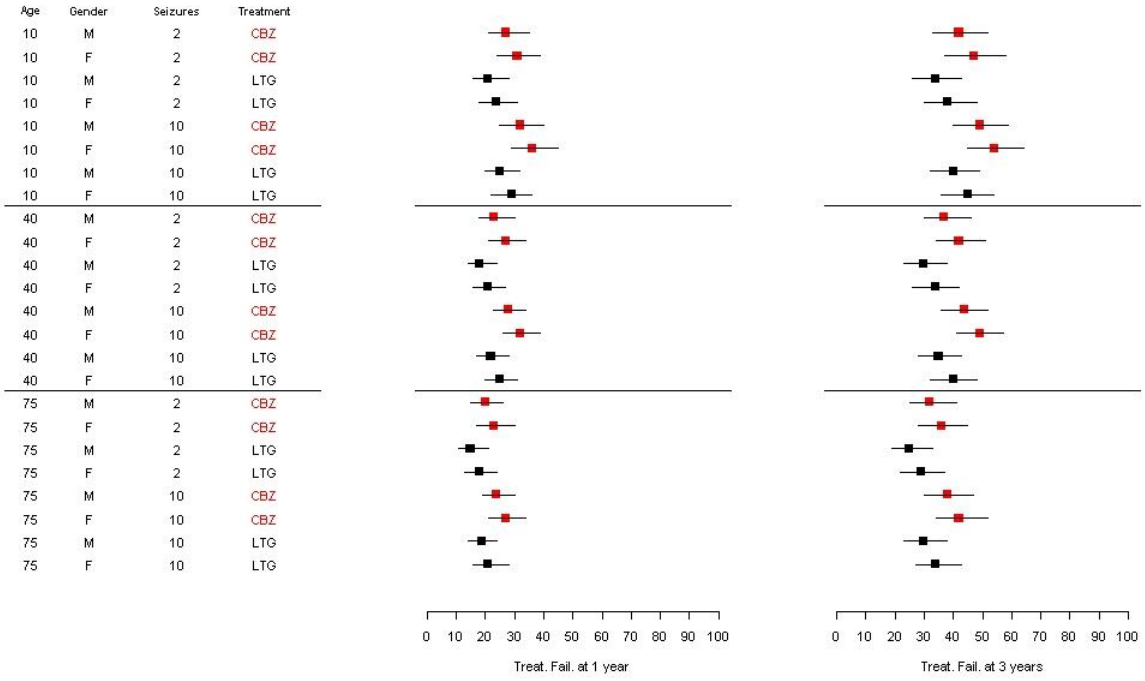


Figure 6: The SANAD Trial – arm A: Combination of Risk Factors for Time to Treatment Failure

(N.B. 'Seizures' is 'Total Number of Seizures before randomisation' variable)

4.3.2 Time to Treatment Failure due to Competing Risks

Table 7: The SANAD Trial – arm A: Multivariable Hazard Ratios and Confidence Intervals for Time to Treatment Failure - Competing Risks Analysis: Full model (all variables)

		Hazard Ratio (95% CI)	
Prognostic Factor	**Comparison**	**Inadequate Seizure Control**	**Unacceptable Adverse Events**
Gender	Female	1.00	1.00
	Male	1.06 (0.85 to 1.31)	0.80 (0.66 to 0.98)
Febrile seizure history	Absent	1.00	1.00
	Present	0.90 (0.55 to 1.49)	1.09 (0.69 to 1.72)
First degree relative	Absent	1.00	1.00
	Present	0.88 (0.62 to 1.27)	1.26 (0.92 to 1.72)
CT/MR scan result	Normal	1.00	1.00
	Abnormal	1.38 (1.08 to 1.77)	0.90 (0.70 to 1.15)
	Not Done	0.77 (0.55 to 1.08)	1.16 (0.87 to 1.56)
Treatment history	Treatment naive	1.00	1.00
	Seizures after remission	0.58 (0.22 to 1.50)	1.77 (0.99 to 3.17)
	Taking non-SANAD AED	1.76 (1.34 to 2.31)	0.92 (0.68 to 1.23)
Age (years) [Linear]	≤10	1.00	1.00
	(10 to 24)	0.84 (0.79 to 0.90)	1.06 (1.01 to 1.12)
	(25 to 36)	0.67 (0.57 to 0.78)	1.15 (1.01 to 1.31)
	(37 to 49)	0.54 (0.42 to 0.68)	1.24 (1.02 to 1.52)
	(50 to 70)	0.40 (0.29 to 0.57)	1.38 (1.03 to 1.84)
	>71	0.29 (0.18 to 0.47)	1.54 (1.04 to 2.28)
Time from 1st seizure (years) [Log]	≤2	1.00	1.00
	(2 to 5)	0.90 (0.86 to 0.95)	1.08 (1.03 to 1.12)
	(6 to 17)	0.83 (0.77 to 0.90)	1.14 (1.05 to 1.23)
	(18 to 59)	0.76 (0.68 to 0.86)	1.21 (1.08 to 1.36)
	(60 to 239)	0.76 (0.67 to 0.86)	1.22 (1.09 to 1.38)
	>240	0.65 (0.54 to 0.79)	1.36 (1.13 to 1.64)
Neurological insult	Absent	1.00	1.00
	Present	0.83 (0.59 to 1.16)	1.07 (0.78 to 1.48)
Total number of seizures [Scale: 100, power: 0]	≤2	1.00	1.00
	(2 to 3)	1.05 (1.04 to 1.07)	0.99 (0.97 to 1.00)
	(4 to 11)	1.20 (1.15 to 1.26)	0.95 (0.91 to 1.00)
	(12 to 50)	1.44 (1.32 to 1.57)	0.91 (0.83 to 1.00)
	(51 to 299)	1.80 (1.56 to 2.09)	0.86 (0.74 to 1.00)
	>300	2.69 (2.11 to 3.44)	0.77 (0.60 to 1.00)
EEG results	Normal	1.00	1.00
	Missing	0.78 (0.50 to 1.22)	1.34 (0.94 to 1.91)
	Non-specific abnormality	0.99 (0.72 to 1.37)	1.20 (0.90 to 1.60)
	Epileptiform abnormality	1.13 (0.88 to 1.46)	1.15 (0.90 to 1.48)
Seizure type	Simple or complex focal only	1.00	1.00
	2° Generalised tonic-clonic	1.07 (0.84 to 1.36)	0.66 (0.52 to 0.82)
	Unclassified	0.77 (0.10 to 6.08)	Could not be estimated
Focal Epilepsy Site of Onset	Temporal	1.00	1.00
	Not Localised	1.19 (0.92 to 1.52)	1.18 (0.93 to 1.48)
	Frontal	0.93 (0.61 to 1.44)	1.31 (0.85 to 2.01)
	Other	1.05 (0.66 to 1.67)	0.76 (0.45 to 1.29)
	Unclassified	1.38 (0.17 to 11.20)	Could not be estimated
Treatment	Carbamazepine	1.00	1.00
	Gabapentin	2.53 (1.85 to 3.47)	0.58 (0.43 to 0.79)
	Lamotrigine	1.08 (0.77 to 1.52)	0.62 (0.45 to 0.84)
	Oxcarbazepine	1.14 (0.73 to 1.77)	0.84 (0.59 to 1.20)
	Topiramate	1.45 (1.04 to 2.03)	1.02 (0.78 to 1.33)

Table 7 shows multivariable results for a competing risks analysis for time to treatment failure for our two treatment failure categories; inadequate seizure control and adverse events. Two models were fitted, one using all thirteen potential prognostic factors and one using only the variables included in the parsimonious multivariable model for overall treatment failure. This latter model was fitted to determine if any of the results seen in the overall model can be explained by the effects of either or both of the competing risks. Results for the model using only the variables included in the parsimonious model for treatment failure can be seen in Table 8. Results for the full multivariable model for the competing risks outcome can be seen in Table 7.

In Table 7 two results could not be estimated. The reason for the unusual results is that the two categories in question have exactly the same patients with treatment failure due to unacceptable adverse events. A sensitivity analysis was undertaken to see the effect of this occurrence and it was found to have no implication on the model - the model was robust to removal of seizure type and focal epilepsy site of onset in turn and also to collapsing of the latter two categories of each of the offending variables in turn.

4.3.2.1 Treatment Failure due to Inadequate Seizure Control

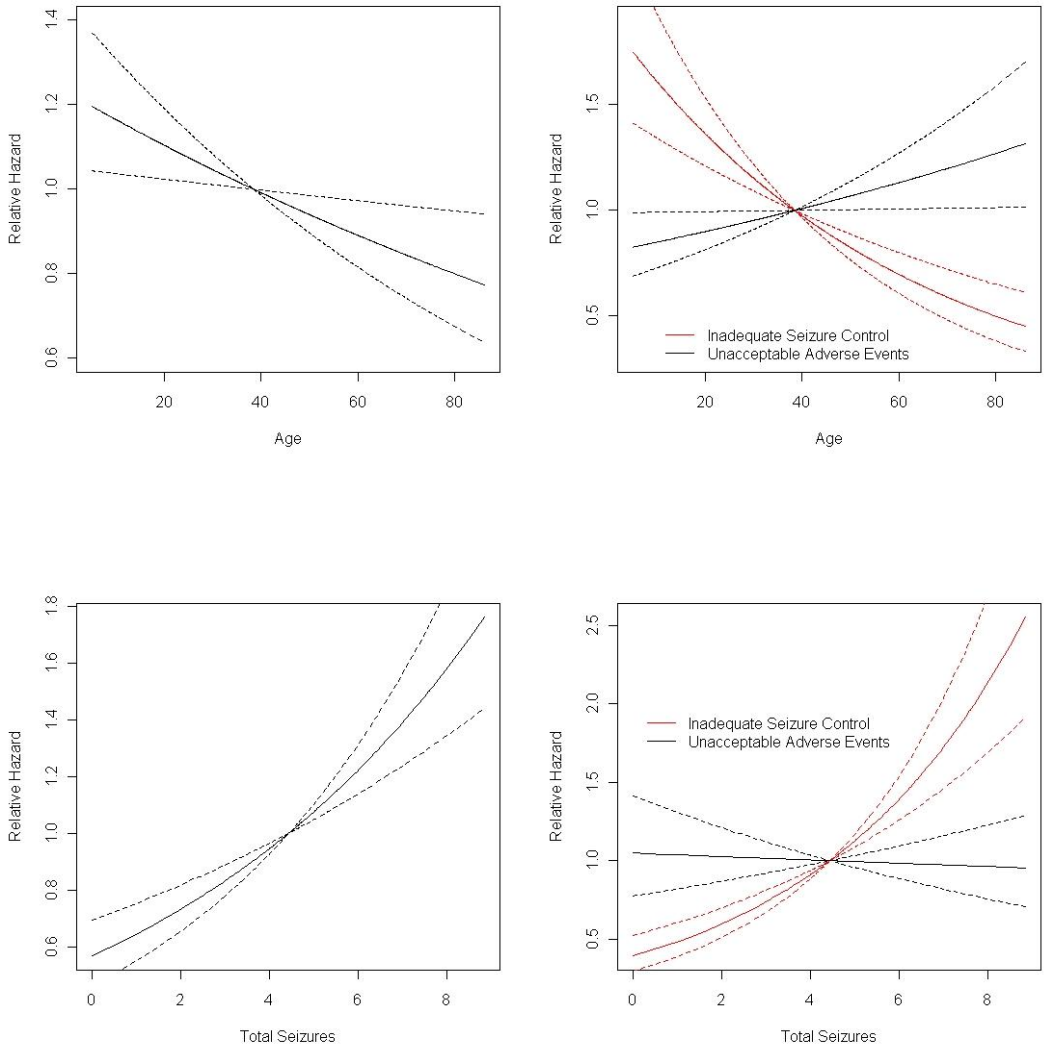
For time to treatment failure due to inadequate seizure control the significant variables (5% level) according to the matched multivariable model (Table 8) were treatment history, age, number of seizures before randomisation, and treatment. Compared to treatment naive patients, patients already taking a non-SANAD AED had a higher treatment failure rate (taking non-SANAD AED vs. treatment naive: HR 1.56 95% CI (1.20 to 2.03)). For age a linear effect was observed with older patients significantly less likely to have a treatment failure than younger patients (see discussion below also), while treatment failure rates increased significantly with increasing number of seizures before randomisation.

4.3.2.2 Treatment Failure due to Unacceptable Adverse Events

For treatment failure due to adverse events the variables significantly (5% level) associated with outcome, according to the matched multivariable model (Table 8), were gender, treatment history, age, seizure type, and treatment. Women were more likely to have treatment failure than men (male vs. female: HR 0.81 95% CI (0.66 to 0.98)). Patients restarting treatment following a period of remission had a higher treatment failure rate than treatment naive patients (seizures after remission vs. treatment naive: HR 2.13 95% CI (1.22 to 3.73)). For age a linear effect was observed with older patients significantly more likely to have a treatment failure than younger patients (see discussion below also). Patients with simple or complex partial seizures were significantly more likely to have a treatment failure than patients with secondary generalised tonic clonic seizures (secondary generalised tonic clonic vs. simple or complex partial: HR 0.69 95% CI (0.55 to 0.86)).

As shown in Figure 7, for time to treatment failure for any reason, according to the multivariable model, as age increases the likelihood of treatment failure decreases. However when the reason for treatment failure is considered an X-shaped plot is produced; the likelihood of treatment failure due to inadequate seizure control decreases with age while the likelihood of treatment failure due to unacceptable adverse events increases with age. More explicitly, younger patients are at a higher risk of withdrawal due to inadequate seizure control whereas younger patients are at a lower risk of withdrawal due to unacceptable adverse events. A similar X-shaped plot is produced for number of seizures before randomisation. In the analysis of treatment failure for any reason the likelihood of treatment failure increased with number of seizures. When the reason for treatment is considered, the likelihood of treatment failure due to inadequate seizure control increases with number of seizures while the likelihood of failure due to adverse events decreases (slightly) with number of seizures, therefore patients with a smaller number of seizures

before randomisation are at a higher risk of withdrawal due to unacceptable adverse events but lower risk of withdrawal due to inadequate seizure control.



Solid lines are used for hazard ratio estimates while dotted lines represent confidence intervals.

Figure 7: The SANAD Trial – arm A: Relative Hazard Plots for Age and Total Number of Seizures for the Outcome Time to Treatment Failure

Table 8: The SANAD Trial – arm A: Univariate and Multivariable Hazard Ratios and Confidence Intervals for Time to Treatment Failure - Competing Risks Analysis

		Hazard Ratio (95% CI)			
		Inadequate Seizure Control		Unacceptable Adverse Events	
Prognostic Factor	Comparison	Univariate	Multivariable	Univariate	Multivariable
Gender	Female Male	1.00 0.96 (0.78 to 1.18)	1.00 1.06 (0.86 to 1.31)	1.00 0.79 (0.65 to 0.97)	1.00 0.81 (0.66 to 0.98)
Febrile seizure history	Absent Present	1.00 0.94 (0.59 to 1.49)	N/A	1.00 1.06 (0.68 to 1.65)	N/A
First degree relative	Absent Present	1.00 0.99 (0.70 to 1.41)	N/A	1.00 1.19 (0.87 to 1.62)	N/A
CT/MR scan result	Normal Abnormal Not Done	1.00 1.18 (0.94 to 1.49) 0.74 (0.53 to 1.03)	N/A	1.00 0.98 (0.77 to 1.24) 1.21 (0.93 to 1.58)	N/A
Treatment history	Treatment naive Seizures after remission Taking non-SANAD AED	1.00 0.61 (0.25 to 1.50) 1.72 (1.34 to 2.19)	1.00 0.48 (0.19 to 1.23) 1.56 (1.20 to 2.03)	1.00 1.89 (1.10 to 3.24) 0.95 (0.72 to 1.26)	1.00 2.13 1.22 to 3.73) 1.00 (0.75 to 1.33)
Age (years) [Linear]	≤10 (10 to 24) (25 to 36) (37 to 49) (50 to 70) >71	1.00 0.85 (0.80 to 0.90) 0.68 (0.60 to 0.78) 0.56 (0.45 to 0.69) 0.43 (0.32 to 0.58) 0.32 (0.21 to 0.48)	1.00 0.85 (0.79 to 0.90) 0.68 (0.58 to 0.78) 0.55 (0.44 to 0.69) 0.42 (0.30 to 0.58) 0.31 (0.19 to 0.48)	1.00 1.07 (1.02 to 1.13) 1.18 (1.04 to 1.34) 1.29 (1.07 to 1.56) 1.45 (1.10 to 1.91) 1.65 (1.13 to 2.39)	1.00 1.06 (1.00 to 1.12) 1.14 (1.01 to 1.30) 1.23 (1.01 to 1.49) 1.35 (1.02 to 1.79) 1.50 (1.02 to 2.20)
Time from 1st seizure (months) [Log]	≤2 (2 to 5) (6 to 17) (18 to 59) (60 to 239) >240	1.00 1.00 (0.96 to 1.03) 0.99 (0.93 to 1.06) 0.99 (0.89 to 1.09) 0.99 (0.89 to 1.10) 0.98 (0.84 to 1.15)	N/A	1.00 1.06 (1.02 to 1.09) 1.10 (1.04 to 1.17) 1.16 (1.05 to 1.27) 1.16 (1.05 to 1.28) 1.26 (1.09 to 1.46)	N/A
Neurological insult	Absent Present	1.00 1.09 (0.80 to 1.47)	N/A	1.00 1.03 (0.76 to 1.40)	N/A
Total number of seizures [Scale: 100, power: 0]	≤2 (2 to 3) (4 to 11) (12 to 50) (51 to 299) >300	1.00 1.04 (1.03 to 1.05) 1.14 (1.10 to 1.18) 1.29 (1.21 to 1.38) 1.52 (1.36 to 1.69) 2.01 (1.67 to 2.42)	1.00 1.04 (1.03 to 1.05) 1.14 (1.09 to 1.18) 1.29 (1.19 to 1.39) 1.51 (1.33 to 1.71) 2.00 (1.62 to 2.47)	1.00 1.00 (0.99 to 1.01) 1.01 (0.98 to 1.05) 1.02 (0.95 to 1.10) 1.04 (0.93 to 1.17) 1.07 (0.88 to 1.30)	1.00 1.00 (0.99 to 1.01) 0.99 (0.95 to 1.04) 0.99 (0.91 to 1.07) 0.98 (0.86 to 1.12) 0.97 (0.77 to 1.21)
EEG results	Normal Not done Non-specific abnormality Epileptiform abnormality	1.00 0.81 (0.53 to 1.24) 1.02 (0.75 to 1.38) 1.34 (1.07 to 1.69)	1.00 0.79 (0.51 to 1.22) 1.01 (0.74 to 1.38) 1.21 (0.94 to 1.55)	1.00 1.33 (0.95 to 1.86) 1.24 (0.94 to 1.64) 1.34 (1.07 to 1.69)	1.00 1.34 (0.95 to 1.88) 1.21 (0.91 to 1.61) 1.11 (0.87 to 1.41)
Epilepsy type	Focal Unclassified	1.00 0.64 (0.43 to 0.96)	N/A	1.00 0.90 (0.64 to 1.27)	N/A
Seizure type	Simple or complex partial only 2° Generalised tonic-clonic Unclassified	1.00 0.98 (0.78 to 1.22) 0.65 (0.42 to 0.99)	1.00 1.01 (0.79 to 1.28) 0.76 (0.08 to 7.29)	1.00 0.71 (0.58 to 0.87) 0.71 (0.47 to 1.02)	1.00 0.69 (0.55 to 0.86) Could not be estimated

		Hazard Ratio (95% CI)			
		Inadequate Seizure Control		Unacceptable Adverse Events	
Prognostic Factor	Comparison	Univariate	Multivariable	Univariate	Multivariable
Focal epilepsy site of onset	Temporal Not localised Frontal Other Unclassified	1.00 1.00 (0.79 to 1.26) 1.04 (0.68 to 1.58) 1.08 (0.69 to 1.68) 0.65 (0.42 to 0.99)	1.00 1.18 (0.92 to 1.51) 1.04 (0.69 to 1.59) 1.04 (0.66 to 1.66) 1.28 (0.13 to 12.72)	1.00 1.04 (0.83 to 1.30) 1.13 (0.75 to 1.70) 0.70 (0.43 to 1.15) 0.91 (0.63 to 1.31)	1.00 1.17 (0.93 to 1.48) 1.24 (0.81 to 1.90) 0.76 (0.45 to 1.26) Could not be estimated
Treatment	Carbamazepine Gabapentin Lamotrigine Oxcarbazepine Topiramate	1.00 2.40 (1.77 to 3.25) 1.10 (0.78 to 1.55) 1.12 (0.73 to 1.71) 1.39 (1.00 to 1.94)	1.00 2.45 (1.80 to 3.34) 1.05 (0.75 to 1.48) 1.12 (0.73 to 1.73) 1.44 (1.03 to 2.00)	1.00 0.58 (0.43 to 0.78) 0.61 (0.45 to 0.82) 0.82 (0.57 to 1.16) 1.01 (0.77 to 1.31)	1.00 0.59 (0.43 to 0.80) 0.62 (0.46 to 0.85) 0.84 (0.59 to 1.20) 1.01 (0.77 to 1.32)

Hazard ratio greater than 1 indicates that treatment failure is more likely in the alternative group than in the baseline group.

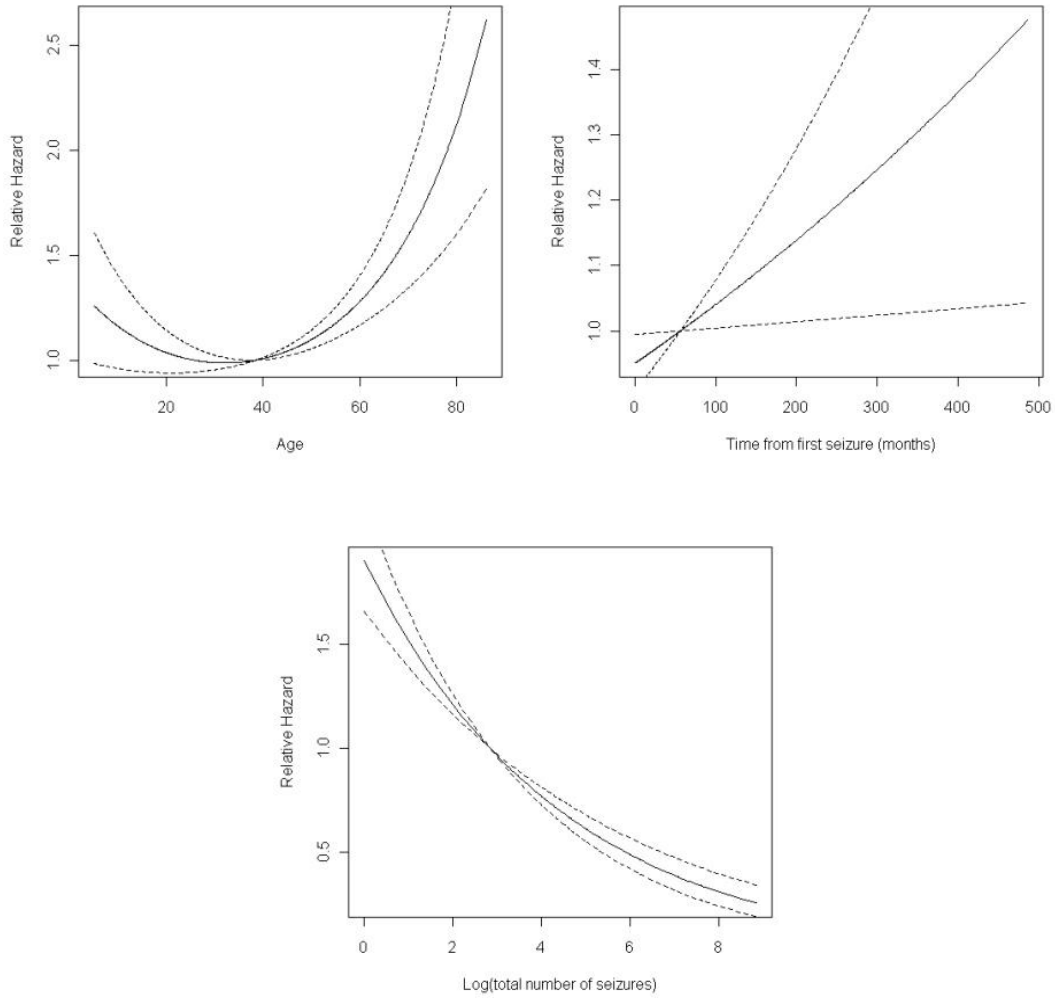
4.3.3 Time to 12 Month Remission

Results for both the univariate and the multivariable analyses for the outcome time to 12 month remission are summarised in Table 9. In the univariate analyses the variables significantly associated with time to 12 month remission at the 5% level were gender, treatment history, age, time from first seizure to randomisation, neurological insult, total number of seizures before randomisation, epilepsy type, seizure type, focal epilepsy site of onset, and treatment. Of these, the parsimonious multivariable model did not include epilepsy type or seizure type but did include CT/MR scan results.

Table 9: The SANAD Trial – arm A: Univariate and Multivariable Hazard Ratios and 95% Confidence Intervals for Time to 12 Month Remission

			Hazard Ratio (95% Confidence Interval)	
Prognostic Factor	Comparison	Univariate p-value	Univariate	Multivariable
Gender	Female Male	0.00	1.00 1.21 (1.07 to 1.37)	1.00 1.19 (1.05 to 1.35)
Febrile seizure history	Absent Present	0.89	1.00 1.02 (0.78 to 1.34)	N/A
First degree relative with epilepsy	Absent Present	0.85	1.00 0.98 (0.80 to 1.20)	N/A
CT/MR scan result	Normal Abnormal Not Done	0.08	1.00 0.89 (0.77 to 1.03) 1.12 (0.94 to 1.33)	1.00 0.88 (0.76 to 1.03) 1.12 (0.93 to 1.34)
Treatment history	Treatment naive Seizures after remission Taking non-SANAD AED	0.00	1.00 1.09 (0.72 to 1.63) 0.55 (0.46 to 0.67)	1.00 1.06 (0.70 to 1.61) 0.64 (0.52 to 0.78)
Age (years) [Scale: 100, powers: 1, 2]	≤10 (10 to 24) (25 to 36) (37 to 49) (50 to 70) >71	0.00	1.00 0.84 (0.77 to 0.93) 0.76 (0.64 to 0.91) 0.79 (0.63 to 0.98) 1.00 (0.81 to 1.24) 1.56 (1.26 to 1.98)	1.00 0.88 (0.79 to 0.97) 0.82 (0.68 to 0.99) 0.85 (0.68 to 1.07) 1.07 (0.86 to 1.34) 1.60 (1.26 to 2.03)
Time from 1st seizure (months) [Scale: 1000, power: 1]	≤2 (2 to 5) (6 to 17) (18 to 59) (60 to 239) >240	0.00	1.00 1.00 (0.99 to 1.00) 0.98 (0.98 to 0.99) 0.95 (0.93 to 0.98) 0.95 (0.92 to 0.97) 0.69 (0.57 to 0.84)	1.00 1.00 (1.00 to 1.01) 1.01 (1.00 to 1.02) 1.03 (1.00 to 1.07) 1.14 (1.01 to 1.29) 1.39 (1.04 to 1.86)
Neurological insult	Absent Present	0.00	1.00 0.72 (0.59 to 0.89)	1.00 0.75 (0.61 to 0.93)
Total number of seizures [Log]	≤2 (2,3) (4,11) (12,50) (51,299) >300	0.00	1.00 0.96 (0.95 to 0.97) 0.87 (0.85 to 0.89) 0.76 (0.72 to 0.80) 0.64 (0.59 to 0.69) 0.47 (0.41 to 0.54)	1.00 0.96 (0.95 to 0.97) 0.87 (0.85 to 0.90) 0.76 (0.72 to 0.81) 0.65 (0.59 to 0.71) 0.48 (0.41 to 0.56)
EEG results	Normal Not done Non-specific abnormality Specific abnormality	0.73	1.00 0.96 (0.76 to 1.22) 1.09 (0.92 to 1.31) 1.00 (0.86 to 1.15)	N/A
Epilepsy type	Focal Unclassified	0.00	1.00 1.59 (1.31 to 1.93)	N/A
Seizure type	Simple or complex partial only 2° Generalised tonic-clonic Uncertain	0.00	1.00 1.01 (0.88 to 1.16) 1.63 (1.32 to 2.01)	N/A
Focal epilepsy site of onset	Temporal Not localised Frontal Other Unclassified	0.00	1.00 0.98 (0.85, 1.13) 1.02 (0.79, 1.32) 1.10 (0.84, 1.42) 1.59 (1.29, 1.96)	1.00 0.87 (0.75, 1.02) 1.12 (0.86, 1.45) 1.02 (0.78, 1.34) 1.15 (0.87, 1.44)
Treatment	Carbamazepine Gabapentin Lamotrigine Oxcarbazepine Topiramate	0.02	1.00 0.75 (0.62 to 0.90) 0.93 (0.78 to 1.11) 0.99 (0.79 to 1.24) 0.84 (0.70 to 1.01)	1.00 0.71 (0.59 to 0.86) 0.90 (0.75 to 1.08) 0.97 (0.77 to 1.22) 0.81 (0.68 to 0.98)

Hazard ratio greater than 1 indicates that 12 month remission is more likely in the alternative group than in the baseline group.



Solid lines represent the hazard ratio estimates while the dotted lines represent the confidence interval estimates.

Figure 8: The SANAD Trial – arm A: Hazard ratio plots for age, time from first seizure to randomisation and total number of seizures for the outcome time to 12 month remission

Focusing on the results of the multivariable model, for age, the relation is U-shaped with patients aged less than or equal to 10 years old or over 71 years having a significantly higher chance of remission (Figure 8). The relationship between the other significant (5% level) variables in the multivariable model and time to 12 month remission are as follows; males had a higher 12 month remission rate than females (male vs. female: HR 1.19 95% CI (1.05 to 1.35)); treatment naive patients had a higher 12 month remission rate than

patients currently taking a non-SANAD antiepileptic drug (taking non-SANAD AED vs. treatment naive: HR 0.64 (0.52 to 0.78)); 12-month remission rates decreased with an increasing number of seizures before randomisation, for example compared to patients with two seizures the estimate for patients with two to three seizures is 0.96 (0.95 to 0.97) and for patients with four to 11 seizures is 0.87 (0.85 to 0.90). Regarding treatment, 12 month remission rates were significantly higher for patients on carbamazepine than for those on gabapentin (gabapentin vs. carbamazepine: 0.71 (0.59 to 0.86)) or topiramate (topiramate vs. carbamazepine: HR 0.81 (0.68 to 0.98)). Plots for the continuous variables within the multivariable model can be seen in Figure 8.

In order to illustrate the range of 12 month remission rates predicted by the multivariable model, Figure 9 shows estimates of the proportion of patients achieving a remission one and three years after randomisation for patients treated with either carbamazepine or lamotrigine together with the associated 95% confidence intervals. Patients were assumed to be treatment naive, have been randomised six months from their first seizure, not have neurological insult, have normal CT/MR result and have temporal lobe epilepsy. In the subset of combinations considered, remission rates are highest in older patients. Males are slightly more likely to achieve remission than women as are patients with fewer seizures. There is a very minor increase in probability of remission in patients on carbamazepine compared to those on lamotrigine.

Risk Factors				Percentage of patients achieving 12 month remission, together with 95% confidence intervals at:	
Age	**Gender**	**Seizures before randomisation**	**Treatment**	**1 year**	**3 years**
10	Male	2	CBZ	47 (39 to 56)	92 (86 to 97)
10	Female	2	CBZ	42 (34 to 50)	89 (81 to 94)
10	Male	2	LTG	44 (36 to 52)	90 (83 to 95)
10	Female	2	LTG	38 (31 to 47)	86 (78 to 92)
10	Male	10	CBZ	36 (30 to 43)	83 (76 to 90)
10	Female	10	CBZ	31 (25 to 38)	78 (70 to 85)
10	Male	10	LTG	33 (27 to 40)	80 (72 to 87)
10	Female	10	LTG	29 (23 to 35)	74 (66 to 82)
40	Male	2	CBZ	43 (36 to 50)	89 (83 to 94)
40	Female	2	CBZ	37 (31 to 45)	85 (78 to 91)
40	Male	2	LTG	39 (32 to 47)	87 (79 to 92)
40	Female	2	LTG	34 (28 to 41)	82 (74 to 88)
40	Male	10	CBZ	32 (27 to 38)	79 (72 to 86)
40	Female	10	CBZ	28 (23 to 33)	73 (66 to 80)
40	Male	10	LTG	29 (24 to 35)	75 (67 to 83)
40	Female	10	LTG	25 (21 to 31)	69 (61 to 77)
75	Male	2	CBZ	63 (53 to 73)	98 (95 to 100)
75	Female	2	CBZ	57 (48 to 67)	97 (92 to 99)
75	Male	2	LTG	59 (49 to 70)	97 (93 to 99)
75	Female	2	LTG	53 (44 to 63)	95 (90 to 98)
75	Male	10	CBZ	50 (42 to 60)	94 (88 to 98)
75	Female	10	CBZ	45 (36 to 53)	91 (84 to 95)
75	Male	10	LTG	47 (38 to 56)	92 (85 to 96)
75	Female	10	LTG	41 (33 to 50)	88 (80 to 94)

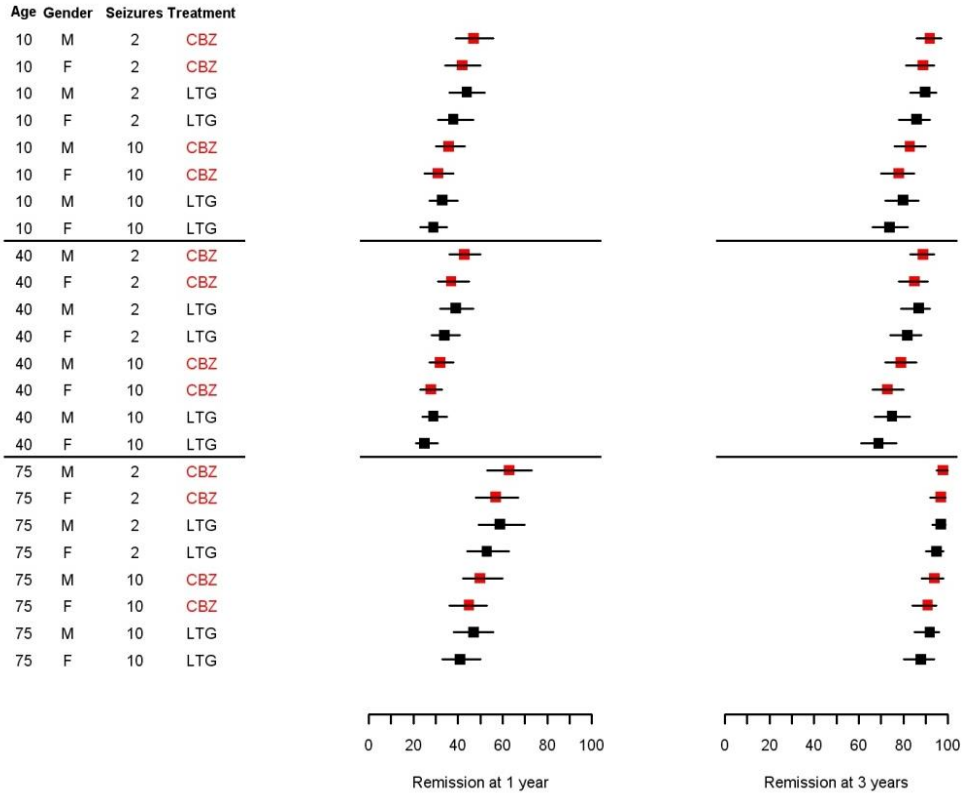


Figure 9: The SANAD Trial – arm A: Combination of Risk Factors for Time to 12 Month Remission

(N.B. 'Seizures' is 'Total Number of Seizures before randomisation' variable)

4.4 Discussion

Predicting outcome in epilepsy is complex but necessary to inform patient counselling about prognosis and treatment choices. Numerous clinical factors have been indentified in this chapter that influence the outcome measures time to treatment failure and time to 12 month remission. Multivariable models have also been produced that will inform clinicians and patients about likely outcomes.

SANAD remains the largest randomised control trial in epilepsy and includes data on long term treatment outcomes which is vital if we are to inform the management of this chronic condition. The dataset provides an unrivalled opportunity to investigate factors that influence treatment outcome for patients starting antiepileptic drug monotherapy. SANAD recruited a heterogeneous group of patients which some have criticised [188, 189] but this can also be argued as a strength, as illustrated by this chapter, as the inclusion of a large heterogeneous population of patients enables a thorough investigation of factors that influence treatment outcome.

Given that SANAD is a large dataset it is perhaps unsurprising that around nine prognostic covariates have been identified. Obviously, of more importance than the statistical significance of these covariates is the ability of the models to predict outcome. In this chapter we only provide a sense of this via the tables of probabilities of events at one and three years for various covariate combinations. In Chapter 7, internal validation of these models is considered which provides a measure of how good they are at predicting the outcomes.

Blinding a long term trial such as SANAD poses significant challenges, for example, patient groups such as women of child-bearing age potentially might be excluded. The blinding would need a double dummy approach which could affect compliance and would come at significant expense such that it would be unaffordable to public funders such as the

National Institute for Health Research Health Technology Assessment programme. SANAD was therefore an un-blinded trial which could have influenced outcome assessment, for example decisions as to whether a treatment had failed, although examination of dosing data indicates that reasonable doses were tried before a decision was made that treatment had failed.

While randomised controlled trials are the best methodology for assessing treatment outcomes, they might recruit a selected population which might influence estimates of prognosis. Ideally SANAD should have recruited a greater proportion of children and elderly patients; nonetheless the analysis has clearly identified the influence of age on outcome. Additionally, the inclusion criteria for SANAD were broad. There were few exclusion criteria and the vast majority of patients with a new diagnosis of epilepsy were eligible for inclusion. Therefore, we believe that there are sufficient data provided for the clinicians to make a judgment. Hence, in general terms the SANAD recruits should be fairly representative of a completely unselected population based cohort, that, ideally, a prognostic model would be based on.

While some advocate separate trials for adults and children [190, 191], assuming that relative treatment responses differ according to age, our results emphasise the importance of trials that recruit a wide age range of patients, allowing a thorough investigation of the influence of age upon treatment outcome. This is reinforced by the X-shaped relationship between age and reasons for treatment failure. The notion that separate trials in children would be an inefficient approach to estimating treatment effects is supported by the EMAs decision to allow data from adults to be extrapolated to children down to the age of four for licensing purposes.

Although randomised controlled trials might be an optimal design to detect treatment effects, they may not be ideal for building prognostic models, given that trials tend to

exclude subjects with comorbidities for safety reasons, and possibly select amongst those that are eligible a more motivated group. However a population based study would be confounded in any assessment of relative treatment effects.

For a prognostic model, it might well be that there are subgroups in which the model does not predict well for example children vs. adults. We explored interactions for the models presented in this chapter. No treatment-covariate interaction terms were found for time to 12 month remission and two were found for time to treatment failure - focal epilepsy site of onset and total number of seizures ever before randomisation. Although these terms could easily be included in the multivariable model, interpretation of the results was very difficult due to the large number (6 levels of seizure category, 5 levels of treatment and 5 levels of lobe = 150) of possible combinations of interaction terms. Both interactions were significant for inadequate seizure control only, and not for failure due to adverse effects. Given that results for 12 month remission were also presented, which is a pure seizure outcome, we have not presented these data.

While we have presented a number of predictive models that can further inform patient counselling and treatment decisions, ideally these models require validation in other similar datasets and the predictive power of the model also needs to be explored. Unfortunately there are no other datasets that are similar to SANAD. The best match is a set of individual participant data collected by the Tudur Smith *et al* [192]. However this data is missing important covariates and the treatments patients were randomised to do not always coincide with SANAD drugs. Therefore more work needs to be undertaken to determine how best to overcome these difficulties – see Chapters 8 and 9.

While other prognostic models have been undertaken in epilepsy [119, 193, 194] no alternative epilepsy monotherapy trial has had sufficient power to thoroughly investigate prognostic factors. Related datasets include the NGPSE study [195] and the MESS study

[133] as described in section 2.6. NGPSE was a large prospective GP based observational study based on an unselected cohort of patients. The only independent predictor of one year and two year remission was the number of seizures experienced by the patient in the six months after the first seizure. It is important to note though that many patients in NGPSE were not prescribed antiepileptic drugs. MESS was an unmasked, multicentre, randomised study of immediate and deferred antiepileptic drug treatment in 1447 patients with single seizures and early epilepsy. Number of seizures of all types at presentation, presence of a neurological disorder, and an abnormal EEG were significant factors in indicating future seizures.

Our results highlight the heterogeneity in terms of outcome in epilepsy and the complex interplay between the factors that influence it, given the numerous factors identified in our models. Patients with differing risks of treatment failure, for unacceptable adverse events or inadequate seizures control, and differing risks of 12 month remission can however be identified at the point in time when antiepileptic drug treatment is initiated. This is emphasised by the data presented in Figure 6 and Figure 9. These data will inform patient counselling and inform decisions about monitoring and follow-up.

While clinical predictors of outcome have been identified, the mechanism by which these factors influence outcome remain poorly understood, and it is likely that many are surrogates for underlying pathological processes. Also, some variability in outcome remains unexplained, and we have not been able to identify factors that influence response to a specific drug. There is much interest in personalised medicine and pharmacogenetics, although to date this endeavour has not identified genetic factors that predict outcome. Studies such as SANAD in which clinical outcome data and DNA are collected alongside one another should provide a powerful resource to investigate clinical and genetic predictors of outcome, and consortia pulling together such cohorts are evolving and should soon be

providing results. We do of course need reliable evidence about outcome and predictors of outcome associated with newer drugs such as levetiracetam, zonisamide and lacosamide, and long term trials assessing these drugs are urgently required to inform clinical decision making and policy.

Chapter 5: Implications for Drivers with a First Seizure

5.1 Introduction

This Chapter begins with an overview of driving regulations both within the U.K. and further afield (section 5.1). In section 5.2 the Multicentre centre for Early epilepsy and Single Seizures (MESS) is introduced and the methods used to calculate risk of seizure recurrence following a first ever seizure are calculated. Results, together with recurrence risks for combinations of risk factors, are presented in section 5.3.

The work in this chapter has been published in the British Medical Journal (BMJ) and I am first author. The BMJ felt that the work was very important. Consequently, they simultaneously published an editorial about the paper. In addition, as a result of the analyses presented in this chapter and in the BMJ article, the DVLA have changed their guidelines concerning time off driving for patients with a first seizure. The work is now being considered by the European Union in an attempt to harmonise driving regulations across the EU.

Driving regulations in the European Union currently differ among member states [196] and there are extremes in regulations worldwide. In Bulgaria, Central African Republic, China, Estonia, Ghana, India, Korea, Pakistan, Portugal, Rwanda, Singapore, Taiwan, Turkey and Uzbekistan, a single seizure of any type, at any time in an individual's life causes indefinite loss of driving privileges. Conversely, in Argentina there are no specific laws prohibiting people with epilepsy from driving although patients generally rely on advice from their physicians [197, 198].

In 1995 and 1996 the International League Against Epilepsy and the International Bureau for Epilepsy organised workshops to discuss driving regulations around epilepsy and seizures, but the recommendations made were never reflected in an official European guideline or in European law. Following further calls for harmonisation across the European Union [199-201] member states have been engaged in a process of identifying minimum standards for driving, and these are now in the process of being implemented in the United Kingdom and other European Union states.

Where possible the United Kingdom based Driving and Vehicle Licensing Agency adopt a risk based approach to determine who can return to driving after an incapacitating event such as an epileptic seizure [202]. Epilepsy comes under the 1988 Road Traffic Act, but single seizures, which by definition do not constitute epilepsy, come under the domestic regulations (1999 Motor Vehicles (Driving Licences) Regulations). People who have had a single unprovoked seizure are usually allowed to regain their ordinary, group 1, driving licence six months after the event provided their risk of a seizure recurrence in the next year is below 20%. Drivers can regain their heavy goods vehicle, group 2, licence after five years provided their risk of a seizure recurrence is below 2% and they have been without the need for antiepileptic drug treatment for at least five years. Although these minimum risk levels of 20% and 2% are somewhat arbitrary, they are supported by other member states [201] and have been adopted in the criteria determining minimum driving standards that are being harmonised across the European Union.

Outside of the European Union there do not appear to be any countries that have such a risk based approach. Within the United States each individual state has its own legislation for driving with epilepsy and single seizures. When surveyed in 2001 [203] 28 states, out of a possible 50, including the District of Columbia, required people with epilepsy to be free of seizures for single fixed periods with a median restriction of six months (range, 3 to 12

months). 23 states adopted more flexible approaches to restricting driving such as varying seizure free restrictions based on individual clinical factors. However, many states allow patients to drive after shorter seizure free periods than stated in law which may be because these laws are usually unpublished and not readily accessible. In 13 states the time was decided by the treating doctor or a medical advisory board but in six of these states clinicians were not legally shielded for their assessments.

When implementing the new European Union legislation, member states may adopt the minimum standard and should not allow a more lenient standard. Member states can, however, adopt a stricter standard provided it can be justified, although it is likely that any state that sets a more stringent standard would face a challenge in court. In October 2009, after a discussion at a meeting of the Secretary of State for Transport's Joint Honorary Medical Advisor Panel on Vision and Neurology Disorders, at which the findings in this chapter were also discussed, the agency altered its guidance [202]. The new guidance stated that after a first unprovoked seizure, drivers may regain their ordinary, group 1, licence once they had been seizure free for six months, provided that their risk of a seizure recurrence in the next year was below the Driving and Vehicle Licensing Agency's 20% cut-off. Note that the 20% risk threshold was unchanged.

Very few published studies have estimated the risk of seizure recurrence and/or investigated factors that modify risk. Publications that do exist [119, 204] have focused on recurrence immediately after a first seizure and none have presented risks of recurrence in the next 12 months at time points such as six or 12 months after an index seizure. Reliable published data are therefore needed to inform Driving and Vehicle Licensing Agency guidance, European Union legislation, and legislation outside the European Union as well as decisions made by clinicians.

MESS [133] was a randomised controlled trial that compared the policies of immediate or deferred treatment in people presenting with a first unprovoked seizure or with early epilepsy. The trial remains the largest reported study of patients with single seizures and early epilepsy, and although the primary purpose of the study was to compare treatment policies, it also provided an important opportunity to examine the risks of seizure recurrence and factors that modify those risks. MESS was used to determine a predictive model for the risk of seizure recurrence after a single seizure [119]. The dataset is therefore ideal to obtain estimates of seizure recurrence within the next year at various time points following an index seizure.

In this chapter, data from participants with only a single unprovoked seizure, referred to as the index seizure, at entry into the MESS study was analysed to estimate seizure recurrence risk in the 12 months after the index seizure. Modelling was used to investigate how antiepileptic drug treatment and several clinical factors influence the risk of seizure recurrence.

5.2 Methods

Patients were eligible for inclusion in the MESS study [133] if they were at least one month old; had a history of at least one clinically definite, spontaneous, unprovoked epileptic seizure (excluding acute symptomatic seizures or febrile convulsions); and if both the clinician and the patient, or carer, were uncertain about whether or not to start antiepileptic drug treatment. Exclusion criteria included previous treatment with antiepileptic drugs or the presence of a progressive neurological disease.

People were recruited to the trial between 1 January 1993 and 31 December 2000 and were followed-up, if possible, from 31 December 2001 to 30 June 2002. After the recruiting clinician had obtained informed consent and enrolled the participant, the

participant was allocated randomly to treatment policy by telephone or facsimile by way of the minimisation method, which was balanced across centre or region and number of seizures at randomisation. For participants assigned to immediate treatment, the clinician selected the optimum antiepileptic drug treatment, dose and duration, based on their usual practice, and started treatment as early as possible. Participants assigned to deferred treatment received no medication until the clinician and patient agreed that it was necessary, mainly after a seizure recurrence.

Baseline clinical information and demographic data was obtained for all randomised participants and those eligible people who did not provide consent for randomisation. EEG was requested for all randomised participants, and brain imaging (CT/MRI) was undertaken if clinically indicated. Follow-up occurred at three, six, and 12 months and successive yearly intervals from randomisation, with additional visits as clinically indicated. At each visit the occurrence and type of any seizures was recorded, together with antiepileptic drugs taken, if appropriate, and any adverse events. Date and cause of death were also obtained as necessary.

In the original MESS study primary seizure outcomes were time from randomisation to first seizure of any type, time from randomisation to first tonic-clonic seizure, time from randomisation to second and fifth seizure of any type, and time from randomisation to two year remission of seizures. Additional primary outcomes were the proportion of patients seizure free for two years between one and three years after randomisation and three and five years after randomisation. Secondary clinical outcomes consisted of adverse events in each group and quality of life, assessed at randomisation, two years and four years.

To make the analyses reported here relevant to those of driving age, only participants aged 16 years or over were included. 16 was chosen as the lower cut-off as, by age 17, after 12 months of follow-up, these participants would be eligible to apply for a provisional licence.

Additionally, interest only lay in participants with a single seizure as the guideline for driving differs between these patients and those classified as having epilepsy, defined as at least two clinically unprovoked, spontaneous seizures.

Patients were classified as remote symptomatic if on entry to MESS the clinician considered their seizure to be caused by a remote disease such as a head injury; meningitis or encephalitis, or both; intracranial surgery; or other disease. Neurological deficit included hemiparesis (weakness on one side of the body) and learning difficulty, while an abnormal EEG was defined as epileptiform activity with focal, as shown in Figure 10, or generalised spikes, as shown in Figure 11, or spike and slow wave activity, as shown in Figure 12 [205, 206].

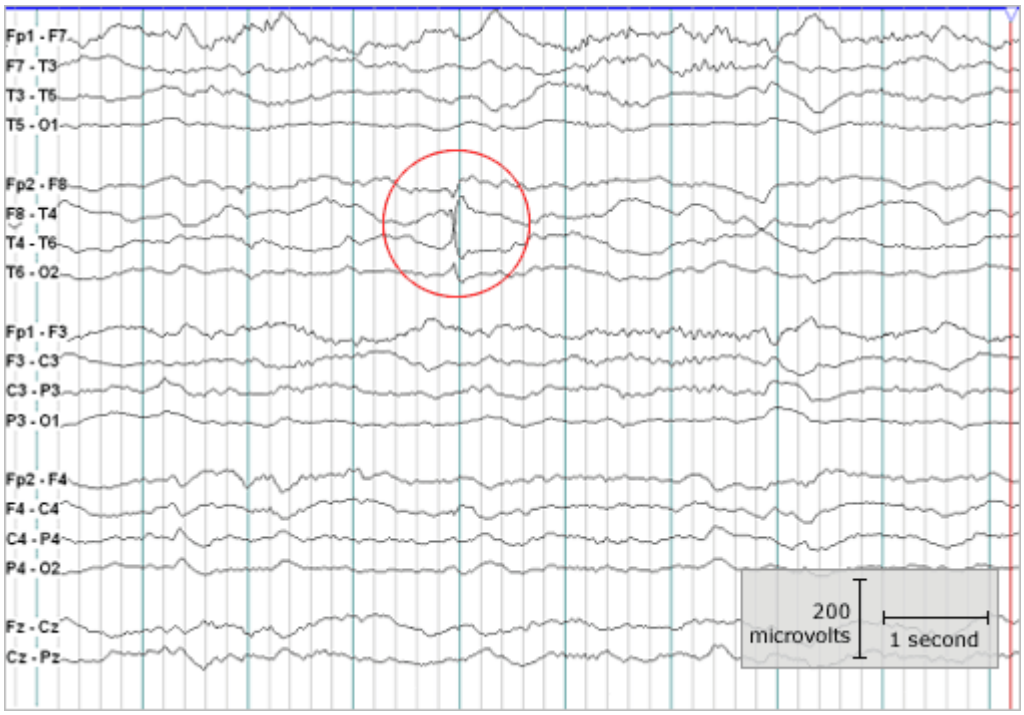


Figure 10: EEG displaying an abnormal discharge called focal spike

Figure 11: EEG displaying an abnormal discharge called a generalised spike and wave

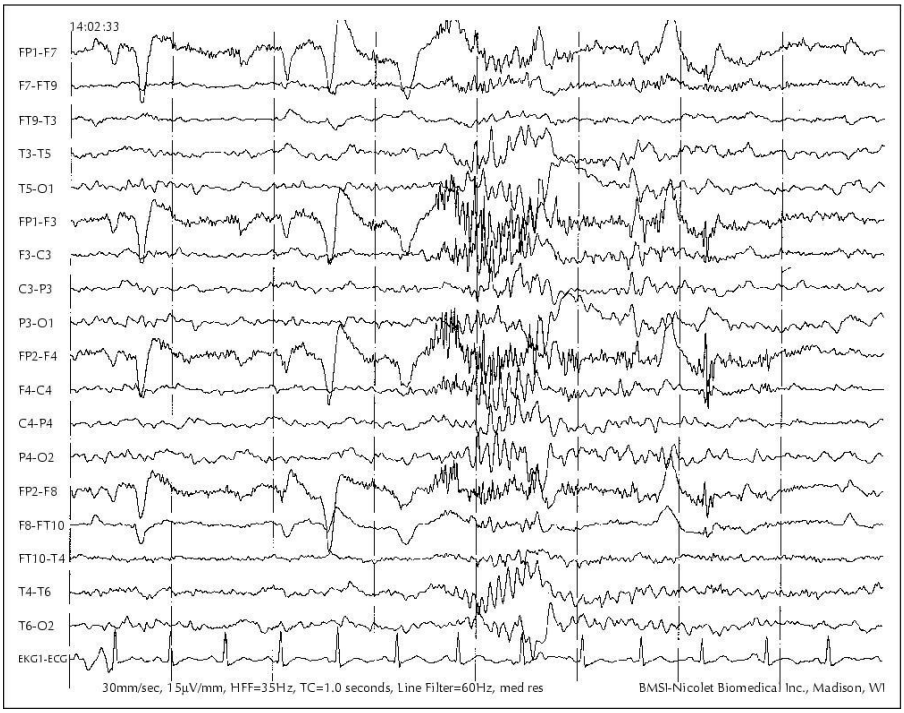


Figure 12: EEG displaying a two-second burst of generalised polyspike-slow wave complex on eye closure in patient with eyelid myoclonia with absenses (idiopathic, myoclonic epileptic syndrome)

5.2.1 Statistical analysis

Analyses were undertaken using *R* version 2.9.2 [182] and SAS version 9 [183].

Although MESS was a randomised trial, in this analysis time to event was measured from the date of the first seizure, occurring before study entry, referred to as the index seizure, and not the date of randomisation. The outcome of interest was the probability of being seizure free for the next 12 months conditional on being seizure free from the index seizure to the time point in question. This outcome is of interest as the Driving and Vehicle Licensing Agency require the probability of being seizure free for the next 12 months to be less than 20% before a person can regain their ordinary, group 1, driving license.

The probability of someone who is seizure free at six months after the index seizure, remaining seizure free throughout months seven to 18 was calculated by dividing the probability of being seizure free for 18 months by the probability of being seizure free for six months. This can be interpreted as the relative probability of being seizure free to 18 months if six months seizure freedom has already been achieved. Conditional probabilities for other time points were calculated similarly [207].

Confidence intervals were estimated around the conditional seizure freedom probabilities. The variance of these rates was estimated as follows. Given $\boldsymbol{m}$ time intervals $[\boldsymbol{0}, \boldsymbol{t_1}], [\boldsymbol{t_1}, \boldsymbol{t_2}], \ldots, [\boldsymbol{t_{m-1}}, \boldsymbol{t_m}]$, let $\boldsymbol{d_k}$ be the number of deaths during interval $\boldsymbol{k}$, $\boldsymbol{r'_k}$ be the adjusted number at risk for interval $\boldsymbol{k}$, $\boldsymbol{S(t)}$ be the probability of surviving past time $\boldsymbol{t}$, and $\boldsymbol{S(t|s)}$ be the conditional probability of surviving past $\boldsymbol{t}$ conditional on surviving past time $\boldsymbol{s}$. Then the variance of the conditional probability of surviving past $\boldsymbol{t_j}$ given survival past $\boldsymbol{t_i}$, with $\boldsymbol{t_i} < \boldsymbol{t_j}$, is given by Equation 10 which is a variation of the usual Greenwood's formula for unconditional survival [207, 208].

Equation 10

$$\text{Var}\left[\widehat{S}(t_j|t_i)\right] = \left[\widehat{S}(t_j|t_i)\right]^2 \sum_{k=i+1}^{j} \frac{d_k}{r'_k(r'_k - d_k)}$$

The 95% confidence intervals were constructed assuming that the conditional rates follow a normal distribution, and are as shown in Equation 11.

Equation 11

$$\left(\widehat{S}(t_j|t_i) \pm \left(1.96 \times \sqrt{\text{Var}\left[\widehat{S}(t_j|t_i)\right]}\right)\right)$$

Variables associated with a higher risk of seizure recurrence in the next 12 months, both univariately and after adjusting for multiple variables, were determined using log-rank tests and Cox proportional hazards modelling methods as described in Chapter 3. A best fitting, parsimonious, multivariable model was produced with variable reduction by Akaike's Information Criterion [178], as also described in Chapter 3. For this model missing values were excluded except in the case of EEG results and CT/MRI results, where a third category was created for missing results, or test not done, as there was too much missing data to ignore and the missingness was informative. The proportional hazards assumption was investigated using Schoenfeld residual plots [173] and incorporation of time-dependent covariate effects [175] as described in Chapter 3.

From the multivariable model the risk of recurrence in the next 12 months for combinations of risk factors was calculated [167] as it is important to determine which groups of patients are at highest risk of a recurrence and similarly, which patient groups are at least risk of a recurrence, not only for driving implications but also for treatment implications. To determine annual recurrence risks for combinations of risk factors the baseline survivor function was estimated from the multivariate model assuming a

piecewise linear assumption. The estimate was subsequently raised to a suitable power calculated from combinations of variable coefficient estimates [167]. For example if the estimated hazard function for the i^{th} patient has the form

$$\hat{h}_i(t) = \exp\{0.5Gender_i + 0.2Relative_i\}$$

where $Gender_i$ equals unity for male patients and $Relative_i$ equals unity for patients with a first degree relative with epilepsy, then the estimated survivor function for the i^{th} individual is given by

$$\hat{S}_i(t) = \left\{\hat{S}_0(t)\right\}^{\exp\{0.5Gender_i + 0.2Relative_i\}}$$

where $\hat{S}_0(t)$ is the baseline survivor function. So, for a female with a first degree relative with epilepsy, the estimated survivor function for this individual is

$$\hat{S}_i(t) = \left\{\hat{S}_0(t)\right\}^{\exp\{0.2\}}.$$

From this, conditional probabilities can be calculated in the manner described above.

An extension of this is to determine the length of time until the conditional seizure recurrence risk falls below the current Driving and Vehicle Licensing Agency threshold of 20% for people with various combinations of risk factors. As in the analysis involving combinations of risk factors, seizure recurrence risks can be calculated at each time point for any combination of variables within the multivariable model. The seizure recurrence risk in the next 12 months for these combinations can also be calculated for each time point making it trivial to determine the time point where the seizure recurrence risk in the next 12 months falls below 20%.

5.3 Results

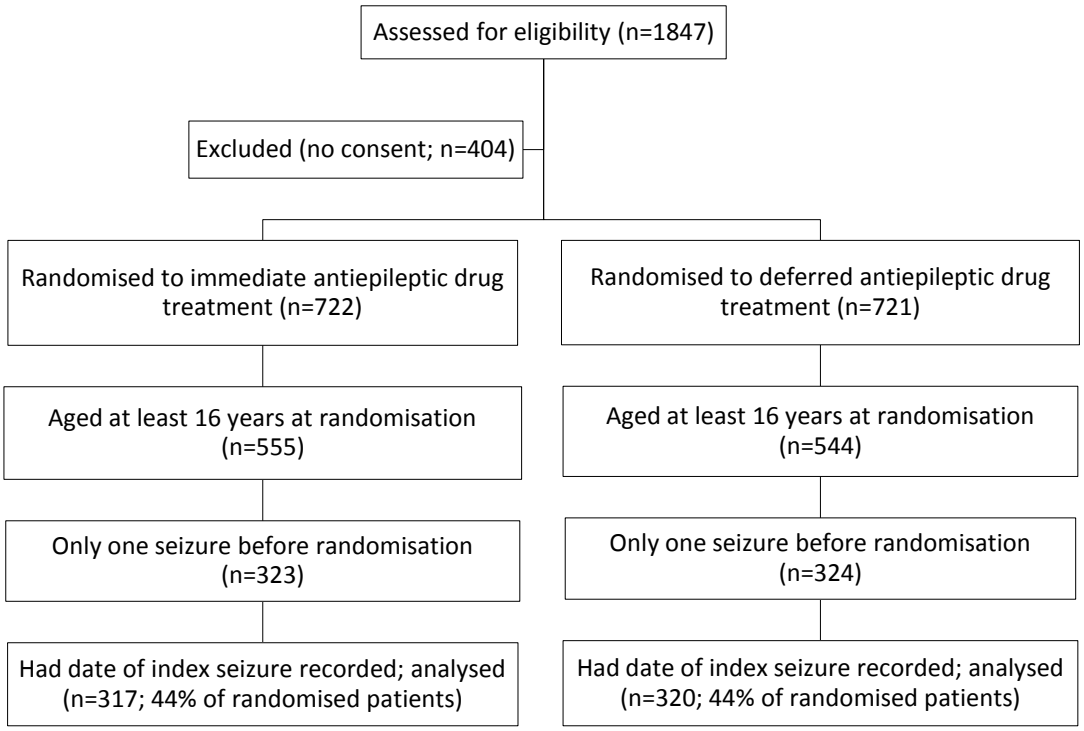


Figure 13: The MESS Study: Flow of patients through the study

Figure 13 shows the flow of patients through the MESS study. Consent was not obtained for 404 of 1847 eligible patients. The remaining 1443 patients were randomised; 722 to immediate antiepileptic drug treatment of whom 317 had a date of index seizure recorded and were consequently analysed, and 721 to deferred antiepileptic drug treatment, of whom 320 were analysed. Table 10 summarises the characteristics of the analysed patients.

65% of the patients were male, the median age was 33, 16% had remote symptomatic seizures, 7% had neurological deficit, 5% had previous acute symptomatic seizures and 10% had a first degree relative with epilepsy. 32% of patients were classified with focal epilepsy and 66% with generalised epilepsy while 18% of patients had seizures only while asleep, 49% had abnormal EEG results and 11% had an abnormal CT/MRI result.

Table 10: The MESS Study: Characteristics of patients; values are numbers (percentages) unless stated

Characteristic	Immediate Treatment (n=317)	Delayed Treatment (n=320)	Total (n=637)
Age at randomisation in years, median (IQR)	32.0 (21.5, 49.5)	33.8 (22.4, 50.6)	32.9 (21.9, 50.3)
Gender			
Male	199 (63)	213 (67)	412 (65)
Female	118 (37)	107 (33)	225 (35)
Aetiology			
Remote symptomatic	55 (17)	49 (15)	104 (16)
Not remote symptomatic	262 (83)	271 (85)	533 (84)
Neurological deficit			
Absent	288 (91)	299 (94)	587 (92)
Present	27 (8)	19 (5)	46 (7)
Missing	2 (1)	2 (1)	4 (1)
Previous acute symptomatic seizures			
Febrile	15 (5)	14 (4)	29 (5)
Other	1 (0)	7 (2)	8 (1)
No	301 (95)	299 (94)	600 (94)
Epilepsy in first degree relative			
Yes	33 (10)	34 (11)	67 (10)
No	284 (90)	286 (89)	570 (90)
Seizures			
Focal	102 (32)	100 (31)	202 (32)
Generalised/not definitely focal	205 (65)	217 (68)	422 (66)
Other	10 (3)	3 (1)	13 (2)
Seizures only while asleep			
Yes	65 (20)	50 (16)	115 (18)
No	252 (80)	269 (84)	521 (82)
Missing	-	1 (0)	1 (0)
EEG results			
Normal	147(46)	136 (43)	283 (44)
Abnormal	149 (47)	160 (50)	309 (49)
Not done/Missing	21 (7)	24 (7)	45 (7)
CT/MRI scan results			
Normal	219 (69)	231 (72)	450 (71)
Abnormal	38 (12)	34 (11)	72 (11)
Not done/Missing	60 (19)	55 (17)	115 (18)

Seizures while asleep are very difficult to determine in people who have experienced few seizures. The DVLA guidelines [202] make special provision for this – a person who has suffered a seizure whilst asleep must refrain from driving for at least one year from the date of seizure, as for patients with seizures while awake. However, if the person has had a seizure whilst asleep more than three years previously, and has not had seizures whilst awake since the original seizure whilst asleep, they may be licensed even though seizures

during sleep may continue to occur. Only if a seizure whilst awake occurs subsequently do the formal epilepsy regulations apply.

Table 11 shows the risk of seizure recurrence in the next 12 months for patients randomised to immediate or delayed treatment at various time points following an index seizure. For patients randomised to the immediate treatment group the unadjusted risk of seizure recurrence in the next 12 months at six months after the index seizure was 14% (95% confidence interval 10% to 18%), significantly below 20%. For the delayed treatment group the estimate was 18% but the confidence interval (13% to 23%) did not exclude a 20% risk of recurrence. At 12 months, however, the risk was reduced to 10% (6% to 15%), significantly below 20%.

Table 11: The MESS Study: Risk of seizure recurrence over 12 months at time points after index seizure: Risk (%, 95% confidence interval)

	Immediate treatment		Delayed treatment	
Time post index seizure (months)	**Number at Risk**	**Risk of seizure in following 12 months**	**Number at Risk**	**Risk of seizure in following 12 months**
6	260	14 (10 to 18)	254	18 (13 to 23)
12	230	7 (4 to 11)	219	10 (6 to 15)
18	211	8 (5 to 12)	197	12 (8 to 17)
24	199	7 (3 to 10)	182	10 (5 to 14)

The univariate and multivariable results can be seen in Table 12. In the univariate model, factors significantly associated with a higher risk of seizure recurrence in the next 12 months were a remote symptomatic seizure (remote symptomatic vs. not remote symptomatic - HR: 1.45, 95% CI: (1.06 to 1.98)), neurological deficit (present vs. absent - HR: 1.80, 95% CI: (1.17 to 2.76)), seizure while asleep (yes vs. no - HR: 1.49, 95% CI: (1.11 to 2.00)), abnormal EEG result (abnormal vs. normal - HR: 1.62, 95% CI: (1.25 to 2.09)), and absence of CT/MRI (not done/missing vs. normal - HR: 1.37, 95% CI: (1.00 to 1.86)). Treatment policy was not a significant factor (immediate vs. delayed – HR: 0.83, 95% CI:

(0.65 to 1.06)) probably owing to lack of power, as the effect estimate for immediate treatment is of a similar magnitude to that in the main analysis of the MESS trial.

Table 12: The MESS Study: Effect estimates from univariate and multivariable analyses

Variable	Comparison	Univariate p-value	Univariate hazard ratio (95% CI)	Multivariable hazard ratio (95% CI)
Age at randomisation	[16,25) [25, 35) [35, 50) ≥ 50	0.71	1.00 0.86 (0.59 to 1.23) 1.00 (0.72 to 1.39) 1.08 (0.79 to 1.48)	N/A
Gender	Male Female	0.42	1.00 0.90 (0.69 to 1.16)	N/A
Aetiology	Not remote symptomatic Remote symptomatic	0.02	1.00 1.45 (1.06 to 1.98)	1.00 1.33 (0.95 to 1.87)
Neurological deficit	Absent Present	0.01	1.00 1.80 (1.17 to 2.76)	N/A
Acute symptomatic seizures	None Febrile seizure(s) Other	0.96	1.00 0.98 (0.55 to 1.75) 1.17 (0.37 to 3.65)	N/A
Epilepsy in first degree relative	No Yes	0.07	1.00 1.37 (0.97 to 1.94)	1.00 1.33 (0.94 to 1.90)
Seizure type	Generalised or not definitely focal Partial Other	0.23	1.00 0.79 (0.61 to 1.04) 0.83 (0.34 to 2.01)	N/A
Seizures only while asleep	No Yes	0.01	1.00 1.49 (1.11 to 2.00)	1.00 1.47 (1.09 to 1.97)
EEG results	Normal Abnormal Not done/Missing	0.00	1.00 1.62 (1.25 to 2.09) 1.48 (0.86 to 2.55)	1.00 1.55 (1.20 to 2.01) 1.29 (0.74 to 2.27)
CT/MRI scan results	Normal Abnormal Not done/Missing	0.08	1.00 1.32 (0.91 to 1.91) 1.37 (1.00 to 1.86)	1.00 1.07 (0.72 to 1.61) 1.29 (0.94 to 1.78)
Treatment policy	Delayed Immediate	0.13	1.00 0.83 (0.65 to 1.06)	1.00 0.82 (0.64 to 1.05)

Table 13 shows the risk estimates for seizure recurrence in the next 12 months at several time points after the index seizure based on the univariate analysis for the variables with significant hazard ratios. At six months the estimate for patients with a neurological deficit was above 20%, although the confidence interval (13% to 49%) did not exclude the possibility of the risk being below 20%. By 12 months the estimate was 20% which is on the threshold set by the DVLA. The risk of recurrence in the next year at six months for

patients with remote symptomatic epilepsy was 20% (10% to 30%) which reduced to 13% by 12 months although this estimate is still not significantly below 20% (4% to 22%).

Table 13: The MESS Study: Risk of seizure recurrence over 12 months at time points after index seizure: univariate model

Variables	Risk of recurrence at time points (95% CI)		
	6 months	12 months	18 months
Cause of seizure:			
Not remote symptomatic	15 (12 to 19)	8 (5 to 11)	11 (7 to 14)
Remote symptomatic	20 (10 to 30)	13 (4 to 22)	9 (1 to 17)
Neurological deficit:			
Absent	15 (12 to 18)	8 (6 to 11)	11 (7 to 14)
Present	31 (13 to 49)	20 (2 to 37)	—
Seizures only while asleep:			
Yes	19 (11 to 28)	10 (3 to 17)	11 (3 to 19)
No	15 (12 to 19)	9 (6 to 12)	10 (6 to 13)
EEG results:			
Normal	13 (9 to 17)	6 (3 to 9)	8 (4 to 12)
Abnormal	19 (14 to 24)	12 (8 to 17)	13 (8 to 18)
Not done or missing*	15 (1 to 29)	10 (0 to 23)	—

*Missing entries arose as a result of insufficient follow-up data or insufficient participant numbers.

As described in Chapter 3 and in section 5.2.1, the multivariable Cox model was developed with variable selection by AIC – the model with the smallest AIC was chosen as the best fitting model – selection starts with the full model and eliminates predictors one at a time, at each step considering whether the criterion will be improved, smaller AIC, by adding back in a variable removed at a previous step [178]. The parsimonious model (Table 12) included variables for aetiology (cause of the seizure), epilepsy in a first degree relative, seizure while asleep, EEG result, CT/MRI result, and treatment policy, which was forced into the model. From the model, seizure while asleep (yes vs. no – HR: 1.47, 95% CI (1.09 to 1.97)), and an abnormal EEG result (not done/missing vs. normal – HR: 1.37, 95% CI (1.00 to 1.86)) were associated with a higher risk of seizure recurrence in the next 12 months, but again treatment policy was not a significant factor (immediate vs. delayed – HR: 0.83, 95% CI (0.65 to 1.06)).

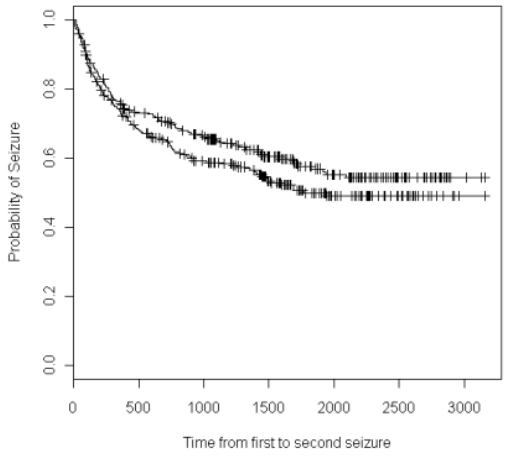


Figure 14: The MESS Study: Survival function over time for treatment policy

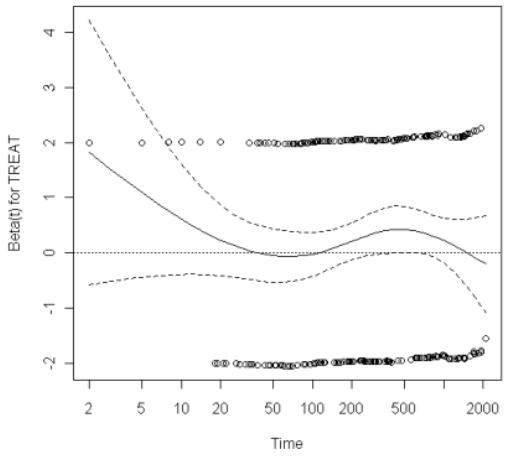


Figure 15: The MESS Study: Schoenfeld residuals for treatment policy

There was no evidence to suggest that the proportional hazards assumption, underlying the Cox model, was invalid – in the case of treatment policy, for example. The graph of the survival function versus the survival time shows almost parallel curves (Figure 14) and the p-value for the test of the Schoenfeld residuals was 0.96 which was not significant so together with Figure 15, there is no reason to doubt the proportional hazards assumption.

The risk of recurrence at 12 months for patients with various combinations of risk factors (i.e. varying characteristics) was estimated from the parsimonious multivariable regression model. As a large number of combinations was possible Table 14 shows selected results for remote and non-remote symptomatic seizure, EEG result, and CT/MRI results assuming that the index seizure was not during sleep and that participants did not have a first degree relative with epilepsy. Appendix A shows the estimates for all possible combinations of factors.

Table 14: The MESS Study: Risk of seizure recurrence in next 12 months estimated from multivariable model at specific seizure free periods. Estimates presented assume seizures not confined to sleep and no first degree relative with epilepsy

Cause of seizure	Test results		Months from index seizure	Immediate treatment		Delayed treatment	
	EEG	CT/MRI		Risk of seizure in next 12 months (%, 95% CI)	Months from index seizure until annual risk falls <20%	Risk of seizure in next 12 months (%, 95% CI)	Months from index seizure until annual risk falls <20%
Not remote symptomatic	Normal	Normal	6	13 (10 to 16)*	1.2	16 (12 to 19)*	3.2
			12	7 (5 to 10)*		9 (6 to 12)*	
Not remote symptomatic	Abnormal	Normal	6	20 (16 to 23)	5.5	23 (20 to 26)	7.0
			12	11 (9 to 14)*		13 (11 to 16)*	
Not remote symptomatic	Normal	Abnormal	6	14 (11 to 17)*	2.1	17 (14 to 20)	3.6
			12	8 (5 to 11)*		10 (7 to 12)*	
Not remote symptomatic	Abnormal	Abnormal	6	21 (18 to 24)	6.1	25 (22 to 28)†	8.1
			12	12 (9 to 15)*		14 (12 to 17)*	
Remote symptomatic	Normal	Normal	6	17 (14 to 20)	3.6	20 (17 to 23)	6.0
			12	10 (7 to 12)*		12 (9 to 14)*	
Remote symptomatic	Abnormal	Normal	6	25 (22 to 28)†	8.7	30 (27 to 32)†	10.7
			12	15 (12 to 17)*		18 (15 to 20)	
Remote symptomatic	Normal	Abnormal	6	18 (15 to 21)	4.8	22 (19 to 25)	6.5
			12	10 (8 to 13)*		13 (10 to 15)*	
Remote symptomatic	Abnormal	Abnormal	6	27 (24 to 30)†	9.2	32 (29 to 34)†	11.3
			12	16 (13 to 18)*		19 (16 to 21)	

*Confidence interval lies completely below 20%. †Confidence interval lies completely above 20%.

Although the unadjusted results suggested that patients who start treatment after a single seizure and are seizure free for six months, on average have a less than 20% risk of recurrence in the next 12 months, our multivariable results indicated that the risk in some subgroups was significantly greater than 20%. This included patients with a remote

symptomatic seizure and an abnormal EEG result whether or not their CT/MRI result was normal. For patients with a non-remote symptomatic seizure, abnormal EEG result and abnormal CT/MRI scan result the estimate was higher than 20% but the lower end of the confidence interval was less than 20% suggesting that some people with that combination of risk factors may have a less than 20% risk of seizure recurrence in the next 12 months.

The estimate from the unadjusted results suggested that patients who had not started AED treatment after a single seizure and who were seizure free for six months had a less than 20% risk of recurrence in the next 12 months. However, the multivariable results indicated that some subgroups had a recurrence risk significantly greater than 20%. This included patients with an abnormal EEG result and abnormal CT/MRI result whether or not their seizure was remote symptomatic. This also included patients with a remote symptomatic seizure and an abnormal EEG result whether or not their CT/MRI result was normal.

For some subgroups the estimate was higher than 20% but the confidence interval included or the lower bound of the confidence interval was less than 20%. For example, patients with a remote symptomatic seizure, an abnormal EEG result and an abnormal CT/MRI result, and patients with a remote symptomatic seizure, a normal EEG result and abnormal CT/MRI result. This again highlights that although the group contains patients with a risk of seizure recurrence in the next 12 months in excess of 20%, some may have a risk that is below this threshold, set by the Driving and Vehicle Licensing Agency.

5.4 Discussion

MESS was a randomised trial comparing the policies of immediate or delayed treatment for patients presenting with a single seizure or early epilepsy. This data can be used to estimate time from a first to second seizure, and for the purposes of informing driving

regulations, the risk of a further seizure following a specific period of seizure freedom can be calculated.

The unadjusted estimates suggest that after a first seizure, the overall risk of a recurrence in the 12 months after a seizure free period of six months is below 20%. The confidence intervals around these estimates suggest that this risk is significantly below 20% for patients who start antiepileptic drug treatment (risk 14%, 95% confidence interval (10% to 18%)), but not for those who do not (risk 18%, 95% confidence interval (13% to 23%)).

The univariate and multivariable analyses allow the identification of patients that are at differing risks of recurrence and hence those that might be at high or low risk of recurrence at specific time points after a first seizure. In the multivariable analyses, patient subgroups were identified whose risk was significantly more than 20% irrespective of the treatment policy. Hence guidance is needed from driving regulators as to how far to individualise risk and whether clinicians should focus on unadjusted results when making recommendations about driving, or on results from univariate or multivariable models.

It could be argued that at the population level the risk, ignoring the confidence interval, of a seizure recurrence is below 20% after a six month seizure free period, thus all patients should be able to regain their driving licence, which would be in keeping with the new European Union standards, accepting that some in the population will have a risk above 20%. Alternatively a more conservative approach could be taken to individualise risk. The multivariable model identified several subgroups that had a risk of recurrence that is significantly greater than 20% after a six month seizure free period, this includes those with a non-remote symptomatic seizure who do not start treatment and have both an abnormal EEG result and an abnormal CT/MRI result and those with a remote symptomatic seizure and an abnormal EEG result, irrespective of the result for CT/MRI and treatment policy. Additionally the risk estimate for several subgroups was greater than 20%, although the

lower confidence limit was below 20%. Similarly, the estimate for some subgroups was below 20% but the upper limit of the confidence interval was above 20%. These results might be used to justify a decision to prevent subgroups with a higher risk of seizure recurrence from returning to driving, which is reflected in the current guidance from the Driving and Vehicle Licensing Agency. Any attempt to utilise these data to inform driving policy could have important implications on resources for health services, as investigations such as EEG may be required that are not otherwise clinically indicated, and patients could choose to take AED treatment, which is not otherwise clinically indicated.

One risk of using a randomised controlled trial such as MESS is the recruitment of a selected population that is not necessarily representative of the general population. This raises questions about the generalisability of results. In MESS, the baseline characteristics of patients not randomised were similar to those randomised, which might argue against a major selection bias when patients made the decision to enter the trial or not. However, patients were only given the opportunity to enter MESS if the clinician was uncertain about the need to start treatment, thus patients entered into MESS might not be representative of the general population of patients presenting with a first seizure. This might have resulted in people at low risk of a seizure recurrence not entering the trial as they did not want to start treatment, and patients at a high risk of a seizure recurrence not entering the trial as they wanted to start treatment. Given the possibility that patients at both low or high risk of a recurrence did not enter MESS it is difficult to predict whether this would have caused an underestimation or overestimation of risk for seizure recurrence in the unadjusted analyses. In the multivariable analyses, this might have resulted in an overestimation of risk for subgroups at low risk of a recurrence and an underestimation of risk for patients at a high risk of a recurrence.

In MESS, most participants (75%) were randomised within three months of their first seizure. Thus another potential selection bias is that some patients had a second seizure whilst waiting to see a specialist. These patients were thus likely to start treatment at that point rather than join the trial therefore MESS might have recruited patients with a lower risk of a seizure recurrence than the general population. It is unlikely that this has biased analyses, however, as patients with a recurrence within six months of their first seizure could not contribute to the risk estimates presented in this chapter. This is because the estimates were conditional on being seizure free for six or 12 months after a first seizure.

In MESS, participants were seen predominantly by neurologists who were experienced at identifying and classifying seizures. However, a further challenge in outpatient based studies of seizures and epilepsy, such as MESS, is that seizures are reported to the clinician by the patient and it is possible that patients under-report the occurrence of seizures although there is no published evidence of this. Validating patient reporting in an outpatient population with infrequent seizures is difficult and to date has not been done. If under-reporting of seizures has occurred, this would result in an underestimate of the risks of recurrence.

Although the possibility and effects of biases from selection and reporting of seizures cannot be ruled out, MESS remains the largest reported study of early epilepsy and single seizures. The analyses presented in this chapter enabled an estimation of risk of seizure recurrence in the next 12 months after periods of seizure freedom and the identification of patients who might be at a high risk of recurrence.

NGPSE [209] is the only other study where risks of recurrence have been estimated after seizure free periods. As explained in sections 2.7 and 4.4, the trial was a population based study that recruited 564 patients with definite new onset seizures. This included 446 with unprovoked seizures, 83 with acute symptomatic seizures, and 35 with alcohol induced

seizures. Overall, 252 patients had a single seizure as their index seizure, whereas the remainder were ascertained after their second or a subsequent seizure. While population based studies might be the ideal design to produce generalisable estimates of seizure recurrence risks, the results of the NGPSE highlight potential problems with ascertainment in epilepsy studies, particularly where the outcome of importance is the risk of recurrence after a first seizure. In their paper assessing recurrence risks after a first seizure, the NGPSE group reported analyses for time from first to second seizure for all patients and for the subgroup of patients where the index seizure was the first seizure. Risk estimates of recurrence were significantly lower for the group where the index seizure was the first seizure: 37% compared with 67% at 12 months [209]. Survival curves were also given for the risk of a recurrence for the whole study population for a seizure free period of six, 12, or 18 months after a first seizure. After six months of being seizure free the risk of a recurrence over the next 12 months was about 35% for the whole study population, substantially higher than the overall estimates in our analyses of MESS: 14% with antiepileptic drug treatment and 18% without treatment.

For the population in NGPSE where the first seizure was the index seizure, which is the comparable population in this analysis, the risk of a recurrence over the next 12 months after a seizure free period of six months was about 18%, similar to the estimate in this analysis of MESS. This highlights the importance of ascertaining patients after their first seizure rather than after a second or subsequent seizure. Multivariable analyses of the NGPSE identified symptomatic seizures, neurological deficit, and no AED treatment as risk factors for a recurrence of seizure after a first seizure, which were also identified in MESS. Data from the NGPSE have not been used to estimate the risks of recurrence in the next 12 months for patients seizure free for various periods for individual risk groups, and with a fairly small sample size validation may not be reliable.

Further research and guidance is required to inform risk based policies, including regulations about returning to driving after a single seizure. To inform recommendations about driving, clinicians and patients require guidance as to whether the focus should be on risk estimates alone or on risks and their associated confidence intervals. If the focus is on the estimate only, the unadjusted estimates presented here suggest that treated and untreated patients are eligible to drive after being seizure free for six months. If the focus is on the risk estimate and its associated confidence interval, a patient who starts treatment will be eligible to drive after being seizure free for six months. For patients who do not start treatment the advice depends on the perspective taken. A conservative approach would allow patients to regain their licence once the upper confidence interval has dropped below 20%. However, this would prevent patients who do not start treatment from returning to driving after being seizure free for six months, and potentially coerce people into taking AED treatment that is not otherwise clinically indicated. A more liberal approach would be to allow people to regain their licence provided that the lower confidence limit is below 20%. This approach could result in those at a high risk of recurrence regaining their licence. An intermediate approach might be to require the risk estimate to be below 20% and also to define the upper limit of the confidence interval that would be acceptable.

The conditional probability survival estimates presented here require that long term follow-up information be available on a large population of patients. The estimates reported here are based on the MESS trial which remains the largest reported study of patients with single seizures and early epilepsy. Even with this large dataset, there were some subgroups for which estimate could not be generated.

Conditional survival estimates appear encouraging for patients in MESS; however different patterns may emerge for other patient groups as these patterns are dependent on the

underlying hazard rates. If the hazard rate is decreasing over time, then the conditional survival will be better than at time 0, as demonstrated here. However, if the hazard rate is increasing over time then the reverse would be apparent. If there are no underlying changes in the hazard rate over time, then the conditional probabilities will be the same as the survival at time 0.

This reanalysis of MESS provides data that can potentially inform decisions about returning to driving for people who have had a first unprovoked seizure. MESS recruited a broad heterogeneous population, which has allowed the investigation of factors that influence seizure recurrence as presented in this chapter. Results should therefore be relevant to the general population, but selection bias may be at play and some caution is required in interpreting results. Ideally, external validation of the models presented here is required and work is ongoing to establishing a consortium to attempt to validate and refine the model presented here. There are additional challenges to externally validating prognostic models, details of which can be found in Chapter 8.

The MESS data allow an estimation of seizure recurrence risk conditional on having been seizure free for six or 12 months. Further guidance is now required from driving authorities as to how these data should be interpreted. In particular, guidance is needed as to whether a population approach should be taken, with a focus on unadjusted results or whether attempts should be made to individualise risk. Guidance is also required as to whether the focus should be on risk estimates only or on the confidence interval as well. If the confidence interval is to be included, guidance is needed as to whether the conservative, liberal, or intermediate approach should be taken.

Chapter 6: Implications for Drivers who Withdraw AED Treatment

6.1 Introduction

In this chapter a further analysis of data from the Antiepileptic Drug Withdrawal Study is presented. The chapter begins with a background to the U.K. driving regulations for people who withdraw treatment following a period of remission from seizures (section 6.1). The Medical Research Council Antiepileptic Drug Withdrawal Study is then introduced (section 6.2.1). A systematic review of AED withdrawal studies is also presented (section 6.2.2).

The risk of seizure recurrence is presented for people who have had their AED treatment withdrawn (section 6.3.2) and for people with a seizure recurrence during or following AED withdrawal who then restart treatment (section 6.3.3).

The work presented in this chapter has been published in the Journal of Neurology, Neurosurgery and Psychiatry, and I am first author. The DVLA are currently considering changing their guidelines for time off driving following treatment withdrawal after a period of remission as a result of this work.

Around 60% to 70% of people with epilepsy will enter a remission from seizures on AED treatment [69] and many of these will become eligible to drive. It is common for individuals who have had two or more years without seizures to consider AED withdrawal. However, a decision to withdraw AED treatment will increase the probability of a seizure occurring. The seizure recurrence risks associated with the policies of continued treatment or withdrawal were examined in the Medical Research Council Antiepileptic Drug Withdrawal Study [132] which found a two year seizure recurrence risk of 22% with continued antiepileptic drug treatment and 41% with withdrawal of treatment.

The Driving and Vehicle Licensing Agency currently advises that patients undergoing antiepileptic drug withdrawal should not drive while the drug is being withdrawn and for a further six months after the last dose has been taken. If an epileptic seizure does occur and treatment is reinstated, a further one year period of seizure freedom [202] is required, as per the Road Traffic Act [210]. Following calls for harmonisation across the European Union [199, 200, 202], member states have been engaged in a process of identifying minimum standards to drive which are now being considered by the United Kingdom and other European Union states. The new European Union standard states that seizures occurring during physician advised change or withdrawal of medication requires three months without driving if the previously effective treatment is reinstated [199]. At present there is little published evidence to support either the United Kingdom or European Union standards. There is therefore a need for reliable published data to inform recommendations made about people with epilepsy who have a seizure recurrence when withdrawing antiepileptic drug treatment with regards to driving.

6.2 Methods

Analyses were undertaken using *R* version 2.9.2 [182].

6.2.1 Antiepileptic Drug Withdrawal Study

The Medical Research Council Antiepileptic Drug Withdrawal Study (MRC AEDWS) was a prospective multicentre randomised study of continued AED treatment versus slow withdrawal. It randomised 1021 patients who had been free from seizures for at least two years. Recruitment began in February 1984 and was completed in June 1988 with follow-up continuing until June 1989, where possible. Patients were eligible for entry in the Antiepileptic Drug Withdrawal Study provided they had a history of two or more definite unprovoked seizures, had been free of all seizures for at least two years, and were taking

antiepileptic drugs. Those who had a progressive neurological disorder or other conditions likely to reduce follow-up below two years were excluded.

Demographic and clinical data on all eligible patients were recorded at entry: presence of developmental delay, neurological handicap or neurological disorder; psychiatric history; results of most abnormal EEG based on review of historical reports when available; dates of first, second and most recent seizures and their classification from clinician's review of historical data; details of AED therapy including starting date, drug and dose at remission of seizures and entry; diurnal pattern of seizures and family history of epilepsy; and previous unsuccessful attempts at AED withdrawal.

Consenting patients were randomised to continued treatment or to slow withdrawal of AEDs by use of the method of minimisation. Allocation to the two treatment policies was balanced by centre and by the seizure-free period before registration. Patients randomised to slow withdrawal had therapy withdrawn according to guidelines suggested by the steering committee with decrements every four weeks. Physicians were encouraged to make smaller decrements in patients receiving small doses of drugs with the aim being to extend withdrawal to a minimum of six months. Patients taking more than one AED had the drugs withdrawn sequentially in an order determined by the clinician. Follow-up was at three, six and 12 months and yearly thereafter from the date of randomisation. At each follow-up appointment information was collected about AEDs, including dose and adverse events, and dates of any seizures together with a clinical description and whether or not AEDs were restarted. Patients with seizures during follow-up could have their antiepileptic drug regimen modified by their clinician. A Consort type diagram of the patient flow through the trial can be seen in Figure 16.

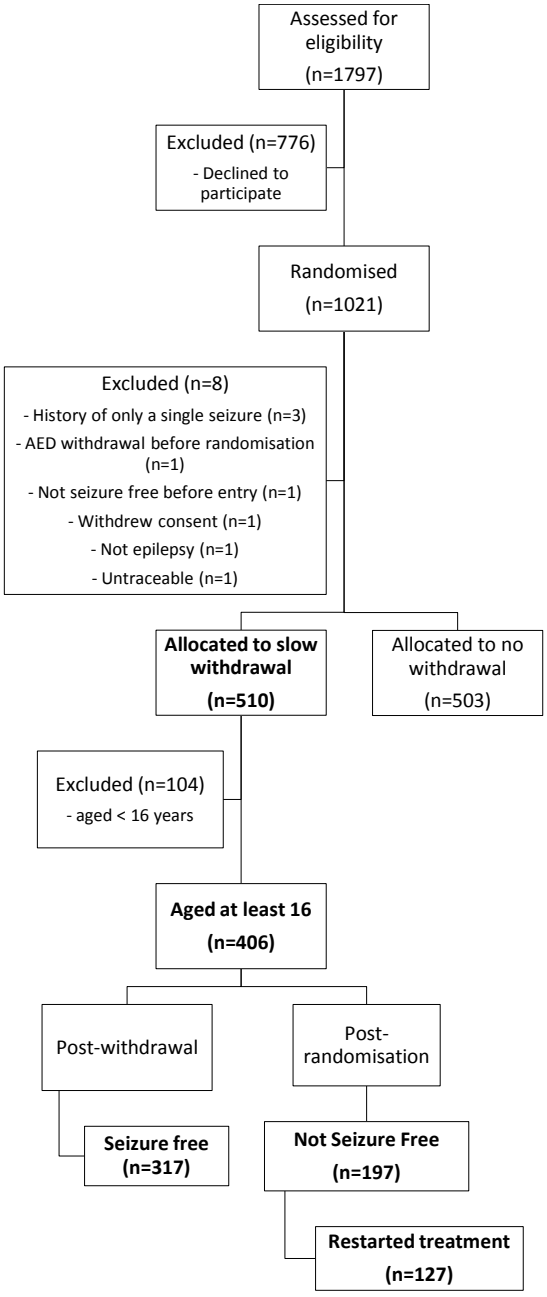


Figure 16: MRC AEDWS: Consort Diagram; Bold figures relate to patients used for these analyses

In this chapter two outcomes were assessed: the risk of a seizure recurrence in the next 12 months at a number of time points following AED withdrawal; the risk of a seizure recurrence in the next 12 months at a number of time points following AED withdrawal, a seizure recurrence and reinstatement of AED treatment. For the purpose of this analysis, withdrawal is assumed to have taken exactly six months for each patient as the date of last dose was not recorded. For the risk of recurrence following AED withdrawal, if a person

had a seizure whilst withdrawing they were excluded from the analysis. In the case of risk following seizure recurrence and reinstatement of AED treatment, patients were removed from the analysis if they: did not restart medication following a seizure; if they did not attempt to withdraw treatment despite being in the withdrawal arm of the study; if they were withdrawing treatment, had a seizure and then continued to withdraw; if they were withdrawing treatment, had a seizure and did not alter their dose; or if they restarted treatment despite not having a seizure.

This analysis was restricted to patients randomised to slow withdrawal of AED treatment and aged at least 16 years. The outcome of interest is the probability of being seizure free for the next 12 months given that a person is seizure free until the time point of interest post-withdrawal. Conditional probabilities were calculated as per Chapter 4 and again confidence intervals for these estimates were calculated utilising a revised version of Greenwood's formula [207, 208, 211]. Univariate analyses used log-rank tests and Cox proportional hazards modelling methods [31].

6.2.2 Systematic Review

The aim of the systematic review was to identify clinical trials that estimated the risk of seizure recurrence following antiepileptic drug withdrawal in patients of driving age who had been free from seizures for at least 24 months whilst on AED treatment. Randomised controlled trials and prospective observational studies, in which at least 50 patients initiated AED withdrawal, were included. Studies had to include patients with epilepsy treated with AED therapy, aged 16 years or over, who had been in remission for at least two years. The search strategy used in MEDLINE can be found in Appendix B.

Trials were assessed for inclusion by two investigators and one investigator extracted the data. The primary outcome measure was time to seizure recurrence while a secondary analysis considered prognostic factors for seizure recurrence. The risk of seizure

recurrence in the next 12 months following AED withdrawal was calculated, where possible, for each trial together with the risk following recurrence and reinstatement of AED treatment. This was done using survival estimates obtained from the Kaplan-Meier curve.

Risk of bias was assessed using standard Cochrane methodology in the case of randomised controlled trials. In the case of observational studies, the Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses was used [212].

6.2.2.1 Cochrane Collaboration Tool

Many tools have been proposed for assessing the quality of studies. Most tools are scales, involving scoring various components of quality and combining them to give a summary score. Alternatively a checklist can be used, in which specific questions are asked [213]. The Cochrane Collaboration's recommended tool for assessing risk of bias [214] is neither a scale nor a checklist. Instead, it is a domain-based evaluation, in which critical assessments are made separately for different domains. It was developed between 2005 and 2007 by a working group of methodologists, editors and review authors. It is a two-part tool, addressing seven specific areas, namely sequence generation, allocation concealment, blinding of participants and personnel, blinding of outcome assessment, incomplete outcome data, selective outcome reporting and 'other issues'.

Each domain in the tool includes one or more specific entries in a 'Risk of bias' table. For each entry, the first part of the tool describes what was reported to have happened in the study, while the second part of the tool assigns a judgement relating to the risk of bias for that entry; either low, high or unclear risk of bias.

6.2.2.2 Newcastle-Ottawa Scale

Nonrandomised studies, including case-control and cohort studies, can be challenging to implement and conduct, however, assessment of the quality of such studies is essential for a proper understanding of nonrandomised studies. The Newcastle-Ottawa Scale (NOS) [212] is an ongoing collaboration between the Universities of Newcastle, Australia and Ottawa, Canada. It was developed to assess the quality of nonrandomised studies. A 'star system' has been developed in which a study is judged on three broad perspectives: the selection of the study groups; the comparability of the groups; and the ascertainment of either the exposure or outcome of interest for case-control or cohort studies respectively. Stars are awarded for high quality choices.

6.3 Results

6.3.1 Antiepileptic Drug Withdrawal Study

Table 15 summarises the patient demographics for the populations under analysis. 49% of all relevant patients, those randomised to slow withdrawal and aged 16 or over, were male, 16% had delayed development, 8% had neurological deficit, 18% had a first degree relative with epilepsy, 9% had a history of febrile seizures and 78% had abnormal EEG results. The median age at randomisation was 34 with mean duration of epilepsy at randomisation 9 years, median duration of antiepileptic drug treatment at randomisation 8 years and median seizure free period at randomisation 3 years.

Table 15: MRC AEDWS: Characteristics of patients undergoing antiepileptic drug withdrawal

Characteristic, n (%)	Slow Withdrawal Group 16 years and older (n=406)	Seizure Free Following AED Withdrawal (n=317)	Seizure Recurrence Following Randomisation (n=197)
Gender			
Male	198 (49)	154 (49)	96 (49)
Female	208 (51)	163 (51)	101 (51)
Development			
Normal	343 (84)	265 (84)	166 (84)
Delayed	63 (16)	52 (16)	31 (16)
Neurological Deficit			
Absent	372 (92)	288 (91)	182 (92)
Present	34 (8)	29 (9)	15 (8)
First Degree Relative with Epilepsy			
No	338 (83)	270 (85)	160 (82)
Yes	66 (16)	45 (14)	36 (18)
Missing	2 (1)	2 (1)	1 (0)
History of Febrile Seizures			
No	377 (93)	298 (94)	179 (91)
Yes	27 (6)	17 (5)	17 (9)
Missing	2 (1)	2 (1)	1 (0)
EEG Results			
Normal	71 (17)	58 (18)	34 (22)
Abnormal	239 (59)	178 (56)	119 (78)
Missing	96 (24)	81 (26)	-
Age at Randomisation (years)			
Median (IQR)	34 (24 to 49)	34 (24 to 50)	34 (24 to 47)
Age at First Seizure (years)			
Median (IQR)	18 (13 to 30)	19 (13 to 30)	16 (13 to 27)
Duration of Epilepsy at Randomisation (years)			
Median (IQR)	9 (5 to 19)	9 (5 to 19)	9 (4 to 20)
Duration of AED Treatment at Randomisation (years)			
Median (IQR)	6 (4 to 13)	6 (3 to 13)	8 (4 to 15)
Period Seizure Free at Randomisation (years)			
Median (IQR)	4 (3 to 6)	4 (3 to 6)	3 (2 to 5)
Seizure Types (n, % answering 'yes')			
Simple Partial	402 (99)	313 (99)	197 (100)
Complex Partial	377 (93)	293 (92)	182 (92)
Simple Partial with 2° tonic-clonic	377 (93)	293 (92)	182 (92)
Complex Partial with 2° tonic-clonic	316 (78)	254 (80)	155 (79)
Tonic-Clonic	158 (39)	120 (38)	75 (38)
Myoclonic	389 (96)	309 (98)	183 (93)
Simple Absence	388 (96)	305 (96)	185 (94)
Complex Absence	394 (97)	309 (98)	189 (96)

6.3.2 Recurrence Risk Following Antiepileptic Drug Withdrawal

Figure 17 shows time to first seizure from the date of complete treatment withdrawal for the 317 patients who remained seizure free during AED withdrawal.

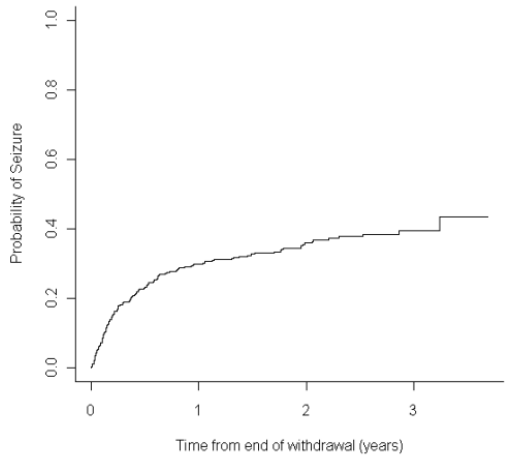


Figure 17: MRC AEDWS: Time to first seizure post-withdrawal

Table 16 shows the probability of having a seizure in the next 12 months given that patients were seizure free at various time points post-withdrawal. At three months following treatment withdrawal the estimated 12 month recurrence risk was 15% with 95% confidence interval (10% to 19%), significantly below the 20% risk bar set by the Driving and Vehicle Licensing Agency. At six months the estimate was 9% with 95% confidence interval (5% to 13%).

Table 16: MRC AEDWS: Probability of having a seizure in next 12 months at time points following antiepileptic drug withdrawal for patients who were seizure free during drug withdrawal

Time post-withdrawal (months)	Number at Risk	Number of Events from baseline	Probability of seizure in next 12-months given seizure free at stated time post-withdrawal (% (95% CI))
0	317	0	30 (25 to 35)
3	257	64	15 (10 to 19)
6	235	84	9 (5 to 13)
12	221	96	9 (5 to 14)
18	188	105	8 (3 to 13)
24	126	113	5(0 to 11)

Univariate regression analyses indicated that the following variables were significantly associated with seizure recurrence risk: duration of antiepileptic drug treatment at randomisation (≥20 years vs. 3 years - HR: 2.50, 95% CI: (1.05 to 5.98)) and myoclonic seizures (present vs. absent - HR: 2.47, 95% CI: (1.01 to 6.06)).

6.3.3 Risk Following Recurrence and Recommencing Treatment

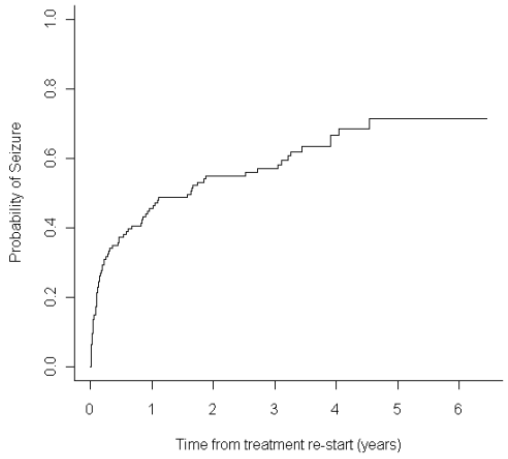


Figure 18: MRC AEDWS: Time to first seizure from treatment restart

Table 17: MRC AEDWS: Probability of having a seizure in next 12 months at time points after reinstating antiepileptic drug treatment for patients with a seizure recurrence following randomisation

Time from treatment restart (months)	Number at Risk	Number of events from baseline	Probability of seizure in next 12-months given seizure free at stated time post-withdrawal (% (95% CI))
0	127	0	45 (36 to 53)
3	87	39	26 (17 to 35)
6	79	47	18 (10 to 27)
12	68	57	17 (8 to 27)
18	62	61	12 (4 to 20)
24	52	68	5 (0 to 11)

Figure 18 shows time from restarting AED treatment to a further seizure recurrence for patients who withdrew treatment, had a further seizure and then reinstated treatment. At three months after reinstating treatment the estimated seizure recurrence risk in the next

12 months was 26% with associated 95% confidence interval (17% to 35%), while the estimate at six months was 18% with 95% confidence interval (10% to 27%) as shown in Table 17.

Univariate regression analyses indicated that the following were significantly associated with a seizure recurrence: history of psychiatric disease (present vs. absent - HR: 1.86, 95% Confidence CI: (1.07 to 3.23)), history of status epilepticus (present vs. absent - HR: 2.47, 95% CI: (1.07 to 5.72)) and complex partial seizures with secondary tonic-clonic seizures (present vs. absent - HR: 2.14, 95% CI: (1.29 to 3.55)).

6.3.4 Systematic Review

In addition to the Antiepileptic Drug Withdrawal study [132], five studies were found that met all the inclusion criteria (Lossius [215], Aktekin [216], Specchio [121], Callaghan [217] and Overweg [218]). Lossius, Antiepileptic Drug Withdrawal study and Callaghan were randomised controlled trials. Lossius and Antiepileptic Drug Withdrawal study compared withdrawal to non-withdrawal while, in the case of Callaghan, the cohort had been recruited into a trial that compared treatment efficacy. Patients were then followed up to the point of AED withdrawal and beyond. Aktekin, Specchio and Overweg were prospective observational studies. Characteristics of the included studies can be seen in Table 18.

Table 18: Studies Meeting the Inclusion Criteria - clinical trials that estimated the risk of seizure recurrence following antiepileptic drug withdrawal in patients of driving age who had been free from seizures for at least 24 months whilst on AED treatment

Study	Design	Number of participants	Setting	Age range	Duration AED treatment	Duration of Remission	Withdrawal policy
Lossius [215]	Randomised Controlled Trial	168	Hospital based neurological outpatient clinics	18-67	Unknown	2+ years	AED dose reduction by 20% of initial dose in 1st 6 weeks and 20% of the initial dose every 2nd week
Aktekin [216]	Observational	79	Hospital based neurological outpatient clinics	Mean ± Standard Deviation Seizure free: 38.43±17.35 Relapsed: 39.39±13.68	~ 7 years	4+ years	1/6th every 2 months
Specchio [121]	Observational	330	Hospital based neurological outpatient clinics	< 15: (n=38) 15-34: (158) 35-54: (22) > 54: (7)	1+ years	2+ years	25% of the maintenance dose at entry was subtracted every 3 months or longer
MRC AED Withdrawal Study Group [132]	Randomised Controlled Trial	1013	Hospital based neurological outpatient clinics	Median (IQR) 27 (17, 42)	Mean ~ 7 years	2+ years	Decrements every 4 weeks
Callaghan [217]	Longitudinal Follow-up of Randomised Controlled Trial	92	Unknown	Mean: 24	Unknown	2+ years	Dosage decreased by 1 unit in intervals of 2 weeks
Overweg [218]	Observational	62	2 Specialised Epilepsy Clinics	18-60	Unknown	3+ years	Dosage decreased by 1 unit in intervals of 2 weeks

Nine other studies (Ohta [219], Shinnar [220], Aldenkamp [221], Mastropaolo [90], Ehrhardt [222], Matricardi [91], Arts [94], Bouma [92] and Todt [93]) were found which might meet the inclusion criteria listed but further clarification from the authors together with individual patient data is required. For example, some studies in children were long-term studies and consequently patients may have been at least 16 years of age with at least two year remission. All six included studies reported seizure relapse or recurrence, and prognostic factors for seizure recurrence were also reported. For additional details see Table 19.

There were potentially important differences among the cohorts that could have influenced the estimated risks of seizure recurrences. Two studies recruited only adults while four had no age limits. Specchio recruited patients of all ages, 7% of which were aged less than 15 and since data were not available for age subgroups calculations in this review were based on the whole study population. The age of patients in the Callaghan study is unclear; the mean age is given as 24 but a standard error is not provided. The patients reported in Callaghan were a subset of 181 patients previously reported in a separate publication by Callaghan [223]. Those 181 patients ranged in age from four to 75, however, no further information was provided therefore, once again, the whole study population was used for the analysis. Four studies required patients to be in a remission of two or more years, one study three or more years and another study four or more years.

Most of the studies had a low risk of bias, as shown in Table 20.

Table 19: Summary of results including predictors for seizure freedom for studies looking at AED withdrawal in patients of driving age in remission for at least 24 months on treatment

Study	Results	Predictors for seizure freedom
Lossius	Seizure relapse within 12 months: 7% (non withdrawers), 15% (withdrawers)	• Prior use of CBZ • Neurological examination results
Aktekin	57% relapsed (although 42.8% did not relapse during the discontinuation programme) For patients discontinuing treatment, the probability of relapse was 21.4% during the tapering period, 28.6% at 1 month, 14.3% at 3 months, 3.6% at 6 months, 7.1% at 12 months, 17.8% at 24 months and 7.1% at 36 months	• Age at onset of epilepsy • Duration of active epilepsy
Specchio	In patients continuing treatment, the probability of remission was 95% at 6 months, 91% at 12 months, 82% at 24 months, 80% at 36 months and 68% at 60 months The corresponding values for patients discontinuing treatment were 88, 74, 57, 51 and 48%	• Duration of active disease • Number of years of remission at study entry • Psychiatric examination results
AED Withdrawal Study Group	35% of the randomised patients had a seizure(s) during follow-up In the slow withdrawal group, 48% seizures occurred in patients during AED reduction and 52% after drug withdrawal In patients continuing treatment, the seizure recurrence per year was 14% up to 6 months, 11% up to 12 months, 13% to 18 months, 11% to 24 months, 9% to 3 years and 4% to 4 years The corresponding values for patients discontinuing treatment were 40, 37, 16, 14, 6 and 3%	• Taking two or more AEDs at randomisation • History of either primary or secondarily generalised tonic-clonic seizure • Period free of seizures at randomisation
Callaghan	Relapse rate 33.7% Average duration of remission was 26 months (range 1-62)	• Number of seizures a patient had before control was achieved • Number of drugs tried as single-drug therapy • Type of treatment withdrawal
Overweg	66% patients who completed the study relapsed. 11% had relapsed by, or prior to, 50% reduction of AEDs, 45% up to the point of complete withdrawal of medication and 10 more in the following year. Three further late recurrences brought the total relapse rate to 66%.	• Age at withdrawal • Age at onset • Age at last seizure • Duration of Epilepsy

Table 20: Quality Assessment of studies looking at AED withdrawal in patients of driving age in remission for at least 24 months on treatment

Study	Item	Author's Judgement	Description
Lossius	Sequence Generation?	"Randomised in blocks of 10"	Low Risk
	Allocation Sequence Concealment?	"Receive blindly active medication or placebo in pre-packed dispensers"	Low Risk
	Blinding?	Double Blind	Low Risk
Aktekin	Representative of exposed cohort?	Representative of the average epileptic in remission in the community	Low Risk
	Assessment of outcome	Self report (but of obvious outcome)	Low Risk
	Adequacy of follow-up	8 years follow-up and only four drop-outs	Low Risk
Specchio	Representative of exposed cohort?	"Patients discontinuing AEDs were younger, and had a poorer education, a shorter disease duration or a normal EEG at entry, less frequent seizure relapse after starting treatment, a shorter duration of active disease and more years of remission"	High Risk
	Selection of the non exposed cohort	Drawn from the same community as the exposed cohort	Low Risk
	Assessment of outcome	Self report (but of obvious outcome)	Low Risk
	Adequacy of follow-up	16 years follow-up and only 9 drop outs (5 from discontinuing group, 4 from continuing group)	Low Risk
AED Withdrawal Study Group	Sequence Generation?	"Randomisation by minimisation"	Low Risk
	Allocation Sequence Concealment?	Telephone Service	Low Risk
	Blinding?	No blinding	High Risk
Callaghan	Sequence Generation?	Randomisation was based on two Latin squares	Low Risk
	Allocation Sequence Concealment?	The drug was identified from the randomisation list by a person not otherwise involved in the study	Low Risk
	Blinding?	The drug was identified from the randomisation list by a person not otherwise involved in the study	Low Risk
Overweg	Representative of exposed cohort?	"Heavily weighted with subjects presenting special problems, including resistance to therapy. 227 patients were nevertheless found who had been seizure-free for 3 years"	Low Risk
	Assessment of outcome	Self report (but of obvious outcome)	Low Risk
	Adequacy of follow-up	All drop-outs accounted for. Follow-up continued beyond 2 years in some cases	Low Risk

There were also important differences in the policy of AED withdrawal and lack of detail as to precisely when AED withdrawal was completed. In the Lossius study, withdrawal commenced one month after recruitment into the study. In the context of this analysis, the mean period of drug withdrawal was 18 weeks which is approximately 4.5 months. This is not a reported time point and very difficult to read from the Kaplan-Meier curve. Therefore, the mean period was rounded up to six months, a reported time point. In the study reported by Specchio withdrawal is assumed to have taken exactly 12 months. In Callaghan exact time to withdrawal is not known however the Kaplan-Meier curve shows the probability of remaining seizure free after the withdrawal of treatment.

For estimation of the risk of a seizure recurrence in the next 12 months following AED withdrawal, Lossius, Specchio, Antiepileptic Drug Withdrawal Study Group and Callaghan provide Kaplan-Meier curves with numbers at risk. Aktekin provides the number of patients who had relapsed at various time points and Overweg by the end of the study. Table 21 shows estimates for seizure recurrence in the next 12 months at zero, three, six, 12 and 18 months after antiepileptic drug withdrawal for Lossius, Specchio, Antiepileptic Drug Withdrawal Study and Callaghan. It was not possible to calculate confidence intervals for Lossius, Specchio and Callaghan as the number of events or number at risk at time points were not given. Due to differences in study design and lack of sufficient data in published literature it was not possible to undertake a meta-analysis of results.

Immediately following AED withdrawal the risk estimate of a seizure in the next 12 months was highest in Antiepileptic Drug Withdrawal Study 30% (95% confidence interval: 25% to 35%), Callaghan gives a similar estimate at 28%, while the estimate for Lossius was 12% and for Specchio, 19%. At three, six and 12 months following AED withdrawal the estimate from the Antiepileptic Drug Withdrawal study is intermediate with Lossius giving a lower

risk, Specchio a higher risk and Callaghan giving both lower and higher risks of recurrence in the next 12 months.

For patients who had a seizure recurrence following AED withdrawal and reinstated treatment, none of the included papers provide data that enabled a calculation of the risk of a seizure recurrence.

Table 21: Probability of having a seizure in the next 12 months at time points following antiepileptic drug withdrawal

Time post-withdrawal (months)	Study	Number at Risk	Number of Events	Percentage probability of seizure in next 12-months given seizure free at stated time post-withdrawal (95% CI)
0	Lossius Specchio MRC AEDW Callaghan	65 164 317 92	Unknown 38 1 Unknown	12 19 30 (25 to 35) 28
3	Lossius Specchio MRC AEDW Callaghan	62 Unknown 257 Unknown	Unknown Unknown 64 Unknown	7 19 15 (10 to 19) 12
6	Lossius Specchio MRC AEDW Callaghan	60 Unknown 235 70	Unknown Unknown 84 Unknown	6 16 9 (5 to 13) 12
12	Lossius Specchio MRC AEDW Callaghan	53 114 221 58	Unknown 50 96 Unknown	4 12 9 (5 to 14) 3
18	Lossius Specchio MRC AEDW Callaghan	53 Unknown 188 Unknown	Unknown Unknown 105 Unknown	6 2 8 (3 to 13) 10

6.4 Discussion

The United Kingdom Driving and Vehicle Licensing Agency take a risk based approach to guidance with an individual usually being allowed to regain their ordinary driving license once their risk of a seizure in the next 12 months falls below 20% [202]. This re-analysis of

data from the Antiepileptic Drug Withdrawal study indicated that for patients who remain seizure free during AED withdrawal, the risk of a seizure recurrence in the next 12 months immediately following drug withdrawal was 30% (25% to 35%). Once the patient has been seizure free for a further three months the risk is significantly less than 20%, 15% (10% to 19%), and similarly at later time points. For patients who have a seizure recurrence during or following AED withdrawal and reinstate treatment, their risk of a seizure recurrence in the next 12 months immediately after restarting treatment was 45% (36% to 53%) and following a further three months seizure freedom the risk was 26% (17% to 35%). Once the patient has been seizure free for six months after restarting treatment the risk estimate was below 20% at 18% (10% to 27%) but not significantly so.

While the Antiepileptic Drug Withdrawal study remains the largest study to assess outcomes following AED withdrawal, it does have a number of limitations when used to estimate absolute risk of seizure recurrences. The study is over 20 years old, and treatment polices might now differ, particularly given the ever increasing number of AEDs available. Thus there might be important differences in the population of patients offered AED withdrawal now compared to 20 years ago. The ideal study design would be a population based cohort study, but the Antiepileptic Drug Withdrawal study was a randomized controlled trial designed to compare the policies of continued, or withdrawal of, AED treatment, and analyses reported here have used only data for those randomised to treatment withdrawal. Patients entered the study if they were prepared to be randomised to either policy and thus selection bias might exist; patients recruited may not be representative of the general population as patients at high risk might not have been offered entry to the study while patients at low risk might have chosen drug withdrawal rather than to join the trial. Any observational study would suffer from similar selection biases as AED withdrawal cannot be imposed upon an unselected population. A further problem with the Antiepileptic Drug Withdrawal Study, and any other outpatient based

study of epilepsy, is that seizures are patient reported events and cannot be validated. There is thus a risk that patients might under report events, and if this occurred would result in an underestimation of the seizure recurrence risk.

The systematic review identified five studies in addition to the Antiepileptic Drug Withdrawal study, three observational studies and two randomised controlled trials. There were potentially important differences among these studies with respect to patient populations and withdrawal policy. Only three of these studies provided data that allowed the calculation of seizure recurrence risks following AED withdrawal and at subsequent time points. Lossius reported consistently lower estimates of a seizure recurrence than the Antiepileptic Drug Withdrawal Study even though duration of drug withdrawal was significantly longer. Specchio reported consistently higher estimates while Callaghan's results were mixed. Immediately following AED withdrawal the risk of a seizure in the next 12 months was estimated at 19% in Specchio, 12% in Lossius and 28% in Callaghan compared to an estimate of 30% (25%, 35%) from the Antiepileptic Drug Withdrawal Study. Following a further three months seizure freedom after AED withdrawal all four studies estimate the risk of a seizure in the next 12 months at below 20%. Similarly risk estimates at subsequent time points are below 20%.

Confidence intervals for the risk estimates in the published papers were unavailable. Parmar *et al* [224] describe a method to estimate the effective number alive at the start of a time interval, the effective number censored in a time interval, the effective number at risk in a time interval and the effective number of deaths in a time interval. The standard adaptation of Greenwood's formula for conditional probabilities, used to estimate the confidence intervals in this chapter, requires the number at risk and the number of events at each (survival) time point. Therefore, to obtain confidence intervals for the estimates using data from Lossius, Specchio and Callaghan, it would be necessary to calculate the

estimates of number of events and number at risk at each possible survival time point between estimates (six and 18 months say). Therefore, although it is possible to calculate confidence intervals from the published data, they have not been calculated here which means we are reliant on only point estimates. Ideally, individual patient data would be obtained for the studies included in the systematic review and therefore a meta-analysis could be performed enabling more reliable estimates concerning risk of seizure recurrence to be obtained.

Data from the Antiepileptic Drug Withdrawal study indicated that patients who withdrew AED treatment and then had a further six months of seizure freedom had a risk of a seizure recurrence in the next 12 months significantly below 20%. The current United Kingdom guidance therefore avoids allowing patients to return to driving whilst their risk is above 20%. In fact the estimate at three months after treatment withdrawal was significantly below 20% at 15% (10% to 19%). It might be therefore that current United Kingdom guidance is too conservative and patients might be eligible to return to driving earlier.

For patients who have a seizure recurrence during or following AED withdrawal and reinstate treatment, data from the Antiepileptic Drug Withdrawal study indicated that the risk of a seizure recurrence in the next 12 months having been seizure free for six months after reinstating treatment was 18%. However, due to the relatively small sample size, the 95% confidence interval (10% to 27%) does not exclude a recurrence risk in excess of 20%. At three months after recommencing AED treatment, the risk of a seizure in the next 12 months was estimated at 26% (17% to 35%) while at 12 months after reinstating treatment the risk was 17% (8% to 27%). If the focus is purely upon the point estimates, these data suggest that the current United Kingdom legislation is too conservative. Conversely, the new European Union minimum standard appears too liberal. These data will need to be

taken into account by European Union member states as they make decisions about adopting the proposed minimum driving standards.

As highlighted by the systematic review, few published prospective studies have assessed seizure recurrence risks following AED withdrawal and none have assessed seizure recurrence risks for patients who have a recurrence and reinstate treatment. Further prospective studies are required to inform driving policy and patient counselling. The epilepsies are a heterogeneous group of disorders and it is likely that some patients will be at a higher risk of seizure recurrence than others [119, 193, 225] and such studies will need to be large enough to allow an examination of predictors of risk. However, further guidance is now required from driving authorities as to if and how to stratify risk when considering fitness to drive.

Chapter 7: Internal Validation of Prognostic Models

7.1 Introduction

This chapter begins by explaining the purpose of a prognostic model and therefore identifies why validation is an important part of the modelling process. The chapter then goes on to summarise methods for evaluating model performance (section 7.2) and for investigating internal validation via apparent (section 7.3.1), split-sample (section 7.3.2), cross (section 7.3.3) and bootstrap (section 7.3.4) validation. Results from the internal validation of the prognostic models developed in Chapter 4 are then presented in section 7.4.

A predictive model's purpose is to provide valid outcome predictions for new patients. When a prediction model is developed it is important to assess how good the model predictions are – this is known as model performance. The distance between the predicted outcome and actual outcome is central to quantifying overall model performance from a statistical perspective. These distances are related to the concept of goodness-of-fit of a model with better models having smaller distances between predicted and observed outcome.

Essentially, the dataset used to develop a model is not of interest other than to learn for the future. Validation is therefore an important aspect of the process of predictive modelling. There are two types of validation – internal which assesses model validity for the setting where the development data originated from, and external which assesses validity in other samples which are fully independent from the development data and

originate from different but plausibly related settings. Internal validity will be considered in this chapter whilst external validation will be considered in Chapters 8 and 9.

7.2 Evaluating Model Performance

A number of performance measures exist for prediction models many of which are related to each other. From a statistical perspective a small distance between the observed outcome and the predicted outcome is desired. Explained variation can then be used to indicate performance and assess the predictability of the outcome. Measures that are in wide use in medical journals today include the concordance statistic (area under the ROC curve) for discrimination and various tests for calibration and goodness-of-fit. Discrimination and calibration are also considered as measures of external validation. Further details about the application of these measures can be found in Chapters 8 and 9.

7.2.1 Measures of Explained Variation

In linear regression the coefficient of determination, R^2, is widely used to indicate how well the covariates, z, in the form of a predictor βz, 'reduce the randomness' in the outcome variable y in a particular dataset. R^2 has many attractive properties [226]; it is the square of the Pearson correlation coefficient between the predictor βz and y, it is maximised by the usual least-square or maximum likelihood estimators, it is asymptotically independent of the sample size and it is easy to understand and can be explained by graphical means. R^2 is a population parameter so it depends not only on β parameters but also on the distribution of z.

For other regression models a similar measure is desired, however, the concept of variance or residual variance is more elusive or not uniquely defined. Generalisation of R^2 to other regression models can be done in many ways depending on which properties one wishes to retain and which definition one wants to generalise.

Various authors have tried to develop a useful R^2 measure for Cox's model however an additional complication in this situation is censoring. Some authors aim at measuring how well the covariates explain the variation in a particular dataset given its censoring structure while some aim at measuring how well the covariates actually predict survival in the wider, underlying population from which the data were sampled.

Schemper and Stare [227] suggest that for measures of explained variation for the Cox model: the expected value of a measure should not be substantially affected by censoring provided that an underlying uncensored population can be assumed [228]; a measure should not be affected by monotonic transformation of the time scale as the estimation of the coefficients in the Cox model is also invariant to such transformation; a measure should have an intuitively clear interpretation [229]; if a dataset could alternatively be analysed by a linear model the resulting values of explained variation should be similar for both approaches; and a measure should also be robust when applied to mis-specified models. There are further criteria [229] which should be met, such as independence from the units of measurements of model variables, and a range of possible values between 0 and 1.

Several measures of explained variation have been suggested for the Cox model. These are outlined in sections 7.2.1.1 to 7.2.1.9 together with a description of their adherence to the criteria.

7.2.1.1 Spearman, Kendall and Somer

When attempting to measure the correlation between two sets of data, various correlation coefficients may be used. In particular, two non-parametric correlation coefficients are Spearman's rank correlation coefficient, r_s, and Kendall's rank correlation coefficient, τ [230]. Of these, Spearman's is more widely used [231]. In significance testing both measures usually produce very similar results [232]. Although Kendall's τ is more demanding from a calculation point of view, its distribution approaches normality more

rapidly than Spearman's r_s. Additionally, although the two coefficients produce similar results, Spearman's r_s tends to be larger than Kendall's τ in absolute value [232]. Somers' D is an ordinal measure of association introduced by Somers [233]. It can be defined in terms of Kendall's τ. Both Somers' D and Kendall's τ are conservative estimates of the degree of relationship between the dependent variable and the independent variable [234]. Somers' D is different from Kendall's τ in that it only makes a correction for tied pairs on the independent variable [235].

Korn and Simon [236] suggest an approach to calculate Somers' D which in general requires numerical integration and leads to results dependent on the choice of T_0, a common censoring time for all survival times, which cannot exceed the maximum observed survival time. Brown *et al*. [237] suggested two generalisations of Kendall's τ for censored data which are both biased towards zero as their 'imputation' for an undefined score from the comparison of two survival times does not take into account the correlation of the covariates.

The definition of R^2-measures based on the squared correlation of the predictor and of survival for the general linear model is appealing for its simplicity. It is obvious that r, denoting any of the standard correlation coefficients, has to be replaced by a measure of monotonic association as the relationship between predictor and outcome is non-linear but usually monotonic. Furthermore, to be consistent with the Cox model, the measure should only use the order of survival times. Therefore, Spearman's r_s, Kendall's τ, Somers' D or related measures are appropriate in this situation.

A hybrid of r and r_s is r_{pr} (partial ranking); this coefficient is the Pearson correlation, ρ, of the predictor x' and the ranks of survival time y. It is more consistent with the semi-parametric character of the Cox model, and its superiority over r_s can be demonstrated for a model with a strong binary and a weak continuous covariate [238].

7.2.1.2 Schemper (& Henderson)

In 1990 Schemper described a measure, V_2 [239], based on a measure of the proportion of explained variation, a generalisation of the sum of squares definition of R^2, with differences being calculated on the axis of survival proportion rather than survival times. Schemper also proposed [239] a measure V_1 which is identical to V_2 except for the squaring of the averaged residuals which are omitted. V_2 is preferred over V_1 because of the consistency with related methodology. Additionally, in the case of V_2, the Kaplan-Meier estimator produces means of the individual survival process under the assumption of non-informative censoring and these means minimise the average squared 'residuals'.

Both measures are, however, dependent on the amount of censoring and therefore violate the first criterion for an assumed uncensored population. Schemper's approaches permit a very intuitive interpretation: the measures quantify the degree to which prediction of the values of an individual survival process can be improved when replacing the unconditional with the conditional expectation (survival functions) of the process.

None of the measures described in sections 7.2.1.1 and neither V_1 and V_2, which relate to the R^2 definition based on the squared correlation of the predictor and of survival, provide an unbiased estimate of the correlation in an uncensored population from a censored population [227]. However, in 1997 Schemper accomplished this using multiple imputation [240]. In this technique, censored survival times are terminated with random residual lifetimes to become uncensored survival times. The algorithm enforces consistency with the estimates from the Cox model and with observed conditional distributions of survival times. Several such augmented datasets are generated, a measure of association between survival and predictor being calculated for each and the squared average association taken as the explained variation.

With the multiple imputation algorithm, proposed by Schemper [240], the computed values of all rank correlation measures considered in section 7.2.1.1, r_s^2, r_{pr}^2, D^2 and τ^2, are unaffected by censoring and are also invariant under transformation of survival time. Only r_{pr}^2 is sensitive to transformation of the explanatory variables, thus being more consistent with the assumptions underlying the Cox model than the others. Also, the criterion of good intuitive interpretability and simplicity is fulfilled.

In 2000, Schemper and Henderson [238] suggested a new measure of the proportion of variation of possibly censored survival times explained by a given proportional hazards model. The proposed measure, termed V, shares several favourable properties with the earlier V_1 but also improves the handling of censoring. The statistic contrasts distance measures between individual survival processes and fitted survival curves with and without covariate information. Schemper and Henderson recommend graphical comparisons of survival curves for prognostic index groups to improve the understanding of obtained values. The main advantages of the new measures are correction for the weakness in the handling of censoring [241] and direct estimation of well-defined population quantities.

7.2.1.3 Korn & Simon

Korn and Simon [236] suggest using the proportion of explained variation measure in the context of survival data. Apart from the squared difference, they consider other possible forms of the distance of the response y_i from a central location parameter, unconditional or conditional on a given model and covariate vector x_i for the i^{th} observation. The problem with survival data is these distances usually cannot be calculated because of the censoring in y. To overcome this problem Korn and Simon [236] do not use the actual survival times, but times derived from the predicted survival functions. Thus the actual survival times are used only in fitting the model, while in calculating the distance for an

individual an expected value of the distance of y from its predictor is calculated where y is distributed according to the estimated survival function for individual i.

This measure is dependent on the amount of censoring so violates the criterion for an assumed uncensored population. Furthermore, this approach also violates another criterion as the required differences in the time scale make their measure dependent on transformations of time; measures for two Cox models with identical regression coefficients can therefore differ. This measure quantifies the degree to which prediction of the time of death, within a range of interest, can be made more precise when replacing unconditional by conditional survival functions.

7.2.1.4 Graf & Schumacher

Graf and Schumacher [241] investigated relationships of Korn and Simon's [236] explained variation measure based on integrated binary squared loss and of Schemper's [239] V_2. They also dealt with the interesting aspect of describing explained variation in the course of time rather than averaging over time. The resulting measures are similar to those of Schemper *et al.* [227, 238-240].

7.2.1.5 Nagelkerke

Nagelkerke [226] suggested the use of likelihood based measures, $R^2_{LR} = 1 - (\frac{L_R}{L_U})^{2/n}$ where L_R is the likelihood for the model without predictors (otherwise known as the intercept model) and L_U is the likelihood for the full model, for the Cox model which will work for time dependent $z(t)$. There are three possible definitions of n; n is the total sample size including censored observations which had been proposed already in 1988 by Kent and O'Quigley [242], n is the number of uncensored observations i.e. (for continuous time) the number of failure times, or n is ^2log (cardinality of the outcome space).

Additional requirements should be imposed to make a choice; the definition of n should be meaningful in the context of conditional logistic regression, for which there are no censored observations, and joining or splitting of datasets from the same population should not (much) affect the value of R^2. This leaves only the number of uncensored observations, failure times, as a valid choice. For tied data Nagelkerke suggests taking the number of failures and not the failure times.

This likelihood based measure is a reasonable measure of how well the covariates reduce the randomness in a particular dataset with particular censoring properties. It is easy to calculate and to interpret, also for time dependent $z(t)$, but it depends on the censoring pattern. The dependency is usually not very much if the proportional hazards assumption is correct and the censoring is random – the distribution of $z(t)$ changes with time and censoring removes more late failure times than early failure times, but the effect is usually small [226].

7.2.1.6 Haegerty & Zheng

Haegerty and Zheng (2005) link predictive accuracy to sensitivity and specificity of a marker M (e.g. βz) at any time point where sensitivity is $P(M_i > c|T_i = t)$ and specificity is $P(M_i \leq c|T_i > t)$. This leads to time dependent ROC curves. The suggested measure is the weighted average of the areas under the curve of such curves.

7.2.1.7 O'Quigley

O'Quigley *et al* [243] derived a 'variance' measure on the basis of squared Schoenfeld residuals [173]. It is based on predicted covariates at failure times rather than predicted survival time. The measure is elegant and stays close to properties of the proportional hazards model. However, it is focused on a specific dataset, does not have an intuitive interpretation and has little meaning outside the proportional hazards setting.

7.2.1.8 Kent & O'Quigley

Kent and O'Quigley [242] apply the information gain approach of Kent [244] to their measure. They use the fact that estimates of the coefficient in the Cox model are invariant under monotonic transformation of the time scale to choose a Weibull model for the baseline hazard. In their paper, Kent and O'Quigley also present a simple approximation for squared multiple correlation coefficients ρ_W^2, $\rho_{W,A}^2$, and a related simple measure ρ_{PM}^2. The usage of expected log-likelihoods makes ρ_W^2 independent of censoring. The measures ρ_W^2, $\rho_{W,A}^2$ and ρ_{PM}^2 fulfil the first and second of Schemper and Stare's [227] criteria.

7.2.1.9 Maddala

Maddala's R_M^2 is appealing as it is easily calculated and is equal to the usual R^2 in linear regression under normally distributed errors [245]. It is not affected by monotonic transformation of the time scale, but it violates the first criteria, its expectation being affected by censoring. Under no censoring, Kent and O'Quigley's ρ_W^2 is not necessarily close to Maddala's R_M^2, because the likelihoods are calculated based on the densities of the extreme value distribution and not using partial likelihoods [239].

7.2.1.10 Conclusion

The proportion of variation in a response variable that is explained by a fitted model is often used in statistical modelling to summarise the fit of the model. A number of measures of explained variation have been proposed for use in modelling survival data, which are analogues of the R^2 statistic that is widely used in linear regression analysis.

There are clearly other possibilities to those listed in section 7.2.1 for measures of explained variation. According to Nagelkerke [226], nobody has looked at the reduction by using covariates in the predicted time span that an individual will die with $p_1 - p_2$ probability. This has direct relevance for doctors and patients and also has meaning

outside the context of specific models. This proposal however lacks all the invariance properties of the proportional hazards model.

Schemper and Stare [227], in a review of new measures for explained variation, conclude that no particular statistic can be recommended for general use. Moreover, the most satisfactory measures are more difficult to compute. For this reason, measures of explained variation are not routinely used in modelling survival data [167].

7.2.2 Discrimination

Harrell's R_H^2 [246] has been shown to be inadequate as a measure of proportion of explained variation [239]. The explained proportion of log-likelihood does not permit an analogous interpretation of explained variation for the dependent variables, its values being implausibly low when compared with all other measures.

Harrell also suggested the c-index [247] which is based on pairwise comparisons of individuals in a sample for which the shorter life-time is uncensored. Each pair is either 'concordant' (one individual has higher values on both covariate and survival time), 'discordant' (the individual with a lower value on one covariate has a higher value on the other) or 'uncertain' (explanatory variable x' is tied). This c-index, the 'explained proportion of concordances' can be transformed to Somers' D. The c-index is frequently used, typically ranges between 0.5 and 1, and cannot be regarded as a measure of explained variation although it can be used as a measure of discrimination.

Discrimination refers to a model's ability to correctly distinguish the two classes of outcome [248]. A model with good discrimination ability produces higher predicted probabilities for subjects who had events than subjects who did not have events. The area under the ROC curve is one of the most used measures for model discrimination [249] and relates to Harrell's c statistic described in further detail in Chapter 8. In the literature a

value of more than 0.70 for the area under the ROC curve is considered satisfactory for discrimination [250].

7.2.3 Calibration

Model calibration describes how well the estimates of survival from the model correspond to the survival from the observed data [251, 252] and can be described as a measure of the extent of bias in a model [184]. A model is well calibrated when predicted and observed values agree for any reasonable grouping of the observation, ordered by increasing predicted values. A common form of the calibration statistics is based on the Pearson χ^2 statistic that compares the observed and expected outcomes within each group defined by the rank ordering of the predicted probabilities [253].

In a survival context the calibration of a model is usually studied at fixed time points. For these time points grouped patients can be considered with sufficient numbers per group to allow for calculation of survival rates with the Kaplan-Meier method [254]. This observed survival is compared with the mean predicted survival from the prognostic model.

7.3 Internal Validation of Prediction Models

Internal validation assesses model validity for the setting where the development data originated from [8]. It is important to obtain an honest estimate of performance for patients that are similar to those in the development sample. Several techniques are available to assess internal validity such as apparent validation, split-sample validation, cross-validation and bootstrap validation.

7.3.1 Apparent Validation

With apparent validation, model performance is assessed directly in the sample from where it was derived, but this leads to a biased estimate of performance as model parameters were optimised for the sample [8]. However all of the available data is used to

develop the model and all of the data is used to test the model hence the procedure gives optimistic but stable estimates of performance.

7.3.2 Split-sample validation

With split-sample validation, the sample is randomly divided into two groups. In one group the model is created and in the other the model performance is evaluated. Typical splits are fifty-fifty or two thirds to one third.

If samples are split fully at random, imbalances may occur with respect to the outcome and the distribution of predictors. For predictors with skewed distributions the consequences may be even worse. Stratifying the random sampling by outcome and relevant predictors may alleviate this problem.

There are many drawbacks to split-sample methods [254-256] - only part of the data is used for model development which leads to less stable model results compared with development with all development data i.e. there are problems related to variance. There are also problems related to bias - an assessment of the performance is obtained when part of the data is used. The performance of a model based on the full sample is actually required.

Split-sample validation is a classical but inefficient approach to model validation [8]. Simulation studies have shown that large sample sizes are required to make split-sample validation reasonable [257] but with large samples, the apparent validity is already a good indicator of model performance. Steyerberg [8] therefore concludes that split-sample validation is a method that works when it is not needed.

7.3.3 Cross-validation

Cross-validation is an extension of split-sample validation aiming for more stability [8]. A prediction model is again tested on a random part that was left out from the sample and

the model is developed in the remaining part of the sample. However in cross-validation this process is repeated for consecutive fractions of patients. For example, the dataset may be split into deciles with model development in nine of the ten and testing in one of the ten, which is repeated ten times. In this way all patients have served once to test the model. The performance is commonly estimated as the average of all assessments [254].

Cross-validation can use a larger part of the sample for model development than split-sample validation which is an advantage. However, the cross-validation procedure may need to be repeated several times to obtain stable results. The most extreme cross-validation is to leave out each patient once, which is equivalent to the jack-knife procedure [258] but with large numbers of patients, this procedure is not very efficient.

Cross-validation may not properly reflect all sources of model uncertainty, such as that caused by automated variable selection methods. The underestimation of variability is easily recognised for jack-knife cross-validation, where the development sample is identical to the full sample except for one patient [258], hence the same predictors will generally be selected in each jack-knife sample as in the full sample. Such model uncertainty can better be reflected with bootstrap validation.

7.3.4 Bootstrap Validation

Bootstrapping reflects the process of sampling from the underlying population. Bootstrap samples, which are of the same size as the original sample, are drawn with replacement from the original sample, reflecting the drawing of samples from an underlying population [258]. In the context of model validation, 100-200 bootstraps may often be sufficient to obtain stable estimates.

For bootstrap validation a prediction model is developed in each bootstrap sample. This model is evaluated both in the bootstrap sample and in the original sample. The difference

in performance indicates the optimism which is subtracted from the apparent performance of the original model in the original sample [254, 257-259].

There are many advantages of bootstrap validation; the optimism-corrected performance estimate is stable since samples of size N are used to develop the model as well as to test the model. This is similar to apparent validation and an advantage over split-sample and cross-validation methods. Compared with apparent validation, some uncertainty is added by having to estimate the optimism. When sufficient bootstraps are taken, this additional uncertainty is however negligible.

Additionally, bootstrap validation can appropriately reflect all sources of model uncertainty, especially variable selection [260]. Disadvantages of bootstrap validation, and other resampling methods such as cross-validation, include that only automated modelling strategies can be used, such as fitting a full model without selection, or following an automated stepwise selection approach.

7.4 Internal Validation of the SANAD Models

Analyses were undertaken using *R* version 2.9.2 [182].

It is of importance to determine whether the models presented in Chapter 4 are good at predicting outcome. The SANAD dataset is a large dataset and therefore it is unsurprising that nine prognostic factors were identified. Of greater importance than the statistical significance of these covariates is the ability of the models to predict outcome. Consequently, internal validation was performed with bootstrap resampling to assess this.

To assess explained variation, the Nagelkerke R^2 measure of model fit was estimated. As described in 7.2.1.5, Nagelkerke's measure is easy to calculate and interpret. It is also

readily available in statistical software packages and the most frequently reported R^2 measure [261]. For these reasons it was our preferred measure.

The bootstrap resampling approach was applied to assess the extent of model over-optimism (overfitting) in the final model using 200 bootstrap resamples. The bootstrap validation method was preferred as the optimism-corrected performance estimate is stable. This is an advantage over split-sample and cross-validation methods. Additionally, bootstrap validation can reflect all sources of model uncertainty and does not lead to biased estimates of performance, unlike apparent validation, if sufficient bootstraps are taken (section 7.3.4). The bootstrap resampled estimate of R^2 was produced to give an improved estimate of model accuracy.

Model discrimination was considered via Harrell's c-statistic, again bootstrapped over 200 resamples. Finally, calibration plots were produced using the same 200 bootstrap resamples. Graphical results can be easier to interpret than p-values, therefore we preferred calibration plots over the use of Pearson's χ^2 statistic [238].

7.4.1 Time to Treatment Failure

The Nagelkerke R^2 measure of model fit was 0.133. This suggested that the covariates reduce the randomness in the SANAD dataset by 13%. This is a low value however, as explained above, the R^2 measure is not ideal for Cox models. The bootstrap resampled estimate of R^2 was 0.139 suggesting that model under-optimism was only 1%.

The concordance statistic was 0.596 for the model and for the bootstrapped model it was 0.600. These values differ by only 0.004 which suggested that there is little under-optimism. The c-statistic of 0.6 showed that the model accurately discriminated patients 60% of the time. Given the desire was to predict patients with poorer outcome our c-statistic represents reasonable internal validation.

Calibration plots at one, two and three years can be seen in Figure 19, Figure 20 and Figure 21 respectively. If the model was perfectly calibrated at each time point the predicted (blue lines) and observed (black lines) values would sit perfectly on each other. In the case of the SANAD model for time to treatment failure, the predicted and observed lines are not perfectly matched although they are close, especially at two and three years. Therefore, there is no reason to doubt the internal validity of the time to treatment failure model for the SANAD data.

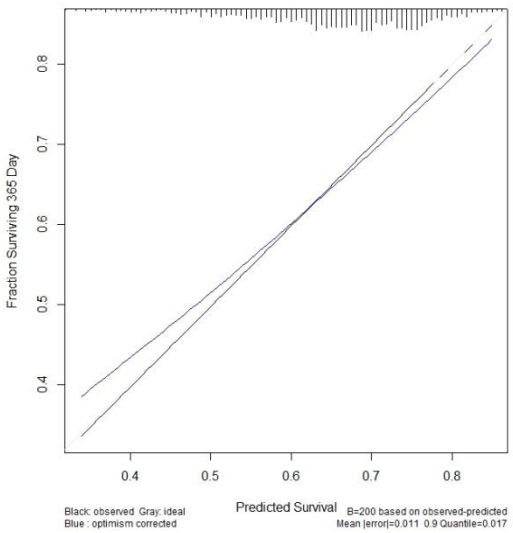


Figure 19: The SANAD Trial – arm A: Calibration plot for time to treatment failure (Table 6, Chapter 4) at 1 year

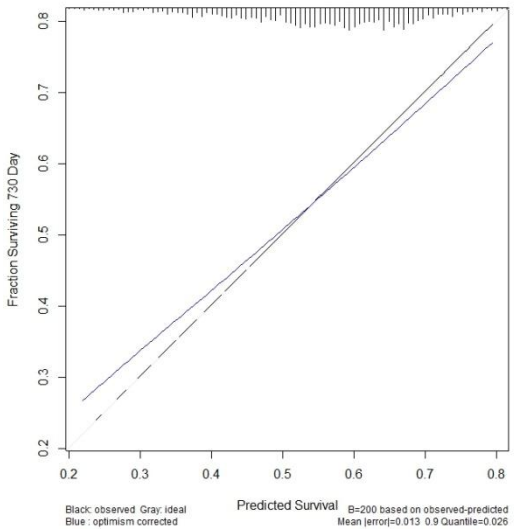


Figure 20: The SANAD Trial – arm A: Calibration plot for time to treatment failure (Table 6, Chapter 4) at 2 years

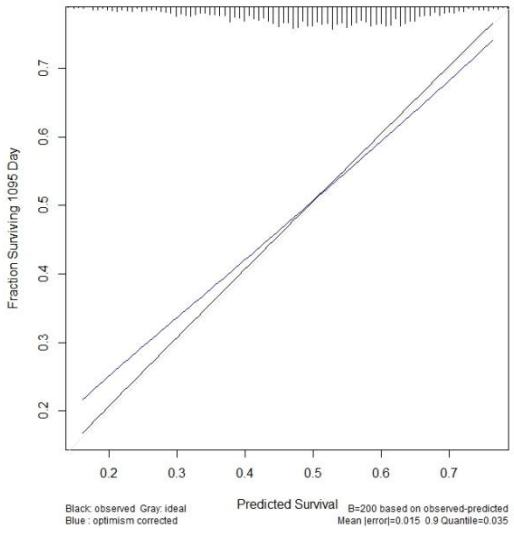


Figure 21: The SANAD Trial – arm A: Calibration plot for time to treatment failure (Table 6, Chapter 4) at 3 years

7.4.2 Time to 12 Month Remission

The Nagelkerke R^2 measure of model fit was 0.118 suggesting that the covariates reduced the randomness in the SANAD dataset by 12%. The bootstrap resampled estimate of R^2 was 0.144 suggesting that model under-optimism was only 3%.

The concordance statistic was 0.665 for the model and for the bootstrapped model it was 0.666. These values differ by only 0.001 which suggested that there was little under-optimism. The *c*-statistic of 0.7 showed that the model accurately discriminated patients 70% of the time which therefore suggested acceptable internal validation [139, 262].

Calibration plots at one, two and three years can be seen in Figure 22, Figure 23 and Figure 24 respectively. In the case of the SANAD model for time to 12 month remission, the predicted and observed lines are very close at two and three years suggesting good model validity. At one year the lines are not very close suggesting that the model does not fit the data particularly well at one year. This may be because approximately 30% of patients achieved remission immediately at 365 days. Therefore more sophisticated techniques may be needed to take account of this in the modelling process. Mixture modelling is used in Chapter 10 to see if this patient subgroup, together with those who achieve remission but at a later time point, and those who do not achieve remission, can be modelled more appropriately.

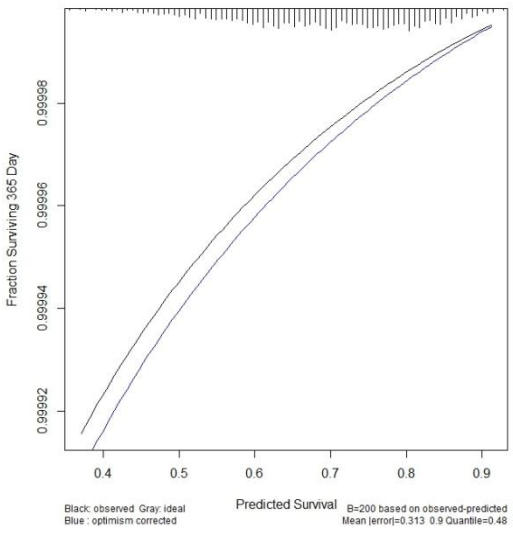


Figure 22: The SANAD Trial – arm A: Calibration plot for time to 12 month remission (Table 9, Chapter 4) at 1 year

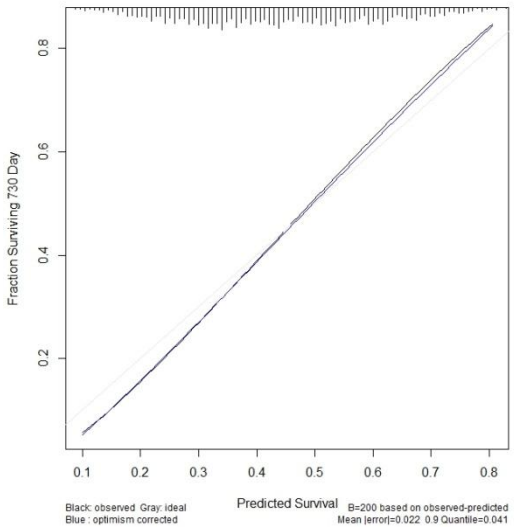


Figure 23: The SANAD Trial – arm A: Calibration plot for time to 12 month remission (Table 9, Chapter 4) at 2 years

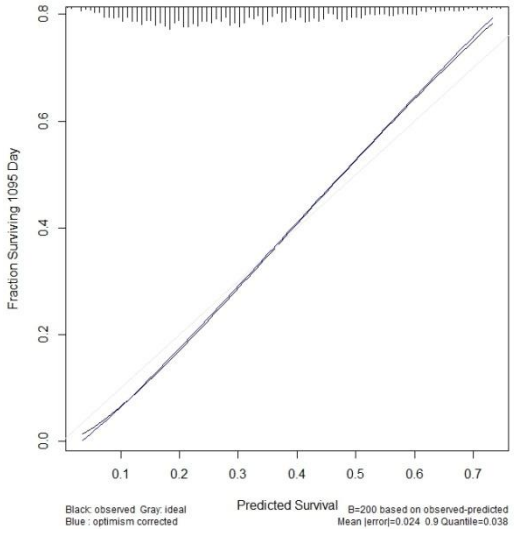


Figure 24: The SANAD Trial – arm A: Calibration plot for time to 12 month remission (Table 9, Chapter 4) at 3 years

7.5 Conclusion

Prognostic models have many uses such as guiding health-care policy, determining study eligibility of patients for new treatments, defining inclusion criteria for clinical trials to control for variation in prognosis, and selecting appropriate tests and therapies in

individual patient management including supporting decisions on withholding or withdrawing therapy. The evaluation of prognostic models has become an increasingly important topic with a rise in the number of such models in the public domain. There has therefore been a requirement to evaluate and validate prediction models. There are three main ways that a model can be evaluated and two ways that a model can be validated.

In section 7.2.1 a number of measures of explained variation were proposed for evaluating survival data, which are variations of the R^2 statistic that is widely used in linear regression analysis. Schemper and Stare [227], in a review of new measures for explained variation, concluded that no particular statistic could be recommended for general use. Moreover, the most satisfactory measures are more difficult to compute [167].

Overall statistical performance measures incorporate both calibration and discrimination aspects. Discrimination was described in section 7.2.2. It relates to how well a prediction model can discriminate those with the outcome from those without the outcome. Calibration, described in section 7.2.3, relates to the agreement between observed outcomes and predictions. Studying discriminative ability and calibration is often more meaningful than an overall measure such as R^2 when the quality of model predictions for individuals needs to be appreciated [8].

A model is developed within a representative sample of patients from an underlying population. This underlying population has specific characteristics. At least the internal validity, or reproducibility, of the predictive model should be determined for the underlying population. This can be done by testing the model in the development sample by either apparent, split-sample, cross or bootstrap validation as outlined in section 7.3. Bootstrapping appears most attractive provided that all modelling steps can be replayed which may sometimes be difficult.

Having applied bootstrap resampling to the SANAD models for time to treatment failure and time to 12 month remission (sections 7.4.1 and 7.4.2), it was found that both models showed signs of fairly good internal validation. There are concerns over the validity of the 12 month remission model at one year. However, in Chapter 10 mixture modelling techniques will be employed to attempt to better account for the patient subgroups – immediate remission at 365 days, delayed remission and no remission.

Further work is now needed to validate the models in external data. A simulation study assessing methods for external validation is presented in Chapter 8 while an application of the results of the simulation study to real data can be found in Chapter 9.

Chapter 8: Externally Validating Prognostic Models – Simulation Study

8.1 Introduction

In Chapters 4, 5 and 6, I developed prognostic models for various patient groups and various outcomes. Before a predictive or prognostic model can be introduced into general practice, it should be externally validated to ensure that it performs satisfactorily in datasets that are fully independent of the development data. In theory the dataset used to externally validate the model will be plausibly related to the development data meaning that both datasets will effectively be samples taken from the same 'super-population' [8]. Therefore this chapter will now examine external validation methods, while Chapter 9 will apply the results of this chapter to real data.

This chapter begins with an overview of types of external validation (section 8.2), a literature review (section 8.2.1) and a summary of the most commonly employed methods of external validation (sections 8.2.2 to 8.2.6). Methods for handling missing data within covariates are then presented in section 8.3 together with possible adaptations of these methods to the situation of handling a covariate missing from the validation dataset. A simulation study which compares the performance of methods of external validation and also compares the performance of methods of handling a covariate missing from the validation dataset is then presented in section 8.4.

As yet, very few authors externally validate their predictive model(s). This may be due to a lack of evidence as to the best method to apply or perhaps because it is difficult to identify comparative studies that provide information on the same set of potential confounders

[263] and consequently the development and validation datasets are unlikely to provide information on the same set of potential confounders. There is no literature available on how to deal with this situation apart from the suggestion of filling in any gaps with a constant value [8].

Initial work has been presented for missing covariates in an individual patient data meta-analysis [263] via way of a bivariate random-effects meta-analysis. For this method information is used from all available cohorts while adjusting for all the potential confounders; fully adjusted and partially adjusted estimated effects are used in cohorts with full confounder information together with an estimate of their within-cohort correlation. This method cannot be used as a method of external validation with a validation dataset missing a covariate because the method [263] assumes that full and partial covariate models are correlated and that this correlation is similar across trials – with one development dataset and one validation dataset this is a big assumption to make.

Prognostic models have many uses such as guiding health-care policy, determining study eligibility of patients for trials of new treatments, defining inclusion criteria for clinical trials to control for variation in prognosis, and selecting appropriate tests and therapies in individual patient management including supporting decisions on withholding or withdrawing therapy. The evaluation of prognostic models has become an increasingly important topic recently with an increase in the number of such models in the public domain. Therefore prognostic models must be externally validated and consequently simple but reliable methods for external validation, especially in the common situation of covariates missing from the validation dataset, must be determined.

In what follows, the data used to derive the model will be referred to as the development dataset. The external, independent, data used to validate the model will be referred to as the validation dataset.

8.2 External Validation

The purpose of a prognostic model is to provide valid outcome predictions for patients. Before a prognostic model is used in clinical practice it is important to have an indication that the model would work well on a second similar cohort to that used to develop the model. External validity, also known as generalisability or transportability, is essential to support general applicability of a prediction model to populations that are 'plausibly related'. The definition of plausibly related populations requires subject knowledge as it is not self-evident. Steyerberg [8] defines such populations as those that can be thought of as parts of a 'super-population' and therefore populations that have matching covariates.

There are different types of plausibly related populations leading to several variations of external validation: temporal, geographic and fully independent [8]. With temporal validation, a predictive model is validated in more recently treated patients. This may be achieved straightforwardly by splitting the development data into a part containing early treated patients, used to develop the model, and another part containing the most recently treated patients, used to assess the performance. An alternative is a prospective application of the model in a specifically collected cohort.

With geographic validation, a predictive model is validated according to site. It can be labelled leave-one-centre-out cross-validation. In geographic validation, the splits are not at random unlike in standard cross-validation. A drawback of geographic validation is that validation samples may get quite small, leading to unreliable results.

With fully independent validation, a predictive model is validated by independent investigators. Other investigators may use slightly different definitions of predictors, outcome and study patients that are differently selected compared with the development setting. Fully independent validation studies often provide more unfavourable results than temporal or geographical external validation [8]. If model performance is found adequate

by fully independent investigators the validation is more convincing than when found by the investigators who developed the model.

There are several main considerations when validating a model such as study design, measuring the intrinsic prognostic information, comparing predictions with observations, quantifying the performance of a model and pre-specifying adequate performance [264]. A system, however, can never be fully validated – you can never be certain that it will apply to the next suitable patient or subject [264, 265].

There are a few inter-related reasons why prognostic models may not perform well such as deficiencies of standard modelling methods, deficiencies in the design of prognostic studies and non-transportability of models [264]. In terms of inadequate model development, the sample size may be relatively small, or patients may be selected from a single centre. Also, statistical analysis may often be suboptimal e.g. with stepwise selection in relatively small samples with many potential predictors, and no shrinkage of regression coefficients to compensate overfitting [8]. Other explanations include true differences between development and validation settings, especially in coding of predictors and outcome. Moreover, variables required for a model may not be available at validation.

External validation is often referred to in the discussion section of a paper however it is unusual for external validation results to be presented. This may be because there is a variety of methods but no direct comparison of them and therefore it is difficult to know which methods to apply. Additionally, it may be because a suitable dataset for validation is missing a significant covariate. Again there are no published guidelines on how to handle this situation. This chapter addresses both of these issues.

8.2.1 Literature Review

There are several well-known examples of predictive models or indices used in medicine. Two such models are the Framingham risk prediction model for atherosclerotic cardiovascular disease [266] and the Nottingham Breast Cancer Index [267]. The performances of the models or indexes have been evaluated differently but authors of both have a common aim to ensure that the models perform sufficiently well and perform suitably in relation to prior knowledge. Neither model has been externally validated in the true sense but both have been validated sufficiently to be accepted into general clinical practice.

The Framingham model [266] was assessed via discrimination and calibration. In addition, the sensitivity of the top quantile resulting from the model was considered and finally, the performance of the model was compared with that of another popular Framingham risk score [268] in terms of survival probabilities. The Nottingham Breast Cancer Index [267] was assessed by comparing the performance of the index with that of the most significant factor alone in terms of survival probability. Additionally the index was compared with an earlier criterion for poor prognosis, again in terms of survival probabilities. In addition they considered a group of patients in the lowest quantile in terms of expected and observed survival and finally they examined the performance of the new index as a predictor of five year survival.

A literature review was undertaken to summarise characteristics of other previous studies that have externally validated a prognostic model. Any studies, excluding laboratory based ones, in which a prognostic model was externally valid were included. Information was collected on the size of the development dataset, the size of the validation dataset, the type of outcome measure, the number of covariates included in the development model, external validation methods and methods for handling covariates missing from the

validation dataset. A summary of the prognostic models was also obtained together with a description of the similarity in the data between the development and validation datasets.

The following strategy was used in MEDLINE which was searched from 2002 to 9th June 2011 – the last ten years was chosen to reflect the large volume of prognostic models published in recent years.

1. (Prognostic adj (model* or factor*)).ti.

2. (Predictive adj (model* or factor*)).ti.

3. 1 or 2

4. validat*.tw.

5. 3 and 4

6. limit 5 to yr="2002 - 2011"

One hundred and nine studies were found that met the inclusion criteria. Of these, 29 externally validated at least one previously published prognostic model, 38 validated a previously unpublished Cox proportional hazards model and 35 validated other unpublished models while seven did not mention the type of model being validated. The unpublished models that were not developed via the Cox model were developed using a variety of methods including Monte Carlo Markov Chains, regression (parametric, logistic, linear, fuzzy logic, bivariate), recursive partitioning, proportional odds, Bayesian networks, Generalised Estimating Equations, Artificial Neural Network and Real Time-Quantitative Polymerase Chain Reaction. Table 22 summaries the size of the development and validation datasets, the number of covariates and the methods of external validation per validated model type: published, unpublished Cox, unpublished other and not specified.

Overall the median size of the development dataset was larger than that of the validation dataset. On an individual level there were many occasions where the validation dataset was larger than the development dataset irrespective of the type of model being validated. In most cases only a few covariates were included in the model, irrespective of the type of model being validated although, of course, there are many exceptions to this. Most authors only employed one method of external validation while some listed up to seven. There was also large variation in the description of the methods.

Table 22: Summary of Literature Review identifying studies which externally validated a prognostic model. Entries are numbers of studies (percentage)

	Published (n=29)	Unpublished Cox (n=38)	Unpublished other (n=35)	Not specified (n=7)	Total (n=109)
Summary Statistics					
Size of Development Dataset, Median (IQR)	395 (150, 1298)	337 (124, 851)	478.5 (213, 2763)	124 (62, 174)	395 (151.5 to 1348)
Size of Validation Dataset, Median (IQR)	280 (117, 1230)	158.5 (110, 449)	280 (116, 1351)	189 (103.5, 306)	280 (115.5 to 1318)
Number of covariates in the model, Median (IQR); Maximum	4 (5, 7); 11	4 (3, 6); 12	5 (4, 7); 20	5 (3, 5); 6	5 (4 to 7); 20
Number of external validation methods used, Median (IQR)	1 (1, 2)	1 (1, 2)	2 (1, 2.5)	1 (1, 1.5)	1 (1, 2)
Methods					
Discrimination					
c-statistic	15	16	20	3	54
Other*	2	4	3	0	9
Repeat analysis and compare results	6	12	6	2	26
Calibration					
Hosmer-Lemeshow	4	1	5	0	10
Plots	3	2	2	1	8
Other*	1	1	3	0	5
Graphical comparison via R^2	1	5	2	0	8
Comparison of analysis product					
χ^2	1	2	3	2	8
Proportion of separation	0	1	2	0	3
t-test	1	0	3	1	5
Other*	5	10	7	1	23
Comparison of accuracy					
Brier	1	0	1	1	3
Sensitivity/Specificity	1	1	5	0	7
Positive predictive value	2	1	4	0	7
Negative predictive value	1	1	4	0	6
Other*	3	3	1	1	8
Other methods (Cochrane-Armitage, trend test, two-way ANOVA, Sheffé procedure, shrinkage etc.)	3	4	4	0	11

In general, discrimination (n=63), calibration (n=23) and comparison of analysis product (n=39) such as prognostic separation, were the most frequently implemented methods of external validation. Other frequently employed methods of external validation were comparison of accuracy (n=31) and repetition of the analysis with comparison of the results (n=26). Accuracy of models is frequently compared via sensitivity, specificity, positive predicted value and negative predictive value. Sensitivity and specificity are statistical measures of the performance of a classification function. Sensitivity measures the proportion of actual positives which are correctly identified as such while specificity measures the proportion of negatives which are correctly identified [269]. The whole point of a prognostic model is to use it to make a prognosis, so it is important to know the probability that the model will give the correct prognosis. The sensitivity and specificity do not give us this information. Instead the data can be approached from the direction of the results, using predictive values - the positive predictive value is the proportion of subjects with positive results who are correctly identified while the negative predictive value is defined as the proportion of subjects with negative results who are correctly identified [270]. The latter tend to be used for diagnostic tests although it can be adapted to other situations.

Many of the methods identified in the literature review are dependent on the type of model being validated. For example, positive and negative predictive values are most appropriate for models concerning diagnosis while the c-statistic can be used to measure the predictive power of any model [271] and the area under the ROC curve can be used to test the ability of a model to correctly classify those with and without the disease or event as appropriate [272].

Only one study considered how to handle a covariate missing from the validation dataset. In Mekhail's study [273] variables with a lot of missing data were not considered when the

model was re-developed using the validation dataset therefore the variables were forced to be missing from the validation dataset. Risk groups were developed using the development data and also using the validation data and were then compared. Any other study that mentioned missing data did so in the context of missing data within a covariate only. In most cases however, the validation datasets were collected specifically for validating a prognostic model, hence missing covariates were not an issue.

There is little evidence in the literature as to how to externally validate prognostic models and even less on which methods should be used in preference. In the literature review four broad methods of external validation were found: repeat analysis with comparison of results, comparison of observed and predicted values, calibration and discrimination. Although we found 109 studies which had made some attempt to externally validate a prognostic model many hundreds or even thousands of models for clinical prediction (prognosis or diagnosis) are published in the medical literature every year [274]. Therefore it seems that external validation is still infrequently implemented and often poorly reported.

Some methods of external validation are now described.

8.2.2 Predictive Performance

Proving that a model has good predictive qualities is not trivial. Several statistical measures have been proposed for this purpose and deciding which is the most suitable can often be challenging and will depend, not only on the type of outcome the model is predicting, but also on the clinical area within which it is being utilised as predictive values calculated in one clinical setting are not necessarily applicable in another. Popular choices of outcomes are continuous and binary however they are not exclusive as time-to-event outcomes are also useful. Censored observations are inherent to time-to-event datasets and survival analysis methods such as the Cox proportional hazards model are required to deal with

them. This censoring also necessitates special consideration when testing the performance of survival models.

Jorgensen [275] summarises such tests for binary and continuous outcomes but does not provide a summary for survival outcomes. In the case of binary outcomes goodness of fit via the Brier score, calibration, and sensitivity and specificity are considered. ROC curves are reviewed which determine a threshold probability above which a drug will not be prescribed, for example. To determine a model's diagnostic accuracy, positive and negative predictive values are suggested while contingency tables are recommended to calculate the measure of discriminatory ability. These methods constitute a toolkit for assessing how well a model fits a dataset, as well as how it performs in reflecting the true situation both at the population and at an individual level. However, none of these methods give consideration to the clinical consequences of applying the model in practice and as a result the conclusions can sometimes be misleading.

A regression model may be externally validated by checking the predictive accuracy of the model in the validation data and the goodness of fit of the model [275]. Some proposals for quantifying the predictive performance of survival models have concentrated on discrimination, while others have considered accuracy of predictions either at the group level, or for individual patients.

8.2.3 Discrimination

Discrimination is the ability of the model to allocate to patients who experience the outcome of interest a higher predicted probability of experiencing the event than that allocated to those who did not experience the event. It is often assessed via the area under the ROC curve. For continuous data the area under the ROC curve can be interpreted as the probability that the results of a randomly selected patient experiencing the event will

be greater than the result of a randomly selected patient without the event, i.e. the probability that the two results are ranked in the correct order [249].

To apply ROC curve methodology to survival data it is necessary to choose a particular time point t^* at which the survival status can be ascertained. Popular choices for t^* include a fixed number of years, the median follow up time or the median survival time. Graf [276] pointed out that when the prediction is made at $t = 0$ the status of t^* is not yet determined, but will evolve in the period from 0 to t^* according to some stochastic mechanism. For many patients the status at t^* will never be known because they are censored before t^*. Another criticism of this approach is that area under the ROC curve measures the discrimination of the model at an arbitrary time point rather than over all the relevant time period.

These issues are addressed by an adaptation of ROC methodology to survival data proposed by Harrell [184]. This method does not depend on choosing a fixed time for evaluation of the model and takes into account censoring of individuals. Harrell's c-index is defined as the proportion of all usable subject pairs in which the predictions and outcomes are concordant. In predicting the time to an event, c is calculated using all possible pairs of subjects, at least one of whom has suffered the event. If the predicted survival time is larger for the subject who actually survived longer, the predictions for that pair are concordant with the outcomes. If one subject suffered the event and the other is known to have survived at least to the survival time of the first, the second subject is assumed to survive the first. When predicted survival times are identical for a pair then 0.5 rather than 1 is added to the count of concordant pairs in the numerator of c, but one is added to the denominator which counts usable pairs. The formula for the concordance statistic is $c = \left(E + \frac{T}{2}\right)/N$ where E is the number of orderings as expected, T is number of tied

predictions and N is number of comparisons. A value of $c = 0.5$ corresponds to the agreement expected by chance and $c = 1$ corresponds to perfect concordance.

Harrell's c is a reparameterisation of Somers' D, similar to the area under the ROC curve. Rather than measuring the ability of a continuous variable to predict disease say, Harrell's c measures the ability of a continuous variable to predict survival. Somers' D is an asymmetric measure of association between two variables and is defined in terms of Kendall's τ_a [230] whose population value is traditionally defined as

$$\tau_{XY} = E\{\text{sign}(X_1 - X_2)\text{sign}(Y_1 - Y_2)\}$$

where (X_1, Y_1) and (X_2, Y_2) are bivariate random variables sampled independently from the same population and $E[.]$ denotes expectation.

This definition can be generalised to left- or right-censored, stratified, clustered, or weighted data in the following way. Suppose that 4-variate observations (X_i, R_i, Y_i, S_i) are sampled from an arbitrary population using an arbitrary sampling scheme. The R_i are censorship indicators for the corresponding X_i and the S_i are censorship indicators for the corresponding Y_i. These censorship indicators are negative for left censorship, positive for right censorship and zero for non-censorship [277].

Define a censored sign difference for two values, u and v with respective censorship indicators p and q as

$$\text{csign}(u, p, v, q) = \begin{cases} 1 & \text{if } u > v \text{ and } p \geq 0 \geq q \\ -1, & \text{if } u < v \text{ and } p \leq 0 \leq q \\ 0, & \text{otherwise} \end{cases}.$$

Given two observations (X_i, R_i, Y_i, S_i) and (X_j, R_j, Y_j, S_j), call the product of $\text{csign}(X_i, R_i, Y_i, S_i)$ and $\text{csign}(X_j, R_j, Y_j, S_j)$ the concordance-discordance difference for the

two observations and say that the two observations are concordant if the product is 1, discordant if the product is -1 and neither concordant nor discordant if the product is 0.

Kendall's τ_a can now be redefined as $\tau_{XY} = E\{\text{csign}(X_i, R_i, X_j, R_j)\text{csign}(Y_i, S_i, Y_j, S_j)\}$ i.e. the mean concordance-discordance difference. The population value of Somers' D [233] is defined as

$$D_{YX} = {}^{\tau_{XY}}/_{\tau_{XX}}.$$

Therefore τ_{XY} is the difference between two probabilities - the probability that the larger of the two X values is associated with the larger of the two Y values and the probability that the larger X value is associated with the smaller Y value. D_{YX} is the difference between the two corresponding conditional probabilities, given that the two X values are known to be unequal. Somers' D is related to Harrell's c index by $D = 2c - 1$ [184, 247].

To determine if there is a significant difference between two prediction models within the same dataset Newson [277] proposes calculating confidence intervals for the difference between Somers' D (equivalently Harrell's c) parameters of the predictors with respect to the outcome. Confidence intervals for Kendall's τ_a and Somers' D are more informative than p-values alone for three reasons. Firstly, confidence intervals show that a high p-value does not prove a null hypothesis. Secondly, for continuous data, Kendall's τ_a can often be used to define robust confidence limits for Pearson's correlation [277] by Greiner's relation [278]. Thirdly, confidence limits can be defined for differences between two Kendall's τ_a's or Somers' D's and these are informative because a larger Kendall's τ_a or Somers' D cannot be secondary to a smaller one. The sampling variances of the point estimates can be estimated using a jack-knife method via Newson's Stata package `somersd` [279].

It is not, however, appropriate to calculate a confidence interval for a difference in concordance statistic between the development and validation dataset. This is because the prognostic model would be fitted using the development dataset and then its predictive power would be measured. In general, confidence interval formula do not cover this situation because the fitting process typically chooses parameter values to maximise the apparent predictive power of the model in the development dataset. This will usually imply that the predictive power observed in the development dataset will be overoptimistic [271].

8.2.4 Calibration

The deviance of a model, D, defined as minus twice the log-likelihood can be thought of as a measure of the lack of fit of the data to the model. The performance of a survival model on external data may be measured by comparing the fit using the parameters estimated from the development dataset ('fixed' parameters) with the fit using the parameters estimated from the validation dataset [280].

Denote the deviance using the fixed parameters as D_F and the deviance using re-estimated parameters as D_R. If the two deviances are similar then the model generalises well. The difference in deviances, $D_R - D_F$, has an approximate χ^2 distribution with degrees of freedom equal to the number of parameters, p, in the model since this is the difference in the number of degrees of freedom between the model fitted with re-estimation and the model fitted without re-estimation of parameters [281].

This methodology was discussed by Spiegelhalter [281] *et al* when they used deviance statistics to choose the best model in the context of a Bayesian framework. The result that the deviance difference is distributed as χ^2 with degrees of freedom equal to the number of re-estimated parameters depends on the independence of the parameters. It is likely that the assumption is violated in this application of external validation. However, a test

based on the assumption of independence will tend to be liberal since the effective number of parameters must be less than or equal to the actual number of parameters and the sum of the degrees of freedom with the corresponding χ^2 reference value still provides a useful benchmark for comparing models [280].

8.2.5 Prognostic Separation

The aim of prognostic classification schemes is to define subgroups of patients with well separated survival distributions [282]. For survival data, the statistical methodology for assessing such schemes consists mainly of a range of *ad hoc* approaches [276]. Measures of separation may be used in situations where the proposed measures of inaccuracy cannot be calculated as a classification scheme is not supplemented by predictions of any kind such as survival time, survival status or survival probability. There are several measures of separation including, SEP [276] and PSEP [264], however no proposal seems to be universally accepted [283].

In survival analysis, a 'natural' approach often used in the medical literature to measure the prognostic ability of a model is to generate a prognostic classification scheme comprising two or more risk groups and to plot the Kaplan-Meier survival curves for each group. This leads to the idea of separation of survival curves as a measure of prognostic information, and led Graf and colleagues to propose a measure that they called SEP [276]. SEP is defined as

$$SEP = \exp\left[\sum_{j=1}^{g} \frac{n_j}{n} \left|\hat{\beta}_j\right|\right]$$

where n_j denotes the number of patients in risk stratum $\mathcal{X}_j$, $\hat{\beta}_j$ is the estimated log-hazard ratio or log-relative risk of patients in risk stratum $\mathcal{X}_j$ with respect to a baseline reference, n is the total number of patients and g is the number of groups of patients. The baseline

reference can be estimated in a Cox model where the dummy covariates for risk strata are centred to have mean zero and therefore the SEP measure is based on the assumption of proportional hazards between risk strata [276].

The measure, SEP, incorporates weights of relative size of the subgroups because a clinically useful categorisation needs to define groups which are not only well separated, but which additionally incorporate a substantial part of the patients. The SEP may be used for a comparison of different schemes but the measure itself is not very informative and gives no estimate of the degree of separation of the different subgroups. In the case of three or more categories, it does not take into account the ordering of the categories and therefore it may be misleading [282].

A simple index of separation, PSEP, can be defined as follows. Suppose the outcome of concern is death within a predefined period following measurement of prognostic factors. Suppose also that a prognostic classification scheme of some kind has been defined in some way. All that is required is that any patient may be classified into one of two or more prognostic groups, and that the groups with best and worst predicted prognosis have been identified. When there are more than two groups it is essential to ensure that the extreme groups, from which PSEP is calculated, are large enough and have sufficient events to allow adequate estimates of p_{worst} and p_{best}. If not, amalgamation of groups may be required [264]. Let

- $p_{\text{worst}} =$ predicted probability of dying for a patient in the group with worst prognosis
- $p_{\text{best}} =$ predicted probability of dying for a patient in the group with best prognosis

Then the predicted prognostic information can be measured by the separation, $PSEP = p_{worst} - p_{best}$. For example [284], if there were three risk groups, low, medium and high, and the respective six month death rates were 0.571, 0.787 and 0.947 then $p_{\text{High}} = 0.947$ and $p_{\text{Low}} = 0.571$ so $PSEP = 0.376$. To compare the PSEP for the original data with the

PSEP for the validation data, it is necessary to use the same data points and consider PSEP agreements at each time point. This method does not however, take into account fewer numbers at risk and increased uncertainty later in the curve.

Royston and Sauerbrei's method of separation [283] is based on the estimation of the underlying spread of the log hazard ratios compared with baseline. The estimate is most accurate when the underlying log hazard ratios are approximately normally distributed between individuals, although the method is applicable more generally.

Suppose the data on n individuals are denoted by $(t_1, \delta_1, \boldsymbol{x}_1), \dots, (t_n, \delta_n, \boldsymbol{x}_n)$, where for the ith individual t_i is the observed time, δ_i is 1 if the event of interest is experienced at t_i or 0 otherwise (right censoring), and $\boldsymbol{x}_i$ is the covariate vector of prognostic factors. The Cox model may be written as

Equation 12

$$\ln \lambda(t_i; \boldsymbol{x}_i) = \ln \lambda_0(t_i) + h_i$$

where $h_i = \boldsymbol{\beta}' \boldsymbol{x}_i$ is the prognostic index for the ith individual [283].

Consider the distribution of the prognostic index values. Defining order statistics $h_{(1)} < \cdots < h_{(n)}$ it is possible to write $h_{(i)} = \mu + \sigma u_i + \varepsilon_i$ where u_i is the ith expected standard Normal order statistic (rankit) in a sample of size n. Ordering the data on the h_i and substituting for $h_{(i)}$ in Equation 12 gives

$$\ln \lambda\left(t_{(i)}; \boldsymbol{x}_{(i)}\right) = \ln \lambda_0\left(t_{(i)}\right) + \mu + \sigma u_i + \varepsilon_i.$$

So far no specific distribution has been assumed for the h_i. Now suppose that the h_i are Normally distributed $N(\mu, \sigma^2)$ where σ is the standard deviation of the prognostic index values. By definition, the regression of the $h_{(i)}$ on the u_i is linear with $E\left(h_{(i)}\right) = \mu + \sigma u_i$ and $E(\varepsilon_i) = 0$ [283].

To a first approximation, ignore the random perturbation ε_i and set $\varepsilon_i = 0$. Then

Equation 13

$$\ln \lambda\left(t_{(i)}; \boldsymbol{x}_{(i)}\right) \cong \ln \lambda_0\left(t_{(i)}\right) + \mu + \sigma u_i.$$

Under the Normality assumption, the special Cox model (Equation 13) is approximately linear in the u_i. On fitting it to the data, the constant μ becomes part of the baseline hazard function and the regression coefficient, σ^*, will estimate σ. Royston and Sauerbrei's proposed measure D is defined as

$$D = \kappa\sigma^*$$

where $\kappa = \sqrt{8/\pi} \cong 1.60$.

Let $z_i = \kappa^{-1}u_i$. The slope from Cox regression (Equation 13) on the z_i is $\kappa\sigma^*$, giving D directly. The scaled rankits $z_1, \ldots, z_n$ may be calculated using Blom's approximation to rankits [285]

$$z_i = \kappa^{-1}u_i = \kappa^{-1}\Phi^{-1}\left(\frac{i - \frac{3}{8}}{n + \frac{1}{4}}\right)$$

where $\Phi^{-1}(.)$ is the inverse standard Normal distribution function. If some of the $\hat{h}_{(i)}$ are tied the relevant z_i are averaged within the groups defined by the tied values.

To calculate a confidence interval for the difference in separation between two models it is possible to take bootstrap resamples and compute the difference in separation between the development model and validation model [258] in each sample. For a large number of bootstrap resamples, the ratio of the mean difference in separation to its standard deviation may be assumed standard Normal under the hypothesis that there is zero difference between the separations. A confidence interval can therefore be calculated in the usual way [283].

8.2.6 Royston-Parmar's Alternative to the Cox Model

The Cox proportional hazards model was designed to estimate the effects of covariates on the hazard function, but not to estimate or predict survival probabilities. By definition, it does not satisfy full validation requirements. Instead Royston and Parmar [277, 286] recommend use of adequately flexible parametric survival models, both for estimation and validation.

The Cox proportional model is a near-universal tool in the analysis of prognostic models in clinical medicine [287]. However, it was not formulated with the concept of full external validation in mind. The main issue is that a parametric estimate of the baseline survival function is not available. Prediction in new data of survival probabilities from a Cox model is therefore problematic. Further, covariate effects must conform to the proportional hazards assumption, which is quite limiting, particularly in data with medium- or long-term follow-up.

Royston and Parmar describe families of flexible parametric survival models resembling generalised linear models with various link functions. The models rely on transformation of the survival function by a link function $g(.)$.

Equation 14

$$g[S(t;\boldsymbol{x})] = g[S_0(t)] + \boldsymbol{x}\boldsymbol{\beta}$$

where $S_0(t) = S(t;\mathbf{0})$ is the baseline survival function and $\boldsymbol{\beta}$ is a vector of parameters to be estimated for covariate vectors $\boldsymbol{x}$.

Exponentiating Equation 14, proportional cumulative hazards models are obtained as

$$H(t;\boldsymbol{x}) = H_0(t)\exp(\boldsymbol{x}\boldsymbol{\beta})$$

where $H(t;\boldsymbol{x})$ is the cumulative hazard function and

Equation 15

$$H_0(t) = H(t; \mathbf{0})$$

is the baseline cumulative hazard function. This is also a proportional hazard model.

Royston and Parmar approximated $\ln H_0(t)$ by a restricted cubic spline function of log time with two 'boundary' knots and m interior knots

Equation 16

$$\ln H_0(t) = \gamma_0 + \gamma_1 \ln t + \gamma_2 v_1(\ln t) + \cdots + \gamma_{m+1} v_m(\ln t)$$

where $\boldsymbol{v}(\ln t) = \ln t \;\; v_1, (\ln t), \ldots, v_m(\ln t)$ is a vector of spline basis functions [288]. Ignoring the constant term γ_0, a spline with m knots has $m + 1$ degrees of freedom. The proportional hazards model obtained by substituting Equation 16 for $H_0(t)$ in Equation 15 is fully specified parametrically and may be written as

Equation 17

$$\ln H(t; \boldsymbol{x}) = \gamma_0 + \boldsymbol{v}(\ln t)\boldsymbol{\gamma} + \boldsymbol{x}\boldsymbol{\beta}.$$

If the prognostic index is written as $\eta = \boldsymbol{x}\boldsymbol{\beta}$ then Equation 17 may be rewritten as

Equation 18

$$\ln H(t; \boldsymbol{x}) = \gamma_0^* + \boldsymbol{v}(\ln t)\gamma + \beta_\eta \eta.$$

Since the term $\boldsymbol{v}(\ln t)\gamma$ in Equation 18 has not changed, the shape of the baseline cumulative hazard function is unaltered, but its location changes through $\gamma_0^* + \beta_\eta \eta$ when $\gamma_0 \neq \gamma_0^*$ or $\beta_n \neq 1$. Estimation of β_η is relevant in validation studies where recalibration is of interest [289]. Regression on the prognostic index in the development data yields $\hat{\beta}_\eta = 1$. In the validation data, a value of $\hat{\beta}_\eta$ much less than 1 may indicate a model with poor performance. If $\hat{\beta}_\eta = 0$ the model is probably useless [290].

In a validation study, interest lies in whether values of γ_0, γ, or η from the development dataset still hold in the validation dataset. Sometimes a simple recalibration, $\beta_\eta\eta$, of η improves model fit in the validation dataset. In the validation dataset the prognostic index η is calculated using values of the parameter vector, $\hat{\beta}$, estimated on the development dataset and applied to the covariate values of the validation dataset. The components of β are not re-estimated on the validation dataset. Similarly, the baseline survival function is calculated in the validation dataset using the parameter vector $(\hat{\gamma}_0, \hat{\gamma})$ estimated on the primary dataset, together with the time values in the validation dataset and the set of spline knots used in the development dataset.

It is possible that the predicted survival probabilities derived from $\hat{\gamma}_0$, $\hat{\gamma}$ and $\hat{\beta}$ are sufficiently accurate in the validation dataset to remove the need to adjust any of the model parameters. This is the ideal situation regarding validation (case 1). There are six additional cases which comprise a structured ordered set of models allowing one to assess, and if necessary improve, the fit of the original model in the validation dataset. The adjustments provide a route to updating a given model in a simple way, thus avoiding complete reconstruction. Of course, in some situations reconstruction may be unavoidable. In any case, an updated model should be re-validated in further independent data [290].

The need for updating, if detected, indicates different types of lack of fit of the original model:

1. Re-estimate γ_0. Altering γ_0 moves the entire baseline distribution function up or down. The shift is sometimes referred to as 'recalibration in the large' [254]. A situation in which γ_0 changes is where the susceptibility of the patients to an event differs systematically between the primary and secondary datasets.

2. Re-estimate γ (but not γ_0). This is a more radical revision. Changing γ alters the shape of the baseline distribution but not its general level.
3. Re-estimate γ_0 and γ. Similar to 2 except that the general level is also changed.
4. Regarding η as a continuous covariate with a linear effect in the secondary dataset, estimate its slope, β_η. This idea, which was proposed for the logistic model by Miller [291], proposes that the effect of η is linear but may be mis-calibrated (i.e. $\beta_\eta \neq 1$). Since $\hat{\beta}_\eta$ is often less than one, β_η is sometimes described as a shrinkage factor.

These simple adjustments to the original model may produce major or minor changes in the goodness of fit (deviance). Such information may guide the interpretation of the results, and where appropriate, the choice of a 'final' updated model for the validation dataset.

Cox's model does not include a transportable estimate of the baseline survival function, making prediction of the latter from the development to the validation dataset infeasible. An alternative approach to validating a Cox model is to regard a Royston-Parmar proportional hazard model as a Cox model with a parametric baseline distribution function. The steps of the analysis are therefore:

1. Fit the Cox model to the development data and estimate the prognostic index η
2. In Royston-Parmar proportional hazard models for the development data with η offset, search for a parsimonious spline function for the baseline cumulative hazard function
3. Apply the above methods to the validation dataset

Since Cox and Royston-Parmar proportional hazard models yield almost identical parameter estimates, one could dispense with fitting the Cox model and regard validating a Royston-Parmar model as equivalent to validating a Cox model with the same covariates.

According to Royston and Parmar [290] validation is not about refitting the predictors in a postulated model on the validation dataset and comparing the estimated β's as some have done. It is about predicting relevant quantities, such as survival distributions, from a model fully specified on the development dataset to a validation dataset and examining its accuracy. Only parametric models support this approach for time-to-event data and therefore Royston and Parmar suggest that since the Cox model does not estimate the baseline distribution, it cannot be fully validated [290].

8.2.7 Conclusion

As shown by the literature review in section 8.2.1 a variety of methods are currently employed by researchers who externally validate their prognostic model(s). The most frequently reported measures encompass repeating the analysis and comparing the results, discrimination, calibration and comparing the products of the analysis, all of which have been described in sections 8.2.2 to 8.2.6.

The literature review did not, however, identify studies that have compared different methods of external validation. Therefore there is a gap in the literature with regards to guidelines for the best method of external validation or indeed which the better methods of external validation are. The simulation study presented in section 8.4 will compare the performance of methods for external validation in an attempt to address this lack of evidence.

8.3 Missing Data

It is often difficult to identify studies, independent from the data used to develop a prognostic model, which provide information on the same set of potential confounders. Consequently the development and validation datasets are unlikely to provide information on the same set of covariates. This is certainly the case for the models developed using SANAD in Chapter 4 and using MESS in Chapter 5.

There are many available strategies for handling missing data within covariates [292, 293] within individual studies. These include the simple deletion approaches of complete case analysis and available case analysis [294] and all case approaches of analysing the missing data as a separate category [295], single imputation [293] and multiple imputation [296, 297]. More complex approaches based on the use of the maximum likelihood can also be effective and practical methods for handling missing covariate data [298]. These approaches are suitable for handling small amounts of missing data within a covariate either by removing cases or by making informed decisions based on non-missing entries of the covariate. There are also methods for handling missing data in individual patient data meta-analyses [299]. There are, however, no guidelines on how to handle the situation of a covariate with every entry missing, equivalent to a missing covariate, apart from the suggestion of filling in any gaps with a constant value [8].

Section 8.3.1 considers the mechanisms of missing data before the methods of handling missing data within a covariate are summarised in section 8.3.2. Adaptations of these methods to handling covariates with every entry missing in the context of external validation are then presented in section 8.3.3.

8.3.1 Mechanisms

In 1976 Rubin [300] rigorously defined different assumptions that might plausibly be made about missing data mechanisms. He defined four assumptions: missing completely at random, missing at random, ignorable missingness and non-ignorable missingness.

8.3.1.1 Missing Completely at Random

Suppose there are missing data on a particular variable Y. The data on Y are said to be missing completely at random (MCAR) if the probability of missing data on Y is unrelated to the value of Y itself or to the values of any other variables in the dataset. When this assumption is satisfied for all variables, the set of individuals with complete data can be regarded as a simple random subsample from the original set of observations. Although MCAR is a rather strong assumption, there are times when it is reasonable, especially when data are missing as part of the research design [298].

8.3.1.2 Missing at Random

A considerably weaker assumption is that the data are missing at random (MAR). Data on Y are said to be missing at random if the probability of missing data on Y is unrelated to the value of Y, after controlling for other variables in the analysis. This means that the conditional probability of missing data on Y, given both Y and X, is equal to the probability of missing data on Y given X alone.

It is impossible to test whether the MAR condition is satisfied - because the values of the missing data are unknown it is not viable to compare the values of those with and without missing data to see if they are systematically different on that variable.

8.3.1.3 Ignorable

The missing data mechanism is said to be ignorable if the data are MAR and the parameters that govern the missing data process are unrelated to the parameters to be estimated.

Ignorability means that there is no need to model the missing data mechanism as part of the estimation process. However, special techniques are needed to utilise the data in an efficient manner [298].

8.3.1.4 Non-Ignorable

If the data are not MAR the missing data mechanism is said to be non-ignorable. In that case, usually the missing data mechanism must be modelled to get good estimates of the parameters of interest. Unfortunately, for effective estimation with non-ignorable missing data, very good prior knowledge about the nature of the missing data process usually is needed, because the data contain no information about what models would be appropriate and the results typically will be very sensitive to the choice of model.

8.3.2 Missing Data within Covariates

Currently there are many strategies for handling missing data within a covariate - complete case analysis, available case analysis, all case approach of analysing the missing data as a separate category, single imputation, multiple imputation and more complex approaches based on the use of the maximum likelihood. A summary of each of these methods is presented in sections 8.3.2.1 to 8.3.2.6.

8.3.2.1 Complete Case Analysis

Complete case analysis restricts attention to cases where all variables are present; any observations with missing values for any of the covariates are deleted. Complete case analysis assumes missingness in the covariates is not associated with the outcome [301]. It is also known as listwise deletion or casewise deletion [302].

Complete case analysis provides a statistical analysis that is fully adjusted for all potential confounders. Advantages of this approach are simplicity, since standard statistical analyses can be applied without modifications, and comparability of univariate statistics since these

are all calculated on a common sample base of cases. Additionally, complete case analysis is the method that is most robust to violations of MAR among independent variables in a regression analysis – specifically, if the probability of missing data on any of the independent variables does not depend on the values of the dependent variable then regression estimates using complete case analysis will be unbiased if all the usual assumptions of the regression model are satisfied [298]. Disadvantages stem from the potential loss of information, and resulting precision, in discarding incomplete cases. The loss in sample size can be considerable, especially if the number of missing variables is large [293]. If the data are not MCAR but only MAR complete case analysis can yield biased estimates.

8.3.2.1.1 Propensity Score Correction

An extension to the idea of complete case analysis is to weight the remaining complete cases after the incomplete cases have been removed so that their distribution more closely resembles that of the full sample or population with respect to auxiliary variables. Pugh *et al* [303] propose a propensity score correction which uses derived sampling weights to deal more appropriately with missing covariate data than blind use of only the complete cases.

The method weights observations based on their likelihood of being incomplete. The underlying idea is to reweight cases from under-represented groups; let $y_i = 1$ if individual i has no missing covariates, 0 otherwise. Logistic regression can be used to estimate each subject's probability of complete data, π_i, or 'propensity score' and in the Cox model $1/\pi_i$ can be used. A subject with complete data and a low propensity score is treated as the representative for a large cohort of similar subjects for whom data collection failed and is thus given a correspondingly large weight. It need not improve on complete case analysis and may do worse - the relative performance of the two approaches depends on the mechanism leading to the missing data [175].

If the probability of a particular covariate entry being missing depends on the possibly unobserved value of the covariate, but not on other covariates or data, then the complete case approach is consistent but the propensity score method need not be [175]. If the probability of missingness depends on the values of completely observed variables, but not on the value of the missing covariate itself then Pugh's method is consistent, but complete case analysis need not be so [175]. If missingness depends on both of these factors both methods may be inconsistent. The completeness propensity score approach is valid when the probability of being missing depends only on the values of observed data, response, or prediction, and not on the values of missing covariates [304].

8.3.2.2 Available Case Analysis

For univariate analyses a natural alternative is to include all cases with complete data for the variables in the fitted model, an option termed available case analysis, also known as pairwise deletion. The idea of available case analysis is to compute each required summary statistic using all the cases that are available. Once the summary measures have been computed they can be used to calculate the parameters of interest such as regression coefficients. Although missing data may cause different observations to be used in calculating different statistics, the missing data do not cause observations to be omitted from the entire analysis as in complete case analysis. For example, in a sample consisting of 10 observations, the correlation between variables X and Y may be computed only by using observations one to five because X is missing for observations six to 10, whereas the correlation between variables Y and Z may be calculated only by using observations six to 10 because Z is missing for observations one to five.

The main problem with pairwise deletion is that the estimated standard errors and test statistics produced by conventional software are biased [298].

8.3.2.3 Additional Category or Missing Indicator Method

In some situations missing data may be a valid response and consequently a larger proportion of responses may be missing than expected. For example, the variable for the results of a clinical test could have normal or abnormal as responses for patients who had the test. Some patients however may not have had the test because it was deemed not to be relevant. Therefore a third category can be introduced for 'not done' or 'missing'. This ensures all cases can be analysed and takes account of the validly missing data.

A variation on this method is to create a dummy variable D that is equal to 1 if data are missing on variable X and equal to 0 otherwise. A new variable X^* can be created such that

$$X^* = \begin{cases} X \text{ when data are not missing} \\ c \text{ when data are missing} \end{cases}$$

where c can be any constant. The dependent variable Y can be regressed on X^*, D, and any other variables in the intended model. This technique, known as dummy variable adjustment or the missing-indicator method, can be extended easily to the case of more than one independent variable with missing data. The apparent virtue of the dummy variable adjustment method is that it uses all the information that is available about the missing data [298]. Unfortunately this method generally produces biased estimates of the coefficients, as proven by Jones [305].

8.3.2.4 Imputation

Except in special circumstances that usually involve only very simple models, the primary alternative to deleting incomplete observations is imputation of the missing values. The basic idea is to substitute some reasonable guess, imputation, for each missing value and then proceed to do the analysis as if there were no missing data. Imputation is a general

and flexible method for handling missing data problems, however, it is not without issues [293, 306].

Decisions need to be made regarding which information to use in computing imputed values for missing data. They can be filled in by sampling non-missing values of the variable, or by using a constant such as the median or mean non-missing value. Imputation algorithms can be based only on external information not otherwise used in the model or they can be derived by only analysing inter-relationships between the predictor variables. Imputations can also use relationships among the predictor and outcome variables and they can take into account the reason for non-response if known [139].

8.3.2.4.1 Single Imputation

In single imputation only one estimate, such as a mean, is imputed and then the resulting data is analysed as if there were no missing data. Single imputation is simple, reduces bias due to non-response compared to complete and available case analyses, and improves precision but it underestimates the standard error and can lead to incorrect p-values. Several variations of single imputation exist: unconditional mean imputation, conditional mean imputation and hot deck imputation.

8.3.2.4.1.1 Unconditional Mean Imputation

A particularly simple form of imputation is to estimate missing values y_{ij} by $\bar{y}_j^{(j)}$, the mean of the recorded values of Y_j. In mean substitution missing values are replaced by the average of the observed values for that item. The average is preserved but other aspects of its distribution are altered with potentially serious consequences [292].

Unconditional mean imputation is simple to implement for any type of variable and in addition, once missing values are imputed and incorporated into the dataset, multiple users can use the data with consistent results. However, one value cannot reflect sampling

variability and marginal distributions and associations are distorted because there is no residual variance after the imputation [293].

8.3.2.4.1.2 Conditional Mean Imputation

A more promising form of imputation is to substitute means that are conditioned on the variables recorded in an incomplete case. This method, proposed by Buck [307], first estimates the mean and covariance from the multivariate normal distribution, from the sample mean and covariance matrix based on the complete cases. These estimates are then used to calculate the linear regressions of the missing variables on the present variables, case by case. Substituting the observed values of the present variables for a case in the regressions yields predictions for the missing values in that case.

In the simple case a multiple regression could be estimated with several independent variables. One of these variables, X, may have missing data for some of the cases. For those cases with complete data, it is possible to regress X on all other independent variables. Using the estimated equation, predicted values for the cases with missing data on X can be generated which are substituted for the missing data and the analysis proceeds as if there were no missing data [298].

Buck's method [307] projects the incomplete cases to the regression line, a process that makes the assumption that the regression is linear. This assumption is particularly tenuous if the imputation involves extrapolation beyond the range of the complete data. However, the filled-in data from Buck's method [307] yields reasonable estimates of means, particularly if the normality assumptions are plausible. The sample covariance matrix from the filled-in data underestimates the sizes of variances and covariances, although the extent of underestimation is less than that obtained when unconditional means are substituted.

Although the regression based imputations in Buck's method [307] appear to require that the variables $Y_1, \ldots, Y_k$ are interval scaled, the method can be applied to categorical variables by replacing each of them by a set of dummy variables, numbering one less than the number of categories. If a categorical variable is completely observed, then the dummy variables always appear as independent variables in the regressions for Buck's method and no problems arise. If it is sometimes missing, then the set of dummy variables also appear as dependent variables in linear regressions. The imputations from the regressions are linear estimates of the probability of falling into the categories represented by each of the dummy variables. Problems can arise from the fact that linear regression is used to predict these probabilities; for example, the predicted values can lie outside the range (0,1). Thus Buck's method has limitations when some of the missing variables are categorical [293].

8.3.2.4.1.3 Hot Deck Imputation

Another example of a simple imputation technique is the hot deck method in which a single value is imputed for each missing observation. This technique selects a value at random, usually with replacement from a donor pool of complete cases, and the missing value is then replaced by the selected value [308]. This strategy introduces variation into the analysis consistent with the range of possible values seen in the complete data [309]. The matching process can be carried out using so-called filter variables. The records match if they have the same values on the filter variables [310].

The main advantage of this non-parametric technique is that it does not require strong distributional assumptions or careful modelling to develop selection criteria for imputing a value – in fact it preserves the distribution of item values and permits the use of the sample weight for all items. Additionally, the results obtained from different analyses are consistent with one another [310]. However the disadvantages of the method are that there are no clear criteria to guide the selection of the donor set of complete cases, and,

although the variability of items is not distorted, it still distorts correlations and other methods of association [309]. In addition, the predictor variable must all be categorical, or treated as such, which imposes serious limitations on the number of possible predictor variables [310].

8.3.2.4.2 Multiple Imputation

The key idea of the multiple imputation procedure is to replace each missing value with a set of M plausible values. Each value is a Bayesian draw from the conditional distribution of the missing observation given the observed data [311]. The imputations produce M 'completed' datasets, each of which is analysed using the method that would have been appropriate had the data been complete. The model for the latter analysis is called the substantive model, while that used to produce the imputations is called the imputation model. A strength of the multiple imputation procedure is that, to a certain extent, the two models can be considered separately.

Multiple imputation is most straightforward to use under MAR, and most software implementations make this assumption. However, it is quite possible to apply it in missing not at random settings, and this is particularly convenient when certain classes of pattern-mixture models are used to construct the imputation model. Multiple imputation involves three distinct phases [296]:

1. The missing values are filled in M times to generate M complete datasets
2. The M complete datasets are analysed by using standard procedures
3. The results from the M analyses are combined into a single inference

Suppose that interest lies in making inferences about the $k \times 1$ parameter vector $\boldsymbol{\beta}$ from the substantive model and that it is possible to make appropriate Bayesian posterior draws from the imputation model. Replacing the missing data by their corresponding imputation samples, M completed datasets are constructed. Denote by $\widehat{\boldsymbol{\beta}}^m$ and V^m respectively the

estimate of $\boldsymbol{\beta}$ and its covariance matrix from the mth completed dataset ($m = 1, \ldots, M$). The multiple imputation estimate of $\boldsymbol{\beta}$ is the simple average of the estimates,

$$\widehat{\boldsymbol{\beta}}^{*} = \frac{1}{M}\sum_{m=1}^{M} \widehat{\boldsymbol{\beta}}^{m}.$$

The advantages of multiple imputation are that a single set of imputed datasets can be used for a variety of analyses and it accounts for missing data uncertainty. However, it is cumbersome to use because of the need to analyse multiple datasets and combine the results to make one overall inference.

8.3.2.5 Maximum Likelihood Algorithms

Theory-based maximum likelihood approaches for dealing with missing data have been known in the technical literature for some time [306, 312, 313]. As the methods require the less restrictive MAR assumption, unbiased parameter estimates should result under both MCAR and MAR [314]. In addition maximum likelihood methods should yield more efficient estimates than listwise and pairwise deletion under MCAR. In theory, likelihood methods are more attractive than ad hoc techniques of case deletion and single imputation [292]. However, they still rest on a few crucial assumptions - the sample is large enough for maximum likelihood estimates to be approximately unbiased and normally distributed, and the likelihood function comes from an assumed parametric model for the complete data. Depending on the application, likelihood methods may or may not be robust to departures from model assumptions.

Intuitively the missing data is 'filled in' with a best guess at what it might be under the current estimate of the unknown parameters, then the parameters are re-estimated from the observed and filled-in data. In a formal sense there is no difference between the maximum likelihood estimation for incomplete data and maximum likelihood estimation for complete data: the likelihood for the parameters based on the incomplete data is

derived and the maximum likelihood estimates are found by solving the likelihood equation [293].

8.3.2.6 The Expectation-Maximisation Algorithm

The expectation-maximisation (EM) algorithm is a convenient and widely applicable computational technique that can be used when the observed data likelihood is awkward and/or difficult to compute. It is a general-purpose iterative algorithm for calculating maximum likelihood estimates in parametric models for incomplete data. The condition for the EM algorithm to be valid, in its basic form, is ignorability, and hence MAR. Within each iteration of the EM algorithm there are two steps called the expectation step, or E step, and the maximisation step, or M step.

The fundamental idea behind the EM algorithm is to associate with the given incomplete data problem a complete data problem for which maximum likelihood estimation is computationally easier. Starting from suitable initial parameter values, the E and M steps are repeated until convergence. Given a set of parameter estimates, such as the mean vector and covariance matrix for a multivariate normal setting, the E step calculates the conditional expectation estimates. This step often reduces to calculating simple sufficient statistics. Given the complete data log-likelihood, the M step then finds the parameter estimates to maximise the complete data log-likelihood from the E step.

The fact that the EM algorithm is guaranteed to converge to a, possibly local, maximum is a great advantage. However a disadvantage is that this convergence is slow and that precision estimates are not automatically provided.

8.3.3 Methods for Handling Missing Covariates

Within an individual study missing covariates have not been considered as there is no data to base the imputation (or similar) on. However, in the case of external validation, extra

data is available that may inform 'replacement' of the covariate. It may be possible to adapt some of the existing methods for handling missing data within covariates to the situation of every entry missing. The situation of a missing covariate may reasonably satisfy the MCAR assumption if missingness is not related to anything within the study design – there may be examples where this is not the case for example if the reason for the missing data was related to severity of disease, say if a CT scan was not indicated. In these cases further consideration would be needed to decide which methods of imputation were valid.

Of the standard methods for handling missing data within a covariate outlined in section 8.3.2, available case analysis, additional category for missing responses, mean imputation and maximum likelihood methods are not adaptable to covariates with every entry missing. In the case of available case analysis this is because all cases would be removed while in the case of an additional category for missing responses it would not be possible to fit a Cox model as the variable with the additional category would have only one category. Mean imputation is not possible as there is no data to estimate the mean from and if the mean from the relevant variable within the development data was imputed, all cases would have the same response and therefore the Cox model could not be fit. Finally, for maximum likelihood methods, the joint distribution of the observed and missing covariates is required but this is unobtainable in the situation of a missing covariate.

Five adaptations of existing methods for handling missing data within a covariate to the situation of handling a covariate missing from the validation dataset are (1) remodelling with variable matching; (2) random selection with replacement; (3) single imputation via estimated proportions; (4) hot deck imputation; (5) random selection with replacement multiple times. These methods will be outlined in sections 8.3.3.1 to 8.3.3.5.

8.3.3.1 Remodelling with Variable Matching

The extension of complete case analysis to external validation involves fitting the model on a reduced covariate set. To externally validate a prognostic model using a validation dataset missing at least one covariate, the model is refit with the development dataset restricted to include only covariates that are available in the validation dataset. This is similar to Mekhail's [273] approach for handling missing covariates by remodelling the data on the reduced dataset as described in the literature review of section 8.2.1.

There are limitations to this approach in terms of how many variables are missing - if the validation dataset has no variables in common with the development data then, obviously, this approach is unsuitable. This is, however, also true for the other imputation methods.

8.3.3.2 Random Selection with Replacement

In random selection with replacement, each entry of the missing variable is imputed by randomly selecting an entry from the equivalent variable in the development dataset. Random selection with replacement is used to ensure that the imputed variable in the validation dataset is not just a re-ordered copy of the variable from the development dataset.

8.3.3.3 Single Imputation via Estimated Proportions

In single imputation via estimated proportions each entry of the missing variable is imputed with a random number. The imputed random number is based on the summary statistic(s) of the equivalent variable in the development dataset. For example, if the missing variable in the validation dataset is a binary variable, then the proportion of positive responses in the equivalent variable within the development dataset is estimated. The missing variable is then imputed with zeros and ones generated according to these proportions. In the case of continuous variables, the mean and standard deviation are estimated from the development dataset and these estimates are used to obtain values that can be imputed

into the validation dataset. Finally, for categorical variables the percentage of each response is used to inform the imputation. Essentially we are assuming that the missing values are random draws from a normal distribution with mean and standard deviation estimated from the development dataset.

8.3.3.4 Hot Deck Imputation

In the case of a missing covariate there is no way to match a missing with a non-missing respondent. The only plausible alternative is to impute the missing covariate values with values recorded for similar respondents in the development data. In our simulation study (section 8.4), similar respondents have matched binary, categorical and continuous variables. Continuous variables are matched to the nearest respondent based on the smallest mean square error between the continuous variable in the development dataset and the equivalent continuous variable in the validation dataset.

8.3.3.5 Multiple Imputation

The conditional distribution of the missing observation is not known in the case of a variable with every entry missing. However a variation is to randomly select, with replacement, values of the missing covariate from the development set to produce M datasets. In the first step the validation data is used to fit the Cox model except for the missing covariate which is substituted with data from the relevant covariate within the development data, using random selection with replacement. This is repeated M times. The outcomes are calculated as usual and finally, results from all the analyses are combined to produce a mean outcome across the repetitions. Rubin [296] claimed that only three to 10 imputations may be needed. He showed that the efficiency of an estimate based on m imputations is approximately

$$\left(1+\frac{\gamma}{m}\right)^{-1}$$

where γ is the rate of missing information for the quantity being estimated. The efficiencies achieved for various values of *m* and rates of missing information are shown in Table 23.

Table 23: Efficiencies achieved for various numbers of imputations and rates of missing information

	γ				
m	0.1	0.3	0.5	0.7	0.9
3	97	91	86	81	77
5	98	94	91	88	85
10	99	97	95	93	92
20	100	99	98	97	96

8.3.3.6 Conclusion

In the literature review (section 8.2.1) we found only one study where there was an issue with a covariate missing from the validation dataset. This was because nearly every study had collected a validation dataset specifically with the intention of externally validating either a previously published prognostic model, or validating a prognostic model they had developed.

As observed by Royston, of the many hundreds of prognostic models published each year, only very few are externally validated [274]. Perhaps a reason for this is a lack of funding to obtain a validation dataset with covariates that match those in the prognostic model. Plausibly related datasets are already available within the field of epilepsy however, and I imagine the same applies to many other medical fields. Unfortunately, many of them are missing covariates found to be significant in a prognostic model. Therefore methods for handling missing covariates in the context of external validation need to be found. In the simulation study of section 8.4 the performance of methods for handling a covariate missing from the validation dataset are compared to potentially address this issue.

8.4 Simulation Study

The simulations were conducted using the *R* statistical programming language versions 2.9.2 and 2.14.1 [182].

As described in Bonnett [315] and in Chapter 4, the SANAD dataset has recently been used to determine if any clinical factors influence outcome and to develop prognostic models. These prognostic models should be validated in plausibly related validation datasets. Unfortunately no other datasets that are similar to SANAD exist. The best match is a set of individual participant data that has been collected [192]. However these data do not include all the covariates that were found to be significant in the multivariable model. Therefore, it is of particular relevance to see how data, based on SANAD, can be validated using independent data with missing covariates.

As shown in the literature review, in section 8.2.1, various methods of external validation exist however there is no evidence to suggest which method is most appropriate. Additionally, methods for handling covariates missing from the validation data do not currently exist. To determine which methods of external validation are appropriate and to investigate strategies for handling covariates missing from the validation dataset, a simulation study was undertaken.

8.4.1 Methods Under Consideration

The most common methods of external validation listed in the literature were repetition of the complete analysis (53%), discrimination (41%), calibration (39%) and comparison of a product of the analysis (35%). Repetition of the analysis involves re-running the model selection process in the validation dataset. Given a dataset it is possible to fit thousands of models and therefore different researchers might discover or learn different things from the same dataset [316]. Additionally, it is possible to gear model building and model selection efforts towards criteria that favour good performance [316]. Therefore repeating

the analysis may lead to biased results as researchers could engineer the model selection process to produce a multivariable model that closely matches that developed using the development model. For this reason, and together with the time-consuming nature of repeating the analysis, repetition of the analysis will not be considered as a method of external validation in the simulation study.

Instead, the other popular methods of external validation will be considered – comparison of discrimination, comparison of calibration and a comparison of products of the analysis. We will consider discrimination via comparison of concordance (Harrell's c or equivalently Somers' D) statistics, calibration via comparison of deviance statistics and comparison of a product of the analysis via a measure of prognostic separation.

As shown in the literature review in section 8.2.1, very few prognostic models are externally validated. In clinical medicine, the focus of this thesis, the Cox proportional hazards model is a near-universal tool in the analysis of prognostic models [287]. Therefore for our purpose it is most appropriate to focus on external validation of Cox models although it would be interesting to extend the simulation study to Royston-Parmar's alternative in the future.

The five methods of handling covariates missing from the validation dataset under consideration in the simulation study are (1) remodelling with variable matching; (2) random selection with replacement; (3) single imputation via estimated proportions; (4) hot deck imputation; and (5) random selection with replacement multiple times, all of which were described in section 8.3.3.

8.4.2 Simulation Procedures – Data Generation

Data used to derive the data-generation process for the simulations consisted of 1644 patients recruited to the SANAD study [180], although development datasets of size 1500

were simulated for simplicity. Time to first seizure after randomisation was the chosen outcome variable as this was the simplest time to event outcome of the SANAD study.

Sections 8.4.2.1 through to 8.4.2.3 describe how the data were simulated.

8.4.2.1 Survival Data

It is common to model survival times through the hazard function. As the Cox proportional hazards model is formulated through the hazard function, the simulation of appropriate survival times is not straightforward [317] - the effect of the covariates have to be translated from the hazard to the survival times because the usual software packages for Cox models require the individual survival time data, not the hazard function. The translation of the regression coefficients from hazard to survival time is simple if the baseline hazard function is constant i.e. the survival times are exponentially distributed [317].

Among the commonly used survival time distributions, only the exponential, Weibull and Gompertz distributions share the assumption of proportional hazards with the Cox model as described in section 1.4.2. The survival function of the Cox model is given by $S(t|x) = \exp[-H_0(t)\exp(\beta' x)]$ where $H_0(t) = \int_0^t h_0(u)\,\mathrm{d}u$ is the cumulative baseline hazard function [30]. Thus the distribution function of the Cox model is $F(t|x) = 1 - \exp[-H_0(t)\exp(\beta' x)]$.

Let Y be a random variable and let the distribution function, F, of Y be constant. Then $U = F(Y)$ follows a uniform distribution on the interval from 0 to 1 [318]. Then let T be the survival time of the Cox model, $h(t|x) = h_0(t)\exp(\beta' x)$.

It therefore follows that $1 - U = \exp[-H_0(t)\exp(\beta' x)] \sim U[0,1]$ and, as both $U \sim U[0,1]$ and $1 - U \sim U[0,1]$ according to the properties of a standard uniform distribution, $U = \exp[-H_0(t)\exp(\beta' x)] \sim U[0,1]$.

If $h_0(t) > 0$ for all t, then H_0 can be inverted and the survival time, T, of the Cox model can be expressed as Equation 19 by taking the logarithm of both sides of the equation, dividing through by $-\exp(\beta' x)$ and inverting $H_0(T)$.

Equation 19

$$T = H_0^{-1}[-\log(U)\exp(-\beta' x)]$$

where U is a random variable with $U \sim U[0,1]$. By applying Equation 19, uniformly distributed random numbers can be transformed into survival times following a specific Cox model. An appropriate cumulative baseline hazard function, such as the exponential, Weibull or Gompertz, then just needs to be inserted into the formula.

Table 24 summarises the survival time and hazard function equations for these three distributions where U is a variable following a uniform distribution on the interval from 0 to 1. By inserting the inverse of the cumulative hazard function for the exponential distribution, $H_0^{-1}(t) = \lambda^{-1}t$, into Equation 19 we transformed uniformly distributed random numbers into survival times following a Cox-exponential model [317]. The Cox-exponential distribution was used for simplicity in our simulation study.

Table 24: Survival time and hazard function equations for the Cox-exponential, Cox-Weibull and Cox-Gompertz distributions

	Model		
Characteristic	Cox-exponential	Cox-Weibull	Cox-Gompertz
Survival time	$T = -\frac{\log(U)}{\lambda\exp(\beta' x)}$	$T = \left(-\frac{\log(U)}{\lambda\exp(\beta' x)}\right)^{1/\nu}$	$T = \frac{1}{\alpha}\log\left[1 - \frac{\alpha log(U)}{\lambda\exp(\beta' x)}\right]$
Hazard function	$h(t\|x) = \lambda\exp(\beta' x)$	$h(t\|x) = \lambda\exp(\beta' x)\nu t^{\nu-1}$	$h(t) = \lambda\exp(\beta' x)\exp(\alpha t)$

8.4.2.2 Censoring Indicator

A censoring indicator is a binary variable which follows a binomial distribution. In this case p is the probability of being censored as a result of either loss to follow-up or study completion without an event. Any survival times equal to or in excess of the maximum follow-up time were censored.

A patient randomised to SANAD could potentially have been followed-up for a maximum of seven years. In the case of time to first seizure the longest follow-up occurred approximately 6.5 years into the trial. Therefore, the maximum possible simulated survival time was 6.5 years. Any survival times equal to or in excess of this time were censored. For all other patients, censoring was in the same proportions as in SANAD.

For each case, if the uncensored survival time for a case was less than or equal to the censored time, then the event was considered to be observed and the survival time equalled the uncensored survival time. Otherwise the event was considered censored and the survival time equalled the censored time.

8.4.2.3 Covariate Data

We assumed that there were three predictor variables. We selected the following variables from the SANAD dataset: gender, age and epilepsy type. For simplicity, covariate data was generated using an underlying multivariate normal distribution [319] with means and covariance matrix as per the SANAD data (Table 25). Age was normally distributed so no transformation was required to make the assumption of normality more appropriate. However, in the clinical setting, continuous variables are often considered to follow the log normal distribution [320]. Therefore, the generated covariate data for the continuous variable was transformed onto the log scale prior to analysis. The continuous covariate (simulated age) was truncated using the upper and lower observed limits of 86 and five

respectively to produce realistic values and reasonable estimates for the mean and standard deviations [321].

Binary and categorical variables were generated as latent normal variables – they were generated as continuous variables and then a cut-point of 0.5 was used to obtain the binary covariate and cut-points of 0.33 and 0.66 were used for the three-level categorical covariate [321]. Two dummy variables were then created accordingly.

Table 25: Data structure for the SANAD dataset and associated means, standard deviations (SDs) and variance-covariance matrix

				Variance-Covariance Matrix		
Covariate	Variable Type	Groupings/ Measurement	Mean (SD)	Gender	Epilepsy Type	Age
Gender	Binary	0 = female 1 = male	0.55 (0.50)	0.25	0.01	-0.15
Epilepsy Type	Categorical	0 = partial 1 = generalised 2 = unclassified	1.21 (0.60)	0.01	0.35	-0.26
Age	Continuous	Years	38.30 (18.34)	-0.15	-0.26	336.36

8.4.3 Simulation Procedures - Scenarios to be Investigated

There are a large number of variations that could be tested, irrespective of the aim - number and combination of covariates present and missing; types of covariates present and missing; size of the development and validation datasets; how closely the 'super-population' assumption is adhered to; length of follow-up etc. We varied the following factors in our simulations: the size of the validation dataset and the missing variable. During the literature review (section 8.2.1), data was collected on the sample size of both the validation and development datasets. The size of the validation dataset could be considered as a percentage of the size of the development dataset. Of those articles that externally validated a Cox model, the 25th percentile of these values was 57% while the 75th percentile was 115%. Therefore, the two sample sizes we considered were 855 and 1725.

With regards to the missing variable, none or one of the binary, continuous and categorical variables were forced to be missing in turn.

The scenarios examined are shown in Table 26 with scenarios 1 and 5 testing the methods of external validation and 2, 3, 4, 6, 7 and 8 testing the methods of imputation. The missing data methods have been compared by comparing each method of external validation (concordance, deviance and prognostic separation) from the validation dataset with missing data to the corresponding method of external validation from the validation dataset without missing data.

Table 26: Simulation Strategies used to consider methods of external validation and methods to handle a covariate missing from the validation dataset

✓ implies the variable is present in the dataset; × implies that the variable is missing from the dataset

Scenario	Dataset to be generated	Number of subjects	Variable present?		
			Binary	Continuous	Categorical
-	Development	1500	✓	✓	✓
1	Validation	855	✓	✓	✓
2	Validation with missing data	855	✓	✓	×
3	Validation with missing data	855	✓	×	✓
4	Validation with missing data	855	×	✓	✓
5	Validation	1725	✓	✓	✓
6	Validation with missing data	1725	✓	✓	×
7	Validation with missing data	1725	✓	×	✓
8	Validation with missing data	1725	×	✓	✓

8.4.4 Simulation Procedures – Statistical Methods

Once a development dataset and a fully independent validation dataset had been randomly generated a Cox model, adjusted for the three simulated covariates, was fitted to each. The prognostic measure of separation for each model was obtained together with the deviance, concordance statistic, regression coefficients and respective standard errors. This process was then repeated 1400 times for each possible combination of size of

validation dataset (855 and 1725) and missing data (none, missing binary, missing categorical, missing continuous). The number of simulations was based on the accuracy of the regression coefficients and was calculated using the following formula:

$$B = \left(\frac{Z_{1-(\alpha/2)}\sigma}{\delta}\right)^2$$

where δ was the specified level of accuracy of the estimate that we were willing to accept, $Z_{1-(\alpha/2)}$ was the $1-(\alpha/2)$ quantile of the standard normal distribution and σ^2 was the variance from the parameter of interest [322]. In our case, a 5% accuracy of the true coefficient was required and $Z_{1-(\alpha/2)} = 1.96$ therefore 1400 simulations were run.

Within each scenario we computed: the proportion of the 1400 differences in prognostic separation between the development and validation datasets within 5% tolerability limits of zero, the proportion of the 1400 p-values for the difference in deviance between the development and validation datasets within 5% of zero (indicating that the models are not externally valid) and the proportion of the 1400 differences in concordance between the development and validation datasets within 5% of zero.

To summarise the estimates once all simulations had been performed, the average regression coefficient over the simulations was calculated as a true estimate of interest where B was the number of simulations and $\widehat{\beta_i}$ was the regression coefficient for simulation i.

$$\tilde{\beta} = \sum_{i=1}^{B} \frac{\widehat{\beta_i}}{B}$$

As the simulation study had been designed to mimic the results that could have been obtained from a single study, an assessment of the uncertainty in the regression coefficient

between simulations, denoted $SE(\tilde{\beta})$, was the empirical standard error calculated as the standard deviation of the regression coefficients from all simulations.

$$\sqrt{[1/(B-1)]\sum_{i=1}^{B}\left(\hat{\beta}_i - \tilde{\beta}\right)^2}$$

Standardised bias was calculated as per Equation 20. A standardised bias of greater than 40% in either direction has been shown [323] to have noticeable adverse impact on the efficiency, coverage and error rate. Therefore, a result of greater than 40% in either direction was taken as a poor result.

Equation 20

$$\left(\frac{\tilde{\beta} - \beta}{SE(\tilde{\beta})}\right) \times 100$$

Mean square error (MSE) was calculated as per Equation 21. It provided a useful measure of the overall accuracy because it incorporated both measures of bias and variability and was calculated for each method. A mean square error of 0 suggests the estimator $\tilde{\beta}$ predicts observations of the parameter β with perfect accuracy [319].

Equation 21

$$\left(\tilde{\beta} - \beta\right)^2 + \left(SE(\tilde{\beta})\right)^2$$

The coverage of a confidence interval is the proportion of confidence intervals that contain the true specified parameter value [319]. The coverage should be approximately equal to the nominal coverage rate, usually taken as 95%. Coverage above 95% suggests that the estimates are too conservative as more simulations will not find a significant result when there is a true effect leading to a loss of statistical power with too many type II errors. In

contrast, coverage below 95% is unacceptable as it indicates inappropriate over-confidence in the estimates since more simulations will incorrectly detect a significant result which leads to higher than expected type I errors.

The coverage should not fall outside of approximately two standard errors of the nominal coverage probability, p where

$$SE(p) = \sqrt{p(1-p)/B}.$$

So, if 95% confidence intervals are calculated using 1400 independent simulations then $SE(\hat{p}) = 0.005825$ and hence between 1314 and 1346 of the confidence intervals should include the true regression coefficient value.

8.4.5 Results

Each method should conclude that the model developed using the development dataset was externally validated by the validation dataset due to the way the data were simulated. Any method which did not reach this conclusion was deemed to be a poor method of external validation. However, a dataset could be chosen that differs from the development dataset just by chance hence bootstrap resampling has been used.

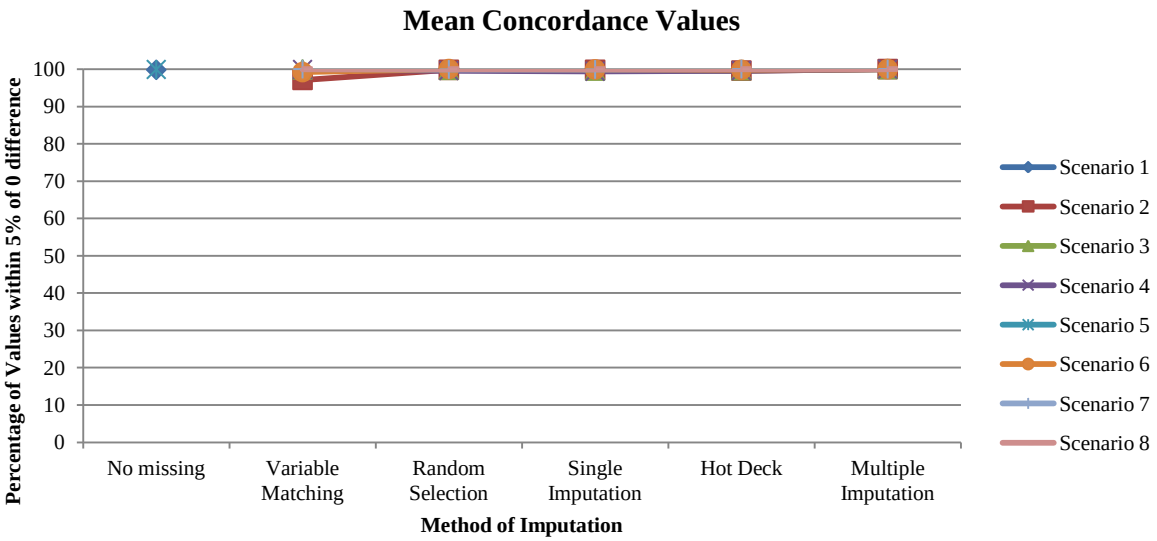
To enable comparison across methods sensitivity and specificity could be calculated. However, this would require further simulation settings so that we could capture situations where the 'truth' is that the model is not externally valid. This is obviously virtually impossible to define and so this simulation study is liable to criticism as we have only assessed performance for cases where the data are related. In pre-emption of this, external validation hinges on satisfying the condition that datasets have come from a super-population. There will, of course, be different degrees of how well this is satisfied in practice and so an extension to this simulation study would consider varying degrees of adherence to the super-population condition.

In the case of the concordance and prognostic separation methods the model would be externally valid if the difference between the development and validation datasets was zero hence the proportion of simulations within 5% tolerability limits of zero have been calculated. In the case of the deviance method, a p-value for the comparison of deviance between the development and validation dataset would be less than 0.05 if the model is not externally valid, although some comparisons would be significant just by chance.

8.4.5.1 Concordance

For ease of interpretation results are presented graphically although numerical results are available in Appendix C. Only the results for the 5% tolerability limits are presented here - the 10% results can also be found in Appendix C. Figure 25 shows the mean concordance values over 1400 simulations. The concordance statistic shows almost perfect agreement between the validation and development datasets for the case of no missing data and for each method of imputation (mean development c-statistic variable matching: 0.55; mean validation c-statistic variable matching: 0.55; mean development c-statistic otherwise: 0.56; mean validation c-statistic otherwise: 0.56). The poorest, although still good, agreement is in the case of variable matching where 97.1% of differences in concordance values between the development and the validation dataset are within 5% of zero difference suggesting that the model is externally valid 97% of the time.

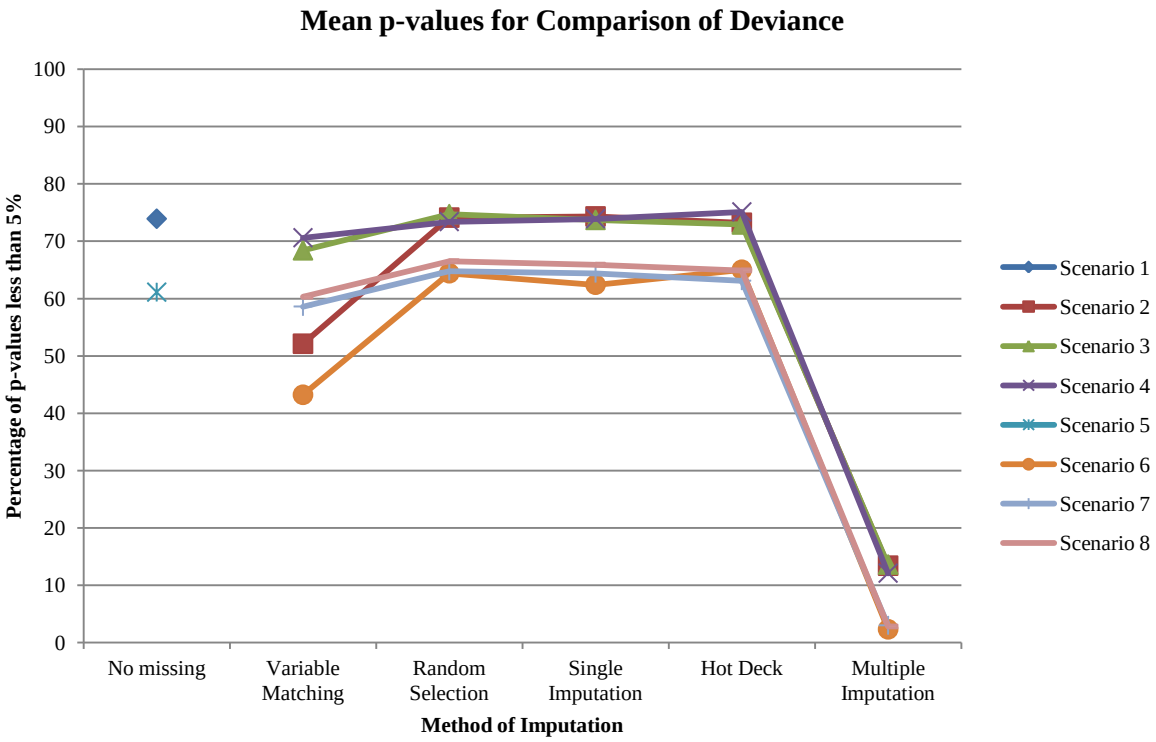
Figure 25: Mean Concordance Values from simulation study examining methods of external validation and methods of handling a covariate missing from the validation dataset



8.4.5.2 Deviance

Figure 26 shows the proportion of p-values less than 5% for the comparison of deviances between the development and validation datasets over 1400 simulations – a p-value less than 5% implies that the model is not externally valid. The deviance statistic seems to suggest that although there is agreement between the development and validation models, it is weak agreement which should not be the case given the method of data simulation. Results for validation datasets which are larger than the development ones are slightly better than for those where the validation dataset is smaller than the development dataset irrespective of the method of imputation. In the case of multiple imputation (random selection with replacement multiple times) only 13.4% of models had a significant p-value (at the 5% level) and were therefore not externally valid. This method of imputation has performed much better than the other four methods – in the case of variable matching at least 43% of models had a significant p-value. For random selection with replacement, single imputation via estimation and hot deck imputation at least 62% of models had a significant p-value and therefore were not externally valid.

Figure 26: Mean p-values for Comparison of Deviance from simulation study examining methods of external validation and methods of handling a covariate missing from the validation dataset

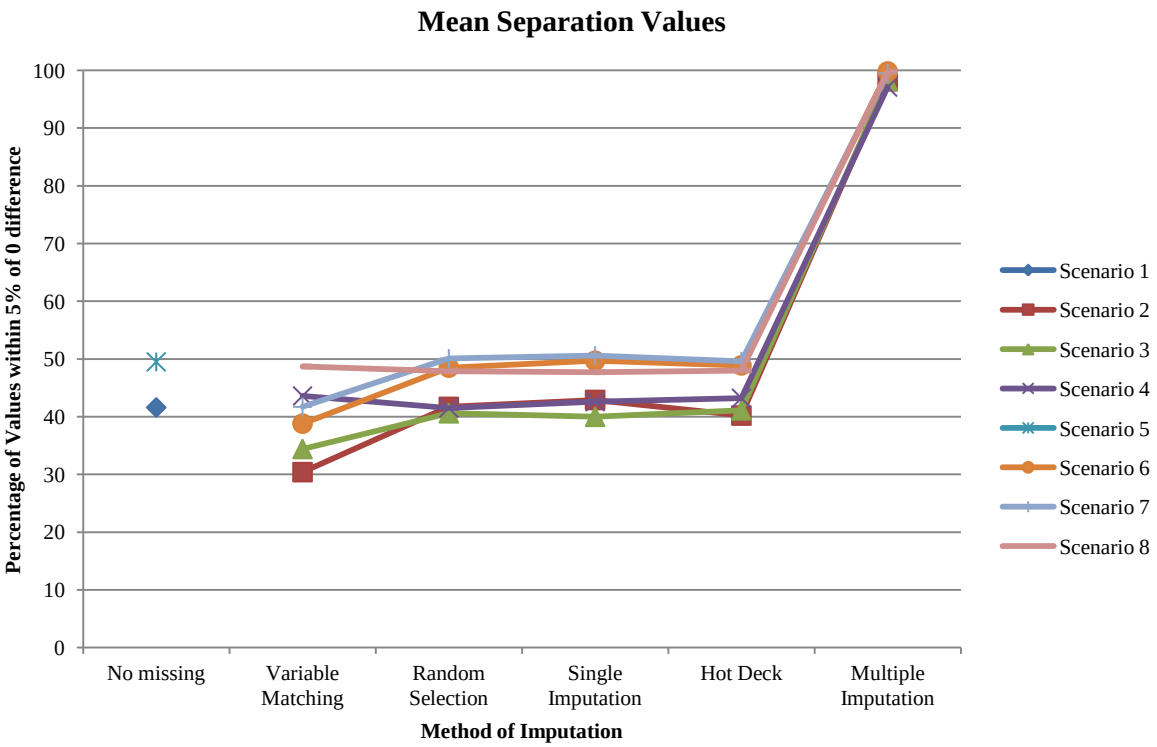


The simulations without missing data gave poorer results than the simulations with missing data. A reason for this is that by using imputation methods, or removing a covariate in the case of variable matching, the validation dataset is a closer match to the development dataset by the nature of the imputation than two randomly generated datasets – when the validation dataset was forced to be identical to the development dataset the method lead to 100% agreement as expected. An additional reason may be the possible violation of the assumption required for the distribution of the deviance difference - the result that the deviance difference is distributed as χ^2 with degrees of freedom equal to the number of re-estimated parameters depends on the independence of the parameters, details of which are explained in section 8.2.4.

8.4.5.3 Prognostic Separation

The mean difference in prognostic separation between the development and validation datasets over 1400 simulations can be seen in Figure 27. Generally the prognostic separation measure showed poor agreement between the validation and development datasets. Validation datasets which were larger than the development datasets tended to show slightly better agreement. In the case of multiple imputation the prognostic separation measure showed near perfect agreement between the validation and development datasets especially with validation datasets which were larger than the development datasets (over 97.1% of mean differences within 5% of 0 for the small validation sets and over 99.5% of difference within 5% of 0 for large validation sets).

Figure 27: Mean Separation Values from simulation study examining methods of external validation and methods of handling a covariate missing from the validation dataset



8.4.5.4 Measures of Accuracy

Average coefficient values, $\tilde{\beta}$, were similar for the development and validation dataset as shown by Figure 28. Standard error values were also similar (not presented – see Appendix C). In Figure 28 values presented are the mean coefficient values over all coefficients per scenario per method.

Figure 29 shows the mean bias results, taken over all coefficients. In most cases standardised bias was very small and certainly within the recommended 40% limits suggesting that bias was not an issue. Only in the situation of variable matching does bias become an issue, especially when a categorical variable is the variable missing from the validation dataset. In this case some individual biases exceed 40%. However, the mean results are within the 40% boundary.

Figure 28: Mean Coefficient Values from simulation study examining methods of external validation and methods of handling a covariate missing from the validation dataset

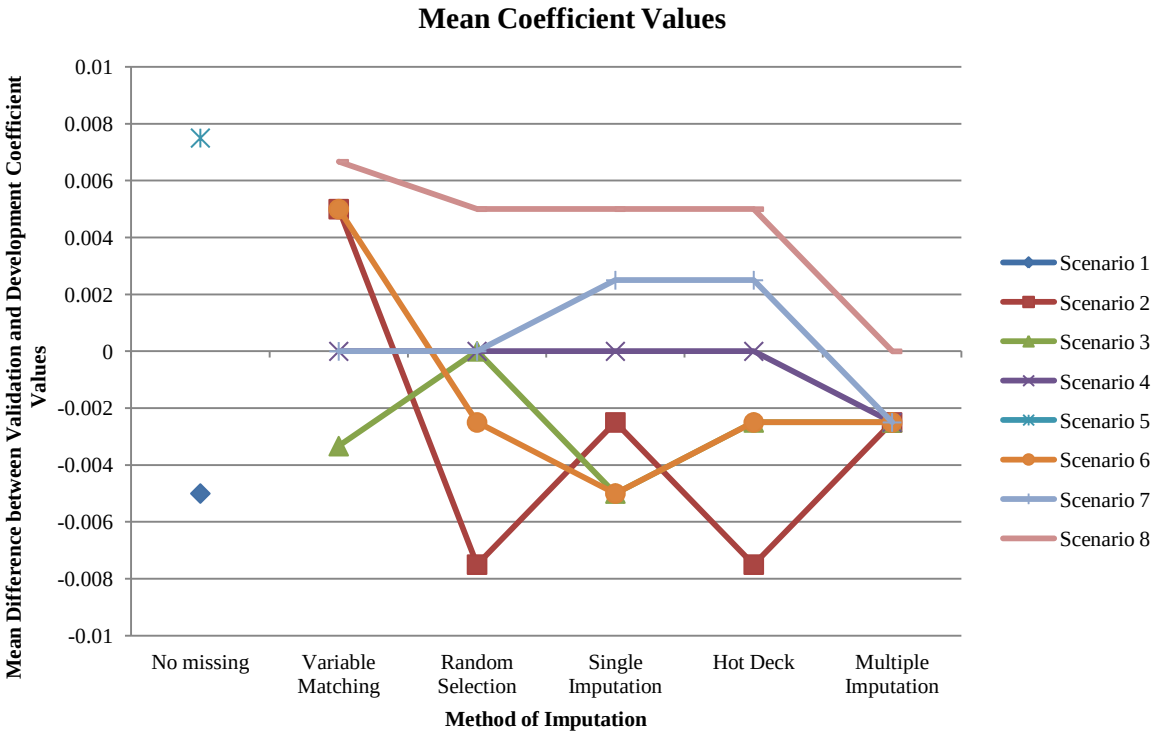
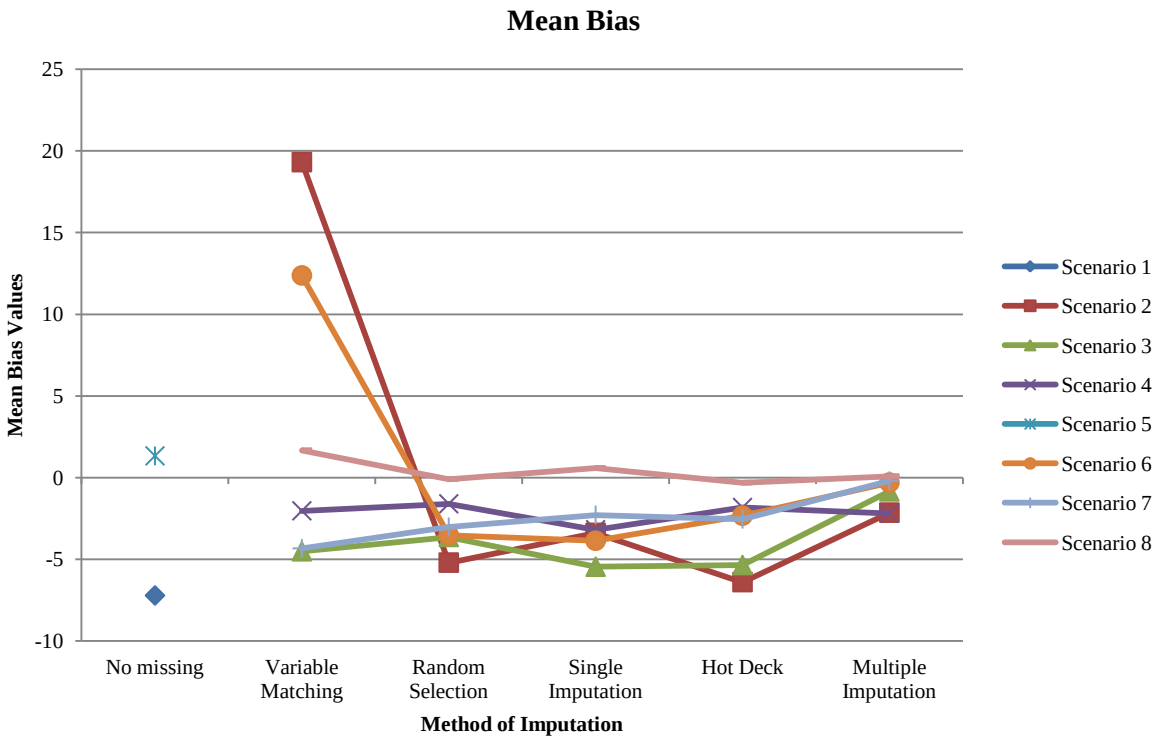


Figure 29: Mean Bias Results from simulation study examining methods of external validation and methods of handling a covariate missing from the validation dataset



Values presented in Figure 30 are the mean MSE values over all coefficients per scenario per method while those in Figure 31 are the mean coverage values over all coefficients per scenario per method. The MSE values are also very close, or equal, to zero which suggests almost perfect accuracy in the results irrespective of the presence of missing data and the strategies employed to impute missing data (Figure 30).

No method led to 95% coverage (Figure 31). The mean coverage values ranged from 80% to 100% with validation datasets that were smaller than the development validation datasets having larger coverage values generally. In the case of multiple imputation the coverage values were 100% irrespective of the size of the validation datasets.

Figure 30: Mean MSE Values from simulation study examining methods of external validation and methods of handling a covariate missing from the validation dataset

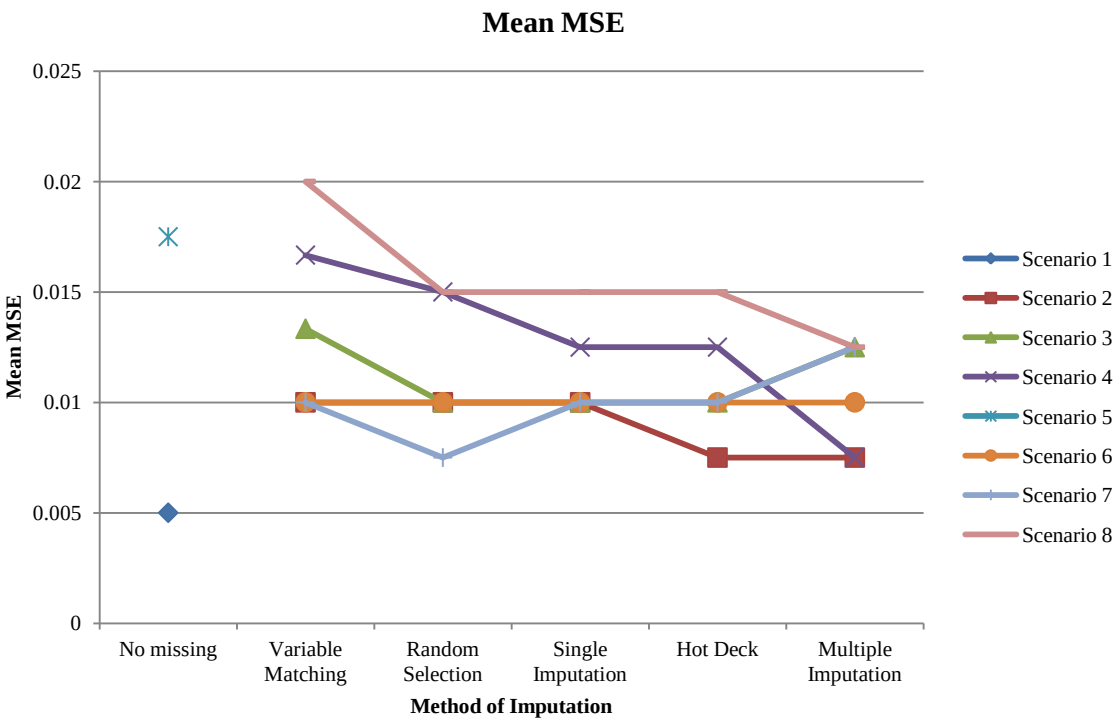
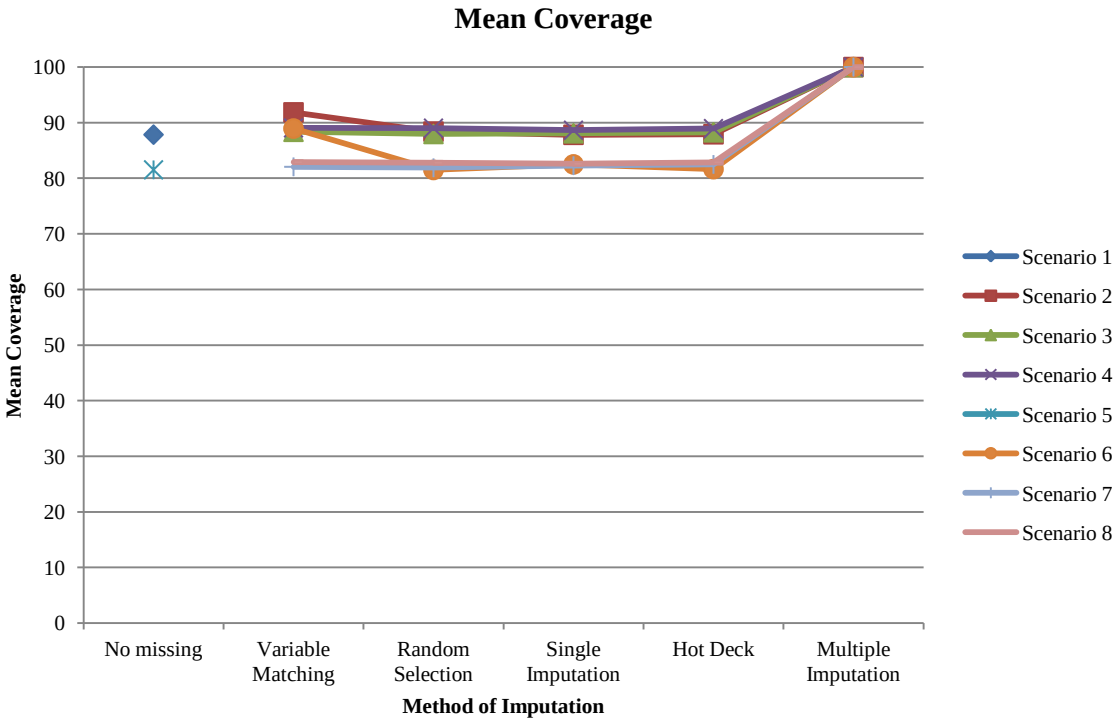


Figure 31: Mean Coverage Values from simulation study examining methods of external validation and methods of handling a covariate missing from the validation dataset



8.5 Conclusion

We conducted a series of bootstrap simulations to examine the performance of methods of external validation and methods for handling covariates missing from the validation dataset. Concordance has been applied to many previous models (as shown by the literature review in section 8.2.1) while deviance and the measure of prognostic separation have been applied sparingly as shown by [251, 280]. A comparison of these methods has not previously been made, nor an assessment of their ease of application to prognostic models. Standard methods of handling missing data within covariates have been regularly applied but this simulation study is the first to vary these methods to handle covariates that are completely missing in the context of external validation of prognostic models.

In the current study, we found that the use of concordance showed most often that the model was externally valid when the development and validation datasets were simulated in the same way. The measure of prognostic separation showed that the models were externally valid for over 90% of the simulations in the case of random selection with replacement multiple times (multiple imputation). This is likely to be as a result of the mechanics of multiple imputation – the imputation process was applied 20 times and the results averaged. Given that the measure of prognostic separation should have shown almost all models to be externally valid, especially in the case of no missing data, it cannot be recommended for general practice.

The deviance measure showed that the models were externally valid for over 80% of the simulations in the case of multiple imputation only. In the case of no missing data, over 60% of models had p-values for the comparison of deviance less than 5% and therefore at least 60% of models were not externally valid according to this measure. This should not be the case given the way the data were simulated and therefore comparison of deviance also cannot be recommended for general practice.

In the case of deviance it is also important to note that a significant p-value suggests that a model is not externally valid. However a non-significant p-value does not necessarily indicate that the model is externally valid. This is obviously not an ideal conclusion to draw from a measure of external validation – we want to conclude that a model is externally valid rather than there not being enough evidence to conclude it is not externally valid. Therefore this is another reason that this method cannot be recommended for general practice.

Variable matching was a poor method of imputation leading to results which were highly biased in this simulation study. Methods random selection with replacement, single imputation via estimation, hot deck imputation and random selection with replacement multiple times were all appropriate methods of imputation in this simulation study as they generally led to unbiased, accurate results. The more data missing from the validation dataset the better the model performed in terms of external validation because the validation dataset was effectively converging to the development dataset. Therefore caution is required when interpreting these results and further investigation is needed to examine the effect of more than one missing covariate.

It is important to note that the methods of external validation examined here are measuring different things – discrimination and calibration in the case of Harrell's *c*-statistic and deviance respectively. Ideally, it would be possible to use any of the methods to examine the external validity of a prognostic model and therefore use multiple methods to confirm the conclusion. However, our results suggested that it may be only concordance which was a reliable method of external validation. It is also worth noting that different degrees of external validation are likely rather than a clear-cut yes or no answer. The degree of external validity may be influenced by what the model will be used for as the consequences of using an 'incorrect' model in practice may have different impacts. For

example a model used to predict a mild adverse event such as a rash or head ache may not need to be as accurate as a model predicting risk of seizure recurrence and hence a lower degree of external validity may be acceptable for the adverse event model than the seizure risk model.

Further limitations of the work in this chapter include the simplicity of the simulated model which is unrealistic in having only three covariates – as shown in the literature review (section 8.2.1) the median number of covariates was five although a maximum of 20 has been observed. Additionally, all the imputation methods rely on having access to the development dataset which is often not possible. The simulation study presented here also makes a large assumption that the development and validation datasets are similar which is often not the case as the purpose of external validation is to evaluate the model for a slight different, though not dissimilar, case-mix. The simulation study as it currently stands, however, highlights the problems of external validation of prognostic models – even in the situation of an over-simplistic development prognostic model with a very similar validation model, two out of the three chosen methods of external validation failed to show validation when it was known to be present. This highlights the need for considerable further research in this area.

In 2004 May [280] measured the quality of predictions from prognostic models for survival time data using three different methods: Harrell's concordance statistic, Brier statistic and a novel method using deviance differences. The results showed that the deviance statistic was able to discriminate between quite similar models and could be used to choose a prognostic model that generalised well to new data. May found Harrell's *c*-statistic worked well in evaluating discrimination but was insensitive to differences in calibration and insensitive to differences between prognostic models [280].

The Brier score compares predicted probabilities with Bernouilli outcomes at a particular time, which is problematic in the context of censored survival data since different models may predict better at different lengths of follow up. May found that it failed to discriminate between the predictive performance of candidate prognostic models [280].

May's newly proposed approach based on deviance differences was able to discriminate between candidate prognostic models, however, further methodological work was recommended to support the use of deviance differences in model validation. In our simulation study we found that deviance differences were not usually appropriate in external validation of prognostic models.

While there is no doubt that models should be validated [264] and some authors do externally validate their model, as highlighted in the literature review, there are no specific guidelines on the best method of external validation. Also, although the recommended practice is to validate a model in a dataset that is plausibly related to the development dataset [8] there is very little literature on how to handle a covariate missing from the validation dataset. The work presented here suggests that concordance may be a simple method to externally validate prognostic models.

With regards to methods for handling covariates missing from the validation dataset in our simulation study, random selection with replacement, single imputation via estimation, hot deck imputation and random selection with replacement multiple times were found to be suitable imputation options. Computation time was extensive for hot deck imputation and multiple imputation. However, for validation of a single model via hot deck imputation, rather than validation of 1400 models as per the simulation study, computational time was up to five minutes, while it was up to one minute for multiple imputation. Results were available instantly for the other methods of imputation. Therefore, although hot deck and multiple imputation take marginally longer to run, there is little practical difference

between them. Therefore hot deck imputation or random selection with replacement multiple times may be the most appropriate methods, as shown by slightly improved results in this simulation study.

A personal decision regarding whether an increased type 1 or type 2 error is more acceptable then needs to be made. If an increased type 1 error is more acceptable then hot deck imputation may be the best method, otherwise random selection with replacement multiple times may be better.

Further work is now required to test these findings in other settings. Application of these methods of external validation and methods for handling missing covariates to real datasets can be seen in Chapter 9.

Chapter 9: Externally Validating Prognostic Models – Case Study

9.1 Introduction

In Chapter 5 a model was developed to determine for how long after a first unprovoked seizure a driver must be seizure-free before the risk of recurrence in the next 12 months falls below 20% enabling them to regain their driving license [193]. In this chapter the methods of external validation considered in Chapter 8 will be applied to real data to examine the externally validity of the prognostic model from Chapter 5. A summary of the developed model can be found in section 9.1 together with the three external validation datasets. Validation of the model is considered in sections 9.2 to 9.4. In section 9.5 simulated results are compared to real data which is forced to be missing to further investigate the methods of handling a missing covariate.

The model developed in Chapter 5 contained variables for aetiology, first degree relative with epilepsy, seizures only while asleep, EEG result, CT/MRI result and treatment policy which was forced into the model. We wished to externally validate this model using plausibly related datasets –NGPSE [131], a dataset collected in local epilepsy clinics in Western Australia (WA) [324] and the FIRST (FIRST) dataset from Italy [204]. NGPSE used the United Kingdom primary care system to obtain comprehensive data on a large and unselected cohort of patients with a possible diagnosis of epilepsy. It was initiated in 1984 and is one of the largest prospective cohort studies of epilepsy undertaken anywhere in the world. The WA dataset was collected in a hospital based study with the majority of patients referred from the emergency room following a first-ever tonic-clonic seizure. The dataset is in excess of 2000 patients who were recruited between 2000 and 2009. The FIRST dataset comprises patients from a randomized clinical trial on the efficacy of

antiepileptic drugs in reducing the risk of relapse after a first unprovoked tonic-clonic seizure.

NGPSE and WA do not have missing covariates and therefore provide a useful case study for the methods of external validation described in Chapter 8. FIRST is missing a covariate and therefore can be used to investigate the methods of imputation also discussed in Chapter 8. In addition to the three methods of external validation and five methods of imputation, and although not strictly a method of external validation, it is of interest to compare the conditional seizure risks across the validation and MESS datasets.

A summary of the demographics of patients in MESS, NGPSE, FIRST and WA, of driving age with only one seizure at randomisation (or recruitment as appropriate) can be seen in Table 27. A combined Kaplan-Meier curve for time from first to second seizure for these datasets can be seen in Figure 32.

Analyses were undertaken using *R* version 2.9.2 [182].

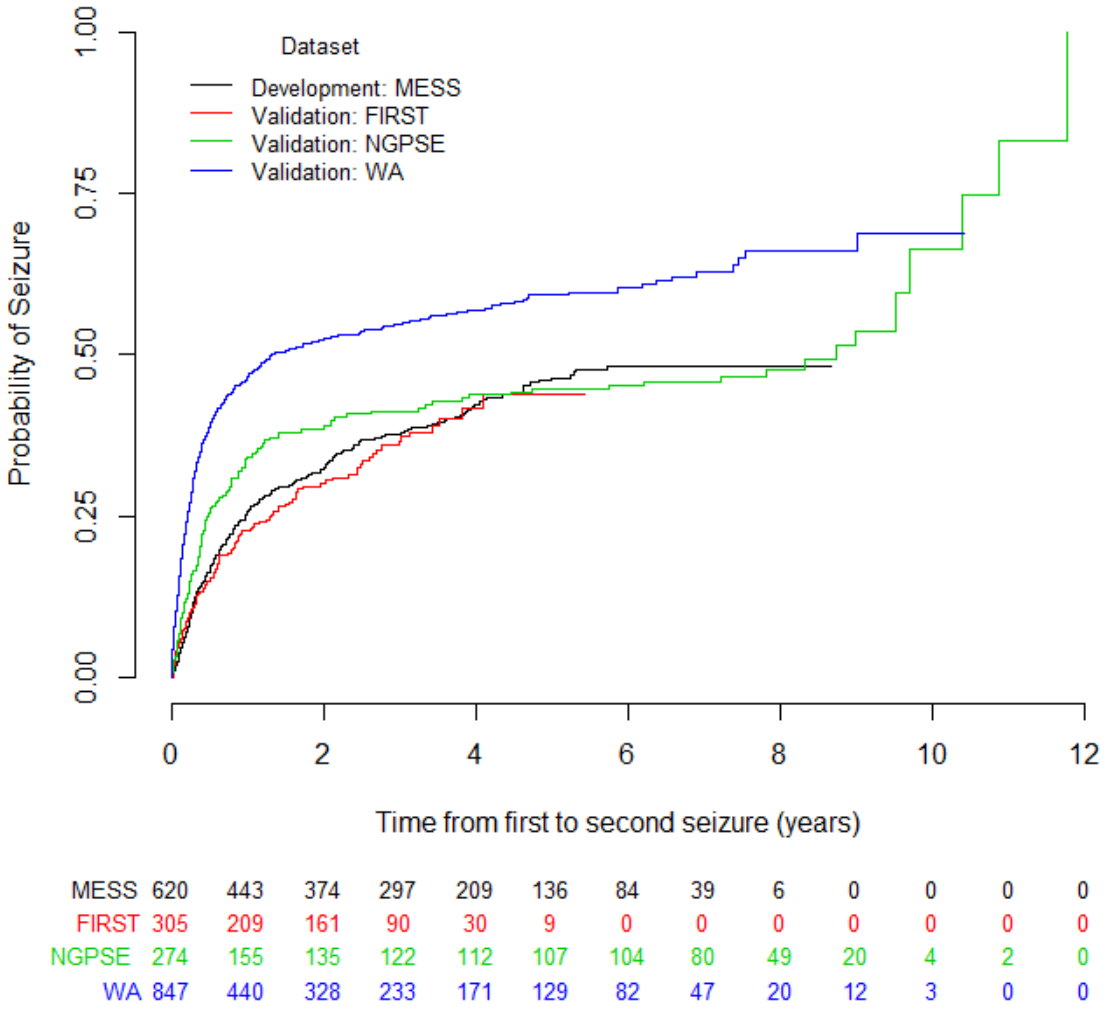


Figure 32: Kaplan-Meier curve for MESS, NGPSE, WA and FIRST including numbers at risk

(Patients of driving age with only one seizure at randomisation/recruitment)

Table 27: Demographics of patients in MESS, NGPSE, FIRST and WA (of driving age with only one seizure at randomisation/recruitment)

Entries in bold relate to variables included in the multivariable model

	MESS			NGPSE			WA			FIRST		
Characteristic	**Immediate Treatment (n=317)**	**Delayed Treatment (n=320)**	**Total (n=637)**	**On Treatment at Recruitment (n=79)**	**Not On Treatment at Recruitment (n=196)**	**Total (n=275)**	**On Treatment at Recruitment (n=235)**	**Not On Treatment at Recruitment (n=621)**	**Total (n=856)**	**On Treatment at Recruitment (n=156)**	**Not On Treatment at Recruitment (n=149)**	**Total (n=305)**
Age at randomisation in years, median (IQR)	32.0 (21.5, 49.5)	33.8 (22.4, 50.6)	32.9 (21.9, 50.3)	53.7 (39.7, 74.6)	44.1 (27.0, 66.6)	50.3 (31.8, 68.7)	48.0 (34.0, 64.5)	36.0 (24.0, 52.0)	39.0 (26.0, 56.0)	29.0 (20.0, 45.0)	28.0 (20.0, 48.0)	28.0 (20.0, 46.0)
Gender												
Male	199 (63)	213 (67)	412 (65)	39 (49)	97 (49)	136 (49)	154 (66)	390 (63)	544 (64)	91 (58)	82 (55)	173 (57)
Female	118 (37)	107 (33)	225 (35)	40 (51)	99 (51)	139 (51)	81 (34)	231 (37)	312 (36)	65 (42)	67 (45)	132 (43)
Aetiology												
Remote symptomatic	**55 (17)**	**49 (15)**	**104 (16)**	**57 (72)**	**100 (51)**	**157 (57)**	**136 (58)**	**138 (22)**	**274 (32)**	**11 (7)**	**11 (7)**	**22 (7)**
Not remote symptomatic	**262 (83)**	**271 (85)**	**533 (84)**	**22 (28)**	**96 (49)**	**118 (43)**	**99 (42)**	**483 (78)**	**585 (68)**	**145 (93)**	**138 (93)**	**283 (93)**
Neurological deficit												
Absent	288 (91)	299 (94)	587 (92)	NA	NA	NA	183 (78)	574 (92)	757 (88)	135 (87)	131 (12)	266 (87)
Present	27 (8)	19 (5)	46 (7)				52 (22)	47 (8)	99 (12)	21 (13)	18 (88)	39 (13)
Missing	2 (1)	2 (1)	4 (1)							0 (0)	0 (0)	0 (0)
Previous acute symptomatic seizures												
Febrile	15 (5)	14 (4)	29 (5)	NA	NA	NA	2 (1)	20 (3)	22 (3)	38 (24)	31 (21)	69 (23)
Other	1 (0)	7 (2)	8 (1)				231 (98)	595 (96)	826 (96)	0 (0)	0 (0)	0 (0)
No	301 (95)	299 (94)	600 (94)				2 (1)	6 (1)	7 (1)	118 (76)	118 (79)	236 (77)

	MESS			NGPSE			WA			FIRST		
Characteristic	**Immediate Treatment (n=317)**	**Delayed Treatment (n=320)**	**Total (n=637)**	**On Treatment at Recruitment (n=79)**	**Not On Treatment at Recruitment (n=196)**	**Total (n=275)**	**On Treatment at Recruitment (n=235)**	**Not On Treatment at Recruitment (n=621)**	**Total (n=856)**	**On Treatment at Recruitment (n=156)**	**Not On Treatment at Recruitment (n=149)**	**Total (n=305)**
Epilepsy in first degree relative												
Yes	**33 (10)**	**34 (11)**	**67 (10)**	**4 (5)**	**17 (9)**	**21 (8)**	**20 (9)**	**73 (12)**	**93 (11)**	**24 (15)**	**12 (8)**	**36 (12)**
No	**284 (90)**	**286 (89)**	**570 (90)**	**75 (95)**	**179 (91)**	**254 (92)**	**207 (88)**	**532 (85)**	**739 (86)**	**132 (85)**	**137 (92)**	**269 (88)**
Missing							**8 (3)**	**16 (3)**	**24 (3)**			
Seizures												
Partial	102 (32)	100 (31)	202 (32)									
Generalised/ not definitely partial	205 (65)	217 (68)	422 (66)	NA	NA	NA	NA	NA	NA	NA	NA	NA
Other	10 (3)	3 (1)	13 (2)									
Seizures only while asleep												
Yes	**65 (20)**	**50 (16)**	**115 (18)**	**19 (24)**	**21 (11)**	**40 (15)**	**57 (24)**	**473 (76)**	**650 (76)**	**NA**	**NA**	**NA**
No	**252 (80)**	**269 (84)**	**521 (82)**	**60 (76)**	**175 (89)**	**235 (85)**	**177 (75)**	**147 (24)**	**204 (24)**			
Missing	**-**	**1 (0)**	**1 (0)**	**0 (0)**	**0 (0)**	**0 (0)**	**1 (1)**	**1 (0)**	**2 (0)**			
EEG results												
Normal	**147(46)**	**136 (43)**	**283 (44)**	**9 (10)**	**41 (21)**	**50 (18)**	**63 (27)**	**363 (59)**	**426 (50)**	**73 (47)**	**71 (52)**	**144 (47)**
Abnormal	**149 (47)**	**160 (50)**	**309 (49)**	**26 (33)**	**45 (23)**	**71 (26)**	**157 (67)**	**250 (40)**	**407 (48)**	**83 (53)**	**78 (48)**	**161 (53)**
Not done/Missing	**21 (7)**	**24 (7)**	**45 (7)**	**44 (57)**	**110 (56)**	**154 (56)**	**15 (6)**	**8 (1)**	**23 (2)**			
CT/MRI results												
Normal	**219 (69)**	**231 (72)**	**450 (71)**	**19 (24)**	**38 (20)**	**57 (21)**	**89 (38)**	**457 (74)**	**546 (64)**	**124 (79)**	**122 (82)**	**246 (81)**
Abnormal	**38 (12)**	**34 (11)**	**72 (11)**	**15 (19)**	**24 (12)**	**39 (14)**	**134 (57)**	**108 (17)**	**242 (28)**	**32 (21)**	**27 (18)**	**59 (19)**
Not done/Missing	**60 (19)**	**55 (17)**	**115 (18)**	**45 (57)**	**134 (68)**	**179 (65)**	**12 (5)**	**56 (9)**	**68 (8)**	**0 (0)**	**0 (0)**	**0 (0)**

In theory, data used to externally validate a prognostic model should be plausibly related to the development dataset, such that the datasets can be considered to come from the same super-population. There are however, no guidelines suggesting how to check whether the development and validation datasets are from a super-population.

In the case of our validation datasets some are more plausibly related to MESS than others. Patients in NGPSE tended to be older than those in MESS and there were more females in NGPSE, although the variables for age and gender were not included in the multivariable model being validated here. Considerably more patients in NGPSE had a remote symptomatic aetiology than those in MESS and many more patients in NGPSE had missing results for EEG and CT/MRI. Characteristics for patients in the WA dataset were mostly similar to those in MESS except for previous acute symptomatic seizures which were more frequently reported in WA. Additionally, many more patients had seizures only while asleep in WA than in MESS.

The FIRST dataset had a similar distribution of characteristics to those of MESS. More patients had previous febrile seizures in FIRST although this variable was not included in the multivariable model for MESS. In FIRST, data on seizures while asleep was not collected. Additionally in FIRST, EEG and CT/MRI result were divided into two categories (normal/not done and abnormal) unlike in MESS where patients could have three result categories: normal, abnormal and not done. Follow-up in FIRST was considerably shorter than in MESS as shown in Figure 32.

Table 28 summarises the unadjusted 12 month seizure recurrence risks at time points following a first seizure for each dataset. The risks, according to NGPSE, were similar to the overall estimates in our analysis of MESS. All estimates from 12 months onwards were significantly below 20% in both MESS and NGPSE. Estimates at six months varied however - according to the NGPSE results, patients who were treated immediately should wait 12

months until their recurrence risk in the next 12 months was significantly below 20% unlike the MESS results which suggested only six months. Patients in NGPSE who were treated immediately had a higher risk of a seizure in the following 12 months generally than those who were not treated. This is probably due to the ability of clinicians to detect patients at high risk of seizure recurrence and to recommend relevant treatment. Confidence intervals for NGPSE were wider than for MESS as a result of the smaller sample size for NGPSE.

The 12 month seizure recurrence risks at time points post index seizure for the WA dataset were similar to those for the MESS data. Instead of concluding that patients who were treated should be allowed to drive after six months and those who were not treated after 12 months, results from WA suggested patients should wait 12 months irrespective of the treatment policy - the conditional risks were not significantly below 20% until 12 months.

The results for FIRST were very similar to those using the MESS data. As for the MESS data the point estimates at six month were less than the 20% risk level set by the DVLA but the result was only significantly below 20% for patients randomised to immediate treatment.

Despite the differences in characteristics and follow-up it is plausible that these datasets came from the same super-population. FIRST is the closest match to MESS in terms of the proportions of patients with certain characteristics, however the follow-up is much shorter and a significant covariate is missing. In the case of NGPSE and WA, information on the same covariates is available but the proportions of patients with some characteristics are not always similar to the MESS data. These datasets are, however, the closest matches available and hence we will continue with the external validation of MESS with NGPSE, WA and FIRST. MESS is validated with NGPSE in section 9.2, with WA in 9.3 and FIRST in 9.4.

Table 28: 12 month seizure recurrence risks at time points post index seizure:

Risk of seizure in next 12 months (%, 95% Confidence Interval)

	MESS				NGPSE				WA				FIRST			
	Immediate treatment		Delayed treatment		Immediate Treatment (Treated)		Delayed Treatment (Not treated)		Immediate Treatment (Treated)		Delayed Treatment (Not treated)		Immediate Treatment (Treated)		Delayed Treatment (Not treated)	
Time post index seizure (months)	**No. at Risk**	**Risk of seizure**	**No. at Risk**	**Risk of seizure**	**No. at Risk**	**Risk of seizure**	**No. at Risk**	**Risk of seizure**	**No. at Risk**	**Risk of seizure**	**No. at Risk**	**Risk of seizure**	**No. at Risk**	**Risk of seizure**	**No. at Risk**	**Risk of seizure**
6	**260**	**14 (10 to 18)**	**254**	**18 (13 to 23)**	56	19 (8 to 29)	136	16 (10 to 23)	128	23 (16 to 30)	382	18 (14 to 22)	139	11 (5 to 16)	106	19 (11 to 27)
12	**230**	**7 (4 to 11)**	**219**	**10 (6 to 15)**	44	5 (0 to 12)	113	8 (3 to 13)	106	10 (4 to 16)	335	11 (8 to 15)	128	9 (4 to 14)	86	11 (4 to 17)
18	**211**	**8 (5 to 12)**	**197**	**12 (8 to 17)**	41	8 (0 to 16)	102	3 (0 to 7)	91	-	286	8 (4 to 11)	114	7 (2 to 12)	74	12 (3 to 20)
24	**199**	**7 (3 to 10)**	**182**	**10 (5 to 14)**	39	6 (0 to 13)	102	3 (0 to 7)	82	-	-	-	106	8 (2 to 14)	65	11 (2 to 20)

9.2 Validation using NGPSE

Of the 792 patients recruited to the NGPSE study, 217 were aged less than 16 years at ascertainment. For 300 patients this was not their first episode so they were removed from the analysis. This left 275 suitable patients in comparison to 637 available for analysis in the MESS study.

9.2.1 Results

The external validation results of MESS via NGPSE can be seen in Table 29. The p-value for the comparison of deviance was significant (0.01). Therefore the prognostic model developed using the MESS data does not appear to generalise well to the NGPSE data according to the deviance statistic. However, in the simulation study, comparison of deviance was found to be a poor method of external validation. The poor validation could therefore be a reflection on the method of validation rather than a lack of agreement between the datasets. Additionally, this may be because of the small sample size of NGPSE.

Table 29: Summary of Results of external validation of MESS model via NGPSE dataset

	MESS	NGPSE
Deviance		
Statistic	30.91	10.26
Degrees of Freedom	8	
p-value	0.01	
Concordance		
Statistic	0.591 (0.556, 0.627)	0.601 (0.547, 0.654)
Difference in Concordance	0.01	
Prognostic Separation		
Statistic	0.636	0.634
Difference (95% CI)	0.00 (-0.02, 0.03)	

As explained in section 8.2.3 it is not appropriate to calculate a confidence interval for a difference in concordance statistic between the development and validation datasets. However it is possible to obtain estimates of confidence intervals for individual

concordance statistics [277]. If two statistics have non-overlapping confidence intervals they are necessarily significantly different. However, if they have overlapping intervals it is not necessarily true that they are not significantly different [325]. In the absence of a better method, we considered the similarity of the concordance statistic across the development and validation dataset via examination of the individual confidence intervals [326]. In the case of MESS and NGPSE, as shown in Table 29, the confidence intervals and estimates were very similar. Therefore it seems reasonable to conclude that the MESS model generalises well to NGPSE according to the concordance measure.

A confidence interval for the difference in prognostic separation between the development and validation datasets was calculated via bootstrap resampling. In the case of MESS and NGPSE the confidence interval for a difference in separation was (-0.02, 0.03) which included 0. Therefore, the MESS model generalises well to the NGPSE dataset according to prognostic separation (Table 29).

Table 30 shows the conditional risk estimates for MESS and NGPSE according to the prognostic model for time to second seizure following a first ever seizure. This model was developed using patients from the MESS study who were of driving age (over 16 at randomisation) and had had only one seizure at recruitment. Full details of the calculations can be found in Bonnett *et al*. 2010 [193] and in Chapter 5. The estimates presented assume that seizures were not confined to sleep and that there were no first degree relatives with epilepsy. The estimates for NGPSE were generally in agreement with the estimates for MESS although they were marginally higher for patients in NGPSE who were treated at ascertainment than they were for the comparable patients in MESS.

Table 30: Seizure recurrence risk in the next 12 months estimated from multivariable model at specific time points of seizure freedom. Estimates presented assume seizures not confined to sleep and no first degree relative with epilepsy – MESS & NGPSE

				MESS				NGPSE			
Variable				Immediate Treatment		Delayed Treatment		Treated		Not Treated	
Aetiology	EEG	CT/MRI scan results	Months from index seizure	Risk of seizure within next 12 months (%, 95% CI)	Months from index seizure until annual risk falls below 20%	Risk of seizure within next 12 months (%, 95% CI)	Months from index seizure until annual risk falls below 20%	Risk of seizure within next 12 months (%, 95% CI)	Months from index seizure until annual risk falls below 20%	Risk of seizure within next 12 months (%, 95% CI)	Months from index seizure until annual risk falls below 20%
Not remote symptomatic	Normal	Normal	6 12	13 (10 to 16) 7 (5 to 10)	1.2	16 (12 to 19) 9 (6 to 12)	3.2	17 (11 to 22) 7 (3 to 11)	4.8	16 (11 to 22) 7 (3 to 11)	4.8
Not remote symptomatic	Abnormal	Normal	6 12	20 (16 to 23) 11 (9 to 14)	5.5	23 (20 to 26) 13 (11 to 16)	7.0	21 (15 to 26) 9 (5 to 13)	6.2	20 (15 to 25) 9 (5 to 13)	6.1
Not remote symptomatic	Normal	Abnormal	6 12	14 (11 to 17) 8 (5 to 11)	2.1	17 (14 to 20) 10 (7 to 12)	3.6	19 (13 to 24) 8 (4 to 12)	5.7	18 (13 to 24) 8 (4 to 12)	5.5
Not remote symptomatic	Abnormal	Abnormal	6 12	21 (18 to 24) 12 (9 to 15)	6.1	25 (22 to 28) 14 (12 to 17)	8.1	23 (18 to 28) 10 (6 to 14)	7.5	23 (17 to 28) 10 (6 to 14)	7.4
Remote symptomatic	Normal	Normal	6 12	17 (14 to 20) 10 (7 to 12)	3.6	20 (17 to 23) 12 (9 to 14)	6.0	19 (14 to 25) 8 (4 to 12)	5.7	19 (13 to 24) 8 (4 to 12)	5.7
Remote symptomatic	Abnormal	Normal	6 12	25 (22 to 28) 15 (12 to 17)	8.7	30 (27 to 32) 18 (15 to 20)	10.7	24 (18 to 29) 10 (6 to 14)	7.5	23 (18 to 28) 10 (6 to 14)	7.5
Remote symptomatic	Normal	Abnormal	6 12	18 (15 to 21) 10 (8 to 13)	4.8	22 (19 to 25) 13 (10 to 15)	6.5	21 (16 to 27) 9 (5 to 13)	6.3	21 (16 to 26) 9 (5 to 13)	6.2
Remote symptomatic	Abnormal	Abnormal	6 12	27 (24 to 30) 16 (13 to 18)	9.2	32 (29 to 34) 19 (16 to 21)	11.3	26 (21 to 31) 12 (8 to 16)	9.4	26 (21 to 31) 11 (7 to 15)	9.3

9.2.1.1 Sensitivity Analysis

External validity involves checking that a model is valid in populations that are plausibly related to the development dataset. In this section a sensitivity analysis is presented which investigated the impact of length of follow-up of the development and validation datasets on external validation.

As shown in Figure 32, there are very few patients at risk after about eight years in MESS. It is therefore of interest to investigate if truncating NGPSE at eight years to match the MESS data makes a difference to the external validity. If patients with follow-up in excess of eight years in NGPSE are influencing seizure risk estimates this may have an impact on the results of external validation. This is because the agreement between the MESS and NGPSE models without influential patients may be reduced.

The external validation results in this truncated setting can be seen in Table 31. The confidence intervals for the concordance statistics were very similar between the MESS and truncated NGPSE datasets. Together with the small difference in concordance it seems that the MESS model generalises well to the NGPSE dataset, truncated at eight years. The confidence interval for a difference in prognostic separation between the development and validation datasets was (-0.02, 0.03) which includes 0. Therefore it seems that the MESS model does generalise well to the NGPSE data. These conclusions are the same as those obtained for the external validation of MESS with the complete NGPSE dataset.

The p-value of 0.02 was significant for the comparison of deviance between the development (MESS) and validation (NGPSE) datasets. Therefore the prognostic model developed using the MESS data does not appear to generalise well to the truncated NGPSE data. This conclusion is as per the comparison of MESS and the complete NGPSE dataset.

Table 31: Summary of Results of external validation of MESS model via a truncated NGPSE dataset

	MESS	NGPSE
Deviance		
Statistic	30.91	13.08
Degrees of Freedom	8	
p-value	0.02	
Concordance		
Statistic	0.591 (0.556, 0.627)	0.599 (0.544, 0.655)
Difference in Concordance	0.01	
Prognostic Separation		
Statistic	0.636	0.632
Difference (& 95% CI)	0.00 (-0.03, 0.02)	

Table 32: Seizure recurrence risk in the next 12 months estimated from multivariable model at specific time points of seizure freedom. Estimates presented assume seizures not confined to sleep and no first degree relative with epilepsy – truncated NGPSE

Variable				Immediate Treatment		Delayed Treatment	
Aetiology	EEG	CT/MRI	Months from index seizure	Risk of seizure within next 12 months (%, 95% CI)	Months from index seizure until annual risk falls below 20%	Risk of seizure within next 12 months (%, 95% CI)	Months from index seizure until annual risk falls below 20%
Not remote symptomatic	Normal	Normal	6 12	26 (19, 33) 12 (6, 18)	9.4	22 (15, 30) 10 (4, 16)	6.9
Not remote symptomatic	Abnormal	Normal	6 12	27 (20, 34) 12 (6, 18)	9.5	23 (16, 30) 11 (4, 17)	7.5
Not remote symptomatic	Normal	Abnormal	6 12	31 (24, 37) 14 (9, 20)	10.7	27 (20, 34) 12 (6, 18)	9.5
Not remote symptomatic	Abnormal	Abnormal	6 12	32 (25, 38) 15 (9, 21)	10.7	28 (21, 35) 13 (7, 19)	9.5
Remote symptomatic	Normal	Normal	6 12	27 (20, 34) 12 (6, 18)	9.5	23 (16, 31) 11 (5, 17)	7.5
Remote symptomatic	Abnormal	Normal	6 12	28 (21, 35) 13 (7, 19)	9.5	24 (17, 31) 11 (5, 17)	8.0
Remote symptomatic	Normal	Abnormal	6 12	32 (25, 38) 15 (9, 21)	10.7	28 (21, 35) 13 (7, 19)	9.5
Remote symptomatic	Abnormal	Abnormal	6 12	33 (27, 39) 16 (10, 21)	10.7	29 (22, 36) 13 (7, 19)	9.5

Truncating NGPSE at eight years did not have an impact on the results of the external validation. However the conditional seizure risks shown in Table 32 were higher than the original NGPSE analyses and were also higher than the MESS results (Table 30). For this

reason it seems better not to truncate the dataset. However, further sensitivity analyses, and potentially simulation studies, are needed to determine if this result is valid over many situations. They are also needed to develop guidelines for checking that the development and validation datasets are from the same super-population.

9.2.2 Conclusion

According to the comparison of deviance method, the MESS model does not generalise well to NGPSE. However the concordance and prognostic separation measures suggest that, in fact, the MESS model does generalise well to NGPSE. In addition the conditional seizure risks were similar across MESS and NGPSE.

Truncating the NGPSE dataset to more closely match the length of follow-up in MESS had no impact on the external validity of the MESS model. It did, however, lead to increased risk estimates in NGPSE which were less similar to MESS. Therefore truncating the NGPSE dataset had a negative impact on the results.

9.3 Validation using WA

In the WA dataset there were 856 suitable patients with one seizure at recruitment and aged over 16 years.

9.3.1 Results

The results of comparison of deviance, concordance and prognostic separation can be seen in Table 33. The p-value of 0.07 for the comparison of deviance was not significant at the 5% level. Therefore there was no evidence (at the 5% level) to suggest that the MESS model was not externally valid although at the 10% level of significance there is such evidence.

The confidence intervals for the concordance statistics for the MESS and NGPSE datasets were very similar. Therefore, together with the small difference in concordance, it seems that the MESS model generalises well to the WA dataset. The confidence interval for a difference in separation between the development and validation datasets did not include 0. Therefore it seems that the MESS model does not generalise very well to the WA dataset according to prognostic separation.

Table 33: Summary of Results of external validation of MESS model via WA dataset

	MESS	WA
Deviance		
Statistic	30.91	45.40
Degrees of Freedom	8	
p-value	0.07	
Concordance		
Statistic	0.591 (0.556, 0.627)	0.590 (0.563, 0.618)
Difference in Concordance	0.00	
Prognostic Separation		
Statistic	0.636	0.556
Difference (& 95% CI)	0.08 (0.06, 0.10)	

Table 34 shows the seizure recurrence risks in the next 12 months at specific time points of seizure freedom as estimated from the multivariable model fitted to the WA dataset. The estimates presented assumed that seizures were not confined to sleep and that there were no first degree relatives with epilepsy. These estimates are generally in agreement with those in MESS (Table 30) for patients with normal and abnormal EEG and CT/MRI results, but lower for patients with abnormal EEG but normal CT/MRI results. Additionally, the risk estimates were generally higher in the WA dataset for patients with abnormal CT/MRI results and normal EEG results than in MESS.

Table 34: Seizure recurrence risk in the next 12 months estimated from multivariable model at specific time points of seizure freedom. Estimates presented assume seizures not confined to sleep and no first degree relative with epilepsy

Variable				Immediate Treatment		Delayed Treatment	
Aetiology	EEG	CT/MRI	Months from index seizure	Risk of seizure within next 12 months (%, 95% CI)	Months from index seizure until annual risk falls below 20%	Risk of seizure within next 12 months (%, 95% CI)	Months from index seizure until annual risk falls below 20%
Not remote symptomatic	Normal	Normal	6 12	19 (16 to 23) 11 (8 to 14)	5.8	22 (19 to 25) 11 (8 to 14)	7.0
Not remote symptomatic	Abnormal	Normal	6 12	6 (2 to 10) 12 (9 to 15)	3.6	10 (7 to 14) 12 (9 to 15)	4.0
Not remote symptomatic	Normal	Abnormal	6 12	29 (26 to 32) 12 (10 to 15)	7.3	32 (29 to 35) 13 (10 to 16)	7.3
Not remote symptomatic	Abnormal	Abnormal	6 12	20 (17 to 23) 13 (10 to 16)	6.0	23 (20 to 27) 14 (11 to 16)	7.3
Remote symptomatic	Normal	Normal	6 12	18 (15 to 22) 14 (11 to 17)	5.6	22 (19 to 25) 14 (11 to 17)	6.5
Remote symptomatic	Abnormal	Normal	6 12	0 (0 to 2) 15 (12 to 18)	3.5	4 (0 to 8) 15 (12 to 18)	3.8
Remote symptomatic	Normal	Abnormal	6 12	31 (28 to 34) 16 (13 to 18)	8.5	35 (32 to 37) 16 (13 to 19)	8.7
Remote symptomatic	Abnormal	Abnormal	6 12	17 (14 to 21) 16 (14 to 19)	5.5	22 (19 to 25) 17 (14 to 20)	6.5

9.3.2 Conclusion

According to comparison of prognostic separation the MESS model does not generalise well to the WA dataset. The comparison of deviance results suggest that the model also does not generalise well at the 10% level of significance but it is alright at the 5% level. The comparison of concordance suggests that the MESS model does generalise well to WA. The conditional risk estimates for WA were similar to those for MESS but there were some subgroups that had very different risk estimates.

In the simulation study of Chapter 8 neither comparison of deviance nor prognostic separation performed as expected and were consequently not recommended as methods of external validation. Based on this, it appears that the MESS model does generalise well to WA as the concordance statistics were very similar across the MESS and WA datasets.

9.4 Validation using FIRST

Of the 420 patients randomised to the FIRST study, 305 were older than 16 at recruitment. Within FIRST there are issues with the EEG and CT/MRI variables which were only recorded as abnormal or normal with the normal response also being used for patients who did not have such a scan in FIRST. Additionally, no data was collected on seizures only while asleep. As the seizures while asleep variable is missing from FIRST, the five methods of imputation described in Chapter 8 were employed: (1) remodelling with variable matching; (2) random selection with replacement; (3) single imputation via estimated proportions; (4) hot deck imputation; and (5) random selection with replacement multiple times.

9.4.1 Sensitivity Analysis

In MESS, patients could have result categories normal, abnormal or not done for EEG and CT/MRI result. In FIRST however, results were only available in the form of normal/not done and abnormal. In the interests of completeness a sensitivity analysis was performed for a single method of imputation, variable matching, to determine if this difference in classification had a significant impact on the external validation results.

When the sleep variable was removed from MESS, to match FIRST, and when EEG and CT were left unchanged in MESS, it was not possible to obtain a comparison of deviance with FIRST. This was because the number of degrees of freedom was different across the models - in MESS there were seven but in FIRST there were only five. When EEG and CT were collapsed to two categories each in MESS to match FIRST, the p-value for the comparison of deviance between MESS and FIRST was 0.86 which was not significant (Table 35). Therefore there is no evidence to suggest that the prognostic model developed using the MESS data (with EEG and CT/MRI collapsed to two categories each) does not generalise well to the FIRST data (with variable matching).

The confidence intervals for concordance were quite similar for MESS and FIRST irrespective of the categorisation of the EEG and CT/MRI variables in MESS. Together with the relatively small differences in concordance, the MESS model seems to generalise fairly well to the FIRST data with variable matching.

Table 35: Summary of Results of external validation of MESS model with EEG and CT variables with 3 categories and the FIRST dataset using variable matching

	FIRST (variable matching)	MESS (EEG/CT as 3 categories)	MESS (EEG/CT as 2 categories)
Deviance			
Statistic	23.03	N/A	21.11
Degrees of Freedom	-	N/A	5
p-value	-	N/A	0.86
Concordance			
Statistic	0.646(0.593, 0.699)	0.581 (0.545, 0.618)	0.576 (0.541, 0.612)
Difference in Concordance		0.07	0.07
Prognostic Separation			
Statistic	0.830	0.578	0.517
Difference (& 95% CI)		-0.27 (-0.28, -0.22)	-0.31 (-0.34, -0.29)

The confidence intervals for the difference in prognostic separation between the development and validation datasets did not include 0, irrespective of the number of categories used to model EEG and CT/MRI in MESS. Therefore the MESS model does not generalise to the FIRST dataset according to this measure.

The external validation results were comparable across the sensitivity analysis. The only difference was that it was not possible to draw a conclusion for external validation via comparison of deviance when EEG and CT/MRI were grouped in three categories in MESS but in two in FIRST. Given that the results for the concordance and separation measures were very similar across the sensitivity analysis, the external validation of MESS via FIRST proceeded with the EEG and CT/MRI variables collapsed to only two categories each for MESS. This ensured that all three methods of external validation could be tested.

Another option would be to remove the differently categorised variable from the development dataset, especially if it has a small influence on the model, re-develop the reduced model and then validation it. Methodologically neither are ideal solutions. It is preferable to leave the developed model untouched during the external validation phase of analysis. Therefore, if other validation datasets were available it would be best to exclude FIRST from the external validation stage, especially as, in addition to the differing categorisations of EEG and CT/MRI, a variable is missing.

9.4.1.1 Conditional Seizure Risks

Table 36 shows the seizure recurrence risks in the next 12 months, estimated from the amended multivariable model for MESS, at specific time points of seizure freedom – EEG and CT/MRI were collapsed to two categories each and the variable for seizures while asleep has been removed from MESS as per the variable matching method of handling a missing covariate. Estimates presented assume no first degree relative with epilepsy. Without the variable for seizures while asleep some confidence intervals for MESS do not overlap which makes them difficult to interpret.

The recurrence risks for patients without a remote symptomatic aetiology were higher in the model for MESS without the sleep variable (Table 36) than in the model with the variable (Table 30). The risks for patients with a remote symptomatic aetiology were lower in the case of the model without the sleep variable than for the model with the variable. This suggests an interaction may exist between seizures while asleep and aetiology. The log-likelihood for the model with aetiology and sleep was 10.7 on two degrees of freedom. For the model inclusive of an interaction term the log-likelihood was 11.3 on three degrees of freedom. The difference in log-likelihood was not statistically significant as the p-value for the difference was 0.53. Therefore there was insufficient evidence of an interaction

between aetiology and sleep. The differences observed between the results in Table 36 and Table 30 are therefore likely to be because of the highly influential sleep variable.

The conditional risks for FIRST were quite different to those for MESS (without the sleep variable and with EEG and CT/MRI categories each collapsed) as can be seen in Table 36. In most cases they were much lower than the MESS results. However, for patients with remote symptomatic aetiology who were in the delayed treatment category, the estimates were generally much higher than for the same patients in MESS.

Table 36: Seizure recurrence risk in the next 12 months estimated from multivariable model at specific time points of seizure freedom. Estimates presented assume no first degree relative with epilepsy. MESS (EEG/CT 2 categories) & FIRST (variable matching)

				MESS				FIRST			
Variable				Immediate Treatment		Delayed Treatment		Immediate Treatment		Delayed Treatment	
Aetiology	EEG	CT/MRI scan results	Months from index seizure	Risk of seizure within next 12 months (%, 95% CI)	Months from index seizure until annual risk falls below 20%	Risk of seizure within next 12 months (%, 95% CI)	Months from index seizure until annual risk falls below 20%	Risk of seizure within next 12 months (%, 95% CI)	Months from index seizure until annual risk falls below 20%	Risk of seizure within next 12 months (%, 95% CI)	Months from index seizure until annual risk falls below 20%
Not remote symptomatic	Normal	Normal	6 12	18 (15, 21) 10 (8, 13)	4.8	22 (19, 25) 12 (10, 15)	6.5	6 (2,11) 4 (0, 9)	<1	14 (9, 18) 10 (5, 14)	0.6
Not remote symptomatic	Abnormal	Normal	6 12	26 (24,29) 15 (13, 18)	9.2	31 (28, 34) 18 (16, 20)	10.8	10 (5, 14) 7 (2, 11)	<1	20 (16,24) 14 (10, 18)	5.3
Not remote symptomatic	Normal	Abnormal	6 12	19 (16, 22) 11 (8, 13)	5.3	22 (19, 25) 13 (10, 15)	6.8	8 (3, 13) 5 (1, 10)	<1	16 (12, 21) 11 (7, 15)	2.2
Not remote symptomatic	Abnormal	Abnormal	6 12	27 (24, 30) 16 (13, 18)	9.2	32 (29, 34) 19 (16, 21)	11.3	11 (7, 16) 8 (4, 12)	<1	23 (19, 27) 16 (13, 20)	7.4
Remote symptomatic	Normal	Normal	6 12	13 (10, 17) 7 (5, 10)	1.2	16 (13, 19) 9 (6, 12)	3.2	10 (6, 15) 7 (3, 11)	<1	21 (17, 25) 15 (11, 19)	6.7
Remote symptomatic	Abnormal	Normal	6 12	19 (16, 22) 11 (8, 14)	5.5	23 (20, 26) 13 (10, 16)	6.9	15 (11, 19) 10 (6, 14)	0.7	30 (26, 34) 21 (18, 25)	12.4
Remote symptomatic	Normal	Abnormal	6 12	14 (10, 17) 8 (5, 10)	1.7	16 (13, 19) 9 (6, 12)	3.4	12 (7, 17) 8 (4, 12)	<1	25 (21, 29) 17 (14, 21)	10.2
Remote symptomatic	Abnormal	Abnormal	6 12	20 (17, 23) 11 (9, 14)	5.7	24 (21, 26) 13 (11, 16)	7.6	18 (13, 22) 12 (8, 16)	2.7	35 (31, 38) 25 (22, 28)	16.0

9.4.2 Methods for Handling a Missing Covariate

The results of external validation via variable matching were presented in section 9.4.1. Following on from the sensitivity analysis of section 9.4.1, the EEG and CT/MRI variables were modeled as two categories each in MESS for the external validation via FIRST.

Comparison of deviance led to non-significant p-values for all methods of imputation (Table 37). Therefore there is no evidence to suggest that the prognostic model developed using the MESS data does not generalise well to the FIRST dataset irrespective of the method used to handle the missing covariate. This conclusion was confirmed by the confidence intervals for the concordance statistics which were fairly similar across all imputation methods. Together with the relatively small differences in concordance between MESS and the imputed FIRST datasets, it seems that the MESS model generalises fairly well to FIRST.

The confidence intervals for the difference in separation between the development and validation datasets did not include 0, the value of no difference. Therefore it seems that the MESS model does not generalise well to the FIRST dataset according to prognostic separation, again irrespective of the method of imputation used.

Tables of the conditional seizure risks for FIRST via these four methods of imputation can be seen in Appendix D - they were almost identical to those presented for FIRST in Table 36.

Table 37: Summary of Results of external validation using MESS with EEG and CT variables having 2 categories each and the FIRST dataset

	MESS	FIRST (random selection)	FIRST (single imputation)	FIRST (hot deck)	FIRST (multiple imputation)
Deviance					
Statistic	27.09	24.70	23.04	25.43	24.40
Degrees of Freedom	-	6	6	6	6
p-value	-	0.88	0.67	0.95	0.85
Concordance					
Statistic (95% CI)	0.587 (0.552, 0.622)	0.648 (0.596, 0.700)	0.651 (0.597, 0.705)	0.655 (0.601, 0.709)	0.648 (0.595, 0.701)
Difference in Concordance	-	0.06	0.06	0.07	0.06
Prognostic Separation					
Statistic	0.587	0.853	0.845	0.931	0.932
Difference (& 95% CI)	-	-0.27 (-0.29, -0.24)	-0.26 (-0.29, -0.23)	-0.35 (-0.37, -0.32)	-0.35 (-0.32, -0.27)

9.4.3 Conclusion

The conditional seizure risks for FIRST are considerably lower than they are for MESS. Irrespective of the method of imputation chosen, comparison of deviance suggests that the MESS model generalises well to the FIRST dataset. This was also confirmed by the comparison of concordance.

The comparison of prognostic separation suggests that the MESS model does not generalise well to the FIRST data however. In the simulation study of Chapter 8, prognostic separation was found to be an inadequate method of external validation. Therefore the poor validation seen here may be as a result of the method of external validation rather than the validity of the model.

9.5 Missing Data – Comparing Simulated and Real Results

In the FIRST dataset there was a variable missing. Therefore, in section 9.4, the five methods of handling missing covariates tested in the simulation study of Chapter 8 were applied. All five performed fairly equally, whereas the simulation study results suggested that variable matching was a poor method of imputation whilst the others were acceptable.

The NGPSE and WA datasets are complete and therefore provide an opportunity to compare the 'true' results with results from the five methods of imputation using real data rather than simulated data. There are only two types of covariate in the MESS model - binary and categorical. One of each of these was forced to be missing in turn from the NGPSE and WA datasets. The five methods of imputation were then applied and the deviance, concordance and separation statistics were calculated. This enabled us to compare the results of the imputed data with the truth, and further investigate methods for handling covariates missing from the validation dataset.

9.5.1 Results

The binary variable chosen to be missing from the MESS model was aetiology and the categorical variable was EEG result. External validation results for each scenario can be seen in Table 38, Table 39, Table 40 and Table 41.

9.5.1.1 Missing Binary Covariate

In the case of a missing binary covariate, the deviance, concordance and prognostic separation measures all showed good agreement between the complete NGPSE dataset and each of the imputed versions. The dataset that had the binary variable replaced via multiple imputation showed closest agreement with the complete dataset according to the deviance and concordance measures. The prognostic separation measure showed closest agreement with the complete dataset when the missing binary variable was imputed by random selection with replacement.

For a binary variable forced to be missing from the WA dataset, the deviance and concordance measures showed good agreement between the complete dataset and each of the imputed versions. In both these cases the closest agreement between the complete and imputed datasets occurred when the binary variable was replaced via multiple imputation. The measure of prognostic separation showed quite poor agreement between the complete and imputed datasets irrespective of the method of imputation – all confidence intervals for a difference between the complete and imputed datasets did not include the value of no difference, 0. However, the differences in prognostic separation were fairly small (≤0.05 in all cases) with the difference between the complete dataset and the one imputed via hot deck imputation being the smallest.

9.5.1.2 Missing Categorical Variable

When a categorical covariate was forced to be missing from the validation dataset, the concordance statistic showed good agreement between the complete dataset and the

imputed versions across all five methods of imputation for both the NGPSE and WA datasets. In the case of NGPSE the closest agreement was found when multiple imputation was implemented. For WA the closest agreement was as a result of hot deck imputation.

For NGPSE, the deviance measures were comparable across the complete and imputed datasets with random selection with replacement leading to the closest agreement. For the WA dataset, the deviance measure showed varying levels of agreement – all methods of imputation except hot deck imputation led to significant p-values. There was consequently sufficient evidence at the 10% level to conclude no agreement between the complete and imputed datasets. Therefore the deviance measure may not be an appropriate method of assessing external validation. It could also be that the methods of handling the missing covariate are not appropriate in this case. This may be due to the large size of the WA dataset, the level of correlation between the variables, or some other reason.

With regards to prognostic separation, the only non-significant difference was for the NGPSE dataset imputed via random selection with replacement. None of the differences in prognostic separation between the complete WA dataset and each imputed version were non-significant. The prognostic separation measure therefore suggested there was no agreement between the complete and imputed datasets for the WA dataset. Again, this suggests that the prognostic separation measure may be inappropriate, but also it may be the case that the methods of imputation are inappropriate in this case.

Table 38: Methods of Imputation tested on a missing binary variable in NGPSE

Missing Binary - NGPSE	Complete	Variable Matching	Random Selection	Single Imputation	Hot Deck	Multiple Imputation
Deviance						
Statistic	10.26	N/A	9.64	10.13	9.7	10.24
Degrees of Freedom	-	N/A	8	8	8	8
p-value	-	N/A	1.00	1.00	1.00	1.00
Concordance						
Statistic	0.601 (0.547, 0.654)	0.597 (0.543, 0.651)	0.596 (0.541, 0.651)	0.596 (0.541, 0.651)	0.595 (0.540, 0.650)	0.598 (0.544, 0.652)
Difference in Concordance	-	0.004	0.005	0.005	0.006	0.003
Prognostic Separation						
Statistic	0.634	0.622	0.639	0.647	0.624	0.645
Difference (& 95% CI)	-	0.01 (-0.02, 0.05)	0.00 (-0.04, 0.03)	-0.01 (-0.05, 0.02)	0.01 (-0.02, 0.04)	-0.01 (-0.05, 0.02)

Table 39: Methods of Imputation tested on a missing binary variable in WA

Missing Binary - WA	Complete	Variable Matching	Random Selection	Single Imputation	Hot Deck	Multiple Imputation
Deviance						
Statistic	45.4	N/A	39.65	38.93	38.94	39.66
Degrees of Freedom	-	N/A	8	8	8	8
p-value	-	N/A	0.68	0.59	0.60	0.68
Concordance						
Statistic	0.590 (0.563, 0.618)	0.579 (0.551, 0.607)	0.581 (0.554, 0.609)	0.579 (0.551, 0.607)	0.580 (0.552, 0.607)	0.598 (0.544, 0.652)
Difference in Concordance	-	0.011	0.009	0.011	0.010	-0.008
Prognostic Separation						
Statistic	0.554	0.505	0.513	0.507	0.522	0.511
Difference (& 95% CI)	-	0.05 (0.03, 0.06)	0.04 (0.03, 0.06)	0.05 (0.03, 0.06)	0.03 (0.02, 0.05)	0.04 (0.03, 0.06)

Table 40: Methods of Imputation tested on a missing categorical variable in NGPSE

Missing Categorical - NGPSE	Complete	Variable Matching	Random Selection	Single Imputation	Hot Deck	Multiple Imputation
Deviance						
Statistic	10.26	N/A	11.02	7.29	6.75	8.27
Degrees of Freedom	-	N/A	8	8	8	8
p-value	-	N/A	1.00	0.94	0.90	0.98
Concordance						
Statistic	0.601 (0.547, 0.654)	0.578 (0.525, 0.632)	0.592 (0.537, 0.648)	0.578 (0.526, 0.631)	0.579 (0.524, 0.634)	0.585 (0.531, 0.639)
Difference in Concordance	-	0.023	0.009	0.023	0.022	0.005
Prognostic Separation						
Statistic	0.634	0.506	0.657	0.576	0.548	0.588
Difference (& 95% CI)	-	0.13 (0.10, 0.16)	-0.02 (-0.06, 0.01)	0.06 (0.03, 0.09)	0.09 (0.06, 0.12)	0.05 (0.01, 0.08)

Table 41: Methods of Imputation tested on a missing categorical variable in WA

Missing Categorical - WA	Complete	Variable Matching	Random Selection	Single Imputation	Hot Deck	Multiple Imputation
Deviance						
Statistic	45.4	N/A	30.29	29.99	35.84	31.27
Degrees of Freedom	-	N/A	8	8	8	8
p-value	-	N/A	0.06	0.05	0.30	0.08
Concordance						
Statistic	0.590 (0.563, 0.618)	0.578 (0.550, 0.605)	0.579 (0.550, 0.607)	0.578 (0.550, 0.605)	0.582 (0.554, 0.610)	0.578 (0.550, 0.606)
Difference in Concordance	-	0.012	0.011	0.012	0.008	0.012
Prognostic Separation						
Statistic	0.554	0.446	0.457	0.450	0.495	0.463
Difference (& 95% CI)	-	0.11 (0.09, 0.12)	0.10 (0.08, 0.11)	0.10 (0.09, 0.12)	0.06 (0.04, 0.07)	0.09 (0.07, 0.11)

9.5.2 Conclusion

We have once again demonstrated that the concordance measure most frequently shows agreement when it is known to be present. Consequently it seems that concordance is the best method of external validation. In terms of handling covariates missing from a validation dataset, all five methods of imputation are fairly comparable, with multiple imputation being the most consistent method across a missing binary or categorical variable.

Given that there are slight differences in the results for missing binary and missing categorical variables, it may not be sensible to infer these results to a missing continuous variable. Ideally, another prognostic model, which contains a binary, categorical and continuous variable, should be obtained together with a suitable validation dataset. The analysis in section 9.5.1 can then be repeated and the performance of the methods of external validation and methods for handling covariates missing from the validation dataset can then be compared across all types of missing data. Guidelines for handling one missing variable of each type may then be produced with possible extensions to more than one missing covariate.

9.6 Summary

We have attempted to validate a prognostic model for time to second seizure following a first ever seizure, in the context of patients who drive, with the NGPSE, FIRST and WA datasets. Following on from the simulation study of Chapter 8, three methods of external validation were tested together with five methods of imputation where necessary.

In the simulation study, concordance was found to correctly identify agreement between the validation and development datasets when they were simulated in the same way. In this case study we found, based on the concordance statistic, that the model developed

using MESS generalised well to the NGPSE and WA datasets. The MESS model did not generalise as well to the FIRST dataset (Table 42).

Table 42: Summary of concordance for external validation of MESS model via NGPSE, WA and FIRST datasets

	Concordance	Difference
MESS	0.591 (0.556, 0.627)	NA
NGPSE	0.601 (0.547, 0.654)	0.01
WA	0.590 (0.563, 0.618)	0.00
MESS: No Sleep Variable	0.576 (0.541, 0.612)	NA
FIRST: Variable Matching	0.646 (0.593, 0.699)	0.07
MESS: Sleep Variable	0.587 (0.552, 0.622)	NA
FIRST: Random Selection	0.648 (0.596, 0.700)	0.06
FIRST: Single Imputation	0.651 (0.597, 0.705)	0.06
FIRST: Hot Deck	0.655 (0.601, 0.709)	0.07
FIRST: Multiple Imputation	0.648 (0.595, 0.701)	0.06

In the case of NGPSE and WA where no variables were missing, the prognostic measure of separation (Table 43) showed that the MESS model did generalise but not very well. The comparison of deviance (Table 44) suggested that the model is not externally valid.

Table 43: Summary of prognostic separation for external validation of MESS model via NGPSE, WA and FIRST datasets

	Separation	Difference	Confidence Interval
MESS	0.636	NA	NA
NGPSE	0.634	0.04	(-0.02, 0.03)
WA	0.553	0.05	(0.06, 0.10)
MESS: No Sleep Variable	0.482	NA	NA
FIRST: Variable Matching	0.782	0.30	(-0.34, -0.29)
MESS: Sleep Variable	0.587	NA	NA
FIRST: Random Selection	0.853	0.27	(-0.29, -0.24)
FIRST: Single Imputation	0.845	0.26	(-0.29, -0.23)
FIRST: Hot Deck	0.931	0.35	(-0.37, -0.32)
FIRST: Multiple Imputation	0.932	0.35	(-0.32, -0.27)

In the case of FIRST where a variable was missing, comparison of deviance suggested no evidence to conclude that the MESS model was not externally valid. Comparison of separation suggested that the MESS model did not generalise well. It is likely that this is as

a result of the method of external validation rather than the model and data in question although it could be as a result of underlying differences in the population as outlined in section 9.1.

Table 44: Summary of deviance for external validation of MESS model via NGPSE, WA and FIRST datasets

	Deviance	P-value
MESS	30.91	NA
NGPSE	10.26	0.01
WA	45.40	0.07
MESS: No Sleep Variable	21.11	NA
FIRST: Variable Matching	23.03	0.86
MESS: Sleep Variable	27.09	NA
FIRST: Random Selection	24.70	0.88
FIRST: Single Imputation	23.04	0.67
FIRST: Hot Deck	25.43	0.95
FIRST: Multiple Imputation	24.40	0.85

In the simulation study of Chapter 8 variable matching was found to be a poor method of imputation. Methods random selection with replacement, single imputation via estimation, hot deck imputation and random selection with replacement multiple times were found to be appropriate. In our external validation of MESS all five methods led to poor comparison of separation although the concordance statistic showed good agreement, and p-values for the comparison of deviance were highly non-significant. When the methods of imputation were tested on real data, all five methods were found to be comparable, with multiple imputation being the most consistent method across a missing binary or categorical variable.

In the case of comparison of deviance, a significant p-value suggests that the deviances are not comparable across the development and validation datasets – in other words the development model is not externally valid. A non-significant p-value, on the other hand, only implies that there is insufficient evidence of a difference in deviances. From this, it is

not possible to infer that the models are externally valid [231] – this confusing interpretation of the results provide an additional justification for not using comparison of deviance as a method of external validation (Chapter 8).

The MESS model appears to be externally valid, to different degrees, using different datasets; the model generalises fairly well to the NGPSE and WA datasets and not quite so well to the FIRST dataset. This may be due to underlying population differences - FIRST is the closest match to MESS in terms of the proportions of patients with certain characteristics, however the follow-up is much shorter and a significant covariate is missing. In addition, EEG result and CT/MRI result were categorised differently in FIRST than in MESS and there may be other differences in unmeasured covariates.

Data used to externally validate a prognostic model should come from the same super-population as the development data. There are however, no guidelines suggesting how to check whether the development and validation datasets are from a super-population. From our analyses it appears that length of follow-up is not too important, as the results of the external validation were unaffected when we truncated NGPSE to match the length of follow-up in MESS. However, the conditional risk estimates became considerably higher than those for the complete NGPSE dataset, and consequently became considerably higher than those for MESS which suggested that truncating was not a sensible option.

If a prognostic model is only externally valid in a population that is too similar, say almost identical, to that used to derive the model then it may not be generaliseable more widely. The three external validation datasets investigated here were plausibly related datasets to MESS but were not a perfect match. Given that the model appears to be externally valid in populations that are slightly different in some respects to the MESS dataset it may be possible to conclude that the model is more useful as a clinical tool.

Inevitably, studies brought together for external validation will differ. This is also the case in systematic reviews and meta-analyses. According to Cochrane guidelines [214], meta-analyses should only be considered when a group of studies is sufficiently homogeneous in terms of participants, interventions and outcomes to provide a meaningful summary. This can be equated to performing external validation when an independent dataset needs to be sufficiently similar to the development dataset to provide a reasonable estimate of validity [8].

The decision of sufficient homogeneity is often based on clinical input in the case of meta-analyses and I recommend a similar process for external validation – if the independent dataset(s) is sufficiently homogenous in terms of participants, interventions and outcomes then it may be suitable for external validation. Obviously, this will vary according to medical area and therefore clinical input would be required.

Chapter 10: Prognostic Models for Remission – Mixture Modelling

10.1 Introduction

In this chapter more complex modelling methods than those used in Chapter 4 will be investigated to assess if a model can be developed which more appropriately accounts for the different patient groups. The chapter begins with a summary of the theory of mixture modelling (section 10.2), especially in the context of survival data (section 10.2.2). Then two alternative methods for modelling time to 12 month remission are presented. In section 10.3.1, the patients are split into two groups - those who achieve immediate 12 month remission and those who achieve delayed 12 month remission. Patients with refractory epilepsy are included in the delayed 12 month remission group as censored observations. Prognostic models for each group are derived. In section 10.3.2, mixture modelling techniques are implemented to derive a single prognostic model for all three patient groups.

In Chapter 7, the prognostic model for time to 12 month remission, derived using a Cox model, showed poor internal validation at one year. This may be because prognostic factors for the patient group achieving remission immediately at 365 days differ from those for the patient group achieving remission after one year.

Considering arm A of the SANAD dataset (described thoroughly in Chapter 4), time to 12 month remission is thought to comprise three distinct sections, as can be seen in Figure 33. The first section is at one year where a spike represents people who achieve remission immediately at 12 months. The second section represents who achieve 12 month remission but not immediately, referred to as the delayed remission group. Finally, the

survival curve flattens out although most Kaplan-Meier curves flatten towards the end of follow-up due to fewer subjects being at risk. However, in the case of epilepsy, clinical knowledge suggests that not all patients achieve remission. Therefore, in Figure 33, the flattening out is likely to represent patients who do not achieve remission, called the refractory epilepsy group (or the 'cure' group in frailty modelling terminology).

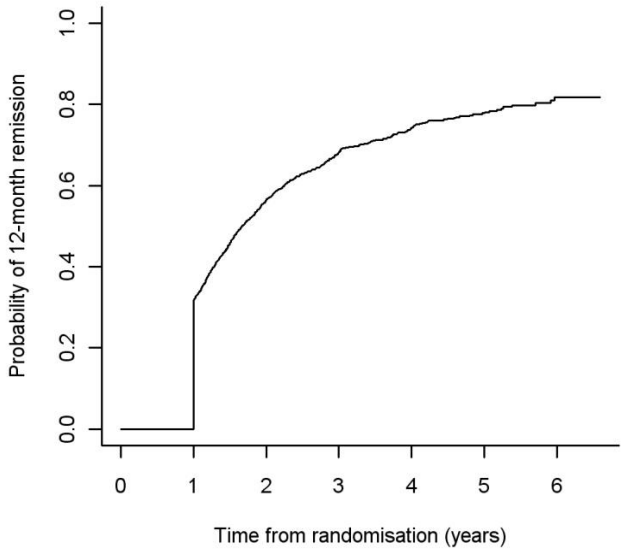


Figure 33: The SANAD Trial – arm A: Kaplan-Meier curve for time to 12 month remission

10.2 Mixture Modelling

Finite mixture models were first proposed over a century ago in the classic paper of Pearson [327]. He fitted a mixture of two univariate Normal components to data on crab measurements provided by his colleague Weldon [328, 329] via moments-based fitting. Pearson's approach [327] involved large amounts of algebra. In the early part of the twentieth century various attempts were made to simplify the method [330]. Over the following 30 years, work continued on the use of the method of moments for this mixture problem - Charlier and Wicksell [331] extended the method to the case of more than two

univariate normal components. In addition, Stromgren [332] considered the use of cumulants. More recently, Cohen [333] showed how an iterative procedure, involving the solving of a cubic equation for a negative root, could be used as an alternative to solve Pearson's [327] nonic. Tan and Chang [334] and Fryer and Robinson [335], and others, showed however that maximum likelihood estimation was, in fact, superior to the methods of moments for this problem. Recent work by Lindsay and Basak [336], amongst others, shows renewed interest in the method of estimation for normal mixtures.

Following the introduction of high-speed computers, interest focused on maximum likelihood estimation of the parameters in a mixture distribution. Rao [337] first used this method for a mixture model applying Fisher's method of scoring for a mixture of two univariate distributions with equal variances. It has, however, been noted [338] that Newcomb [339] predates Pearson's early attempt at mixture models with the method of moments. He suggested an iterative reweighting scheme, which can be interpreted as an application of the EM algorithm of Dempster [306], to compute the maximum likelihood estimate of the common mean of a mixture in known proportions of a finite number of univariate normal populations with known variances.

After Rao's paper [337] maximum likelihood estimation was not pursued further until Hasselblad [340, 341] who addressed the problem, initially for a mixture of g univariate normal distributions with equal variances, and then for mixtures of distributions from the exponential family. Convergence properties of the maximum likelihood solution for the mixture problem were established theoretically after Dempster *et al.* [306] had formalised this iterative scheme in a general context. In addition, their paper proved to be a spring-board for further research into the applications of finite mixture models which is continuing to this day.

10.2.1 General Finite Mixture Models

Finite mixture densities are a family of probability density functions of the form

$$f(\boldsymbol{x};\boldsymbol{p},\boldsymbol{\theta}) = \sum_{i=1}^{c} p_i g_i(\boldsymbol{x};\boldsymbol{\theta}_i)$$

where $\boldsymbol{x}$ is a p-dimensional random variable, $\boldsymbol{p}' = [p_1, p_2, \dots, p_{c-1}]$ and $\boldsymbol{\theta}' = [\boldsymbol{\theta}_1', \boldsymbol{\theta}_2', \dots, \boldsymbol{\theta}_c']$ with the p_i being the mixing proportions, or mixing weights, and the $g_i, i = 1,2, \dots, c$ the component densities of the mixture, with density g_i parameterised by $\boldsymbol{\theta}_i$. The mixing proportions are non-negative and are such that $\sum_{i=1}^{c} p_i = 1$. The number of components forming the mixture is c [342]. In most applications of mixture densities, the g_i are assumed to take the same specified form, such as a univariate Gaussian. In some applications it is better to allow different forms as i varies. A particular case of this, referred to as a nonstandard mixture, is for $c = 2$ where one of the component densities is concentrated on a single value.

Finite mixture densities are frequently used in either the situation where the population, whose distribution is to be modelled, is known to consist of well-defined subpopulations but the individual class memberships are unavailable or too expensive to obtain, or in the situation where subpopulations are only suspected and finite mixture models are used to explore the data for any potentially informative groupings. A simple example of the first scenario would be if clinical measurements were available for patients, but disease classifications were not. In the second case, finite mixture densities act as a relatively sound model for cluster analysis [342].

10.2.1.1 Estimating the Parameters in Finite Mixture Densities

One of the main challenges when applying finite mixture densities is the estimation of the parameters of the mixture, which may or may not include the number of components, c. Given a sample of observations, $\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n$, from the mixture density, the log-likelihood

function, L, is defined as $L = \sum_{i=1}^{n} \ln f(\boldsymbol{x}_i; \boldsymbol{p}, \boldsymbol{\theta})$ [342]. The likelihood equation $\frac{\partial L(\Phi)}{\partial \Phi} = 0$, where $\Phi' = [p', \theta']$, can be solved to give estimates of the parameters in the model. In addition, it has been shown [340, 341, 343-346] that the likelihood equations can be rearranged such that the likelihood estimate of Φ, $\widehat{\Phi}$ satisfies the equations

Equation 22

$$\hat{p}_j = \frac{1}{n} \sum_{i=1}^{n} \hat{P}(j|\boldsymbol{x}_i)$$

and

Equation 23

$$\frac{\sum_{j=1}^{c} \sum_{i=1}^{n} \hat{P}(j|\boldsymbol{x}_i)\, \partial \ln g_j(\boldsymbol{x}_i; \boldsymbol{\theta}_j)}{\partial \widehat{\boldsymbol{\theta}}} = 0$$

where $\hat{P}(j|\boldsymbol{x}_i)$ is the estimated posterior probability of an observation, $\boldsymbol{x}$, arising from component density, j.

In the situation of a mixture where the j^{th} component density is multivariate normal with mean, $\boldsymbol{\mu}_j$, and covariate matrix, $\boldsymbol{\Sigma}_j$, Equation 22 and Equation 23 become [342]

Equation 24

$$\hat{p}_j = \tfrac{1}{n} \textstyle\sum_{i=1}^{n} \hat{P}(j|\boldsymbol{x}_i),$$

Equation 25

$$\widehat{\boldsymbol{\mu}}_j = \tfrac{1}{n\hat{p}_j} \textstyle\sum_{i=1}^{n} \boldsymbol{x}_i\, \hat{P}(j|\boldsymbol{x}_i),$$

and

Equation 26

$$\widehat{\boldsymbol{\Sigma}}_j = \frac{1}{n\hat{p}_j}\sum_{i=1}^{n}(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}_j)(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}_j)'\hat{P}(j|\boldsymbol{x}_i).$$

The estimation of the unknown parameters, Φ, on the basis of the observations $\boldsymbol{x}_i$ is only meaningful if Φ is identifiable [347]. In general, a parametric family of densities $f(\boldsymbol{x}_i;\Phi)$ is identifiable if

Equation 27

$$f(\boldsymbol{x}_i;\Phi) = f(\boldsymbol{x}_i;\Phi^*),$$

if and only if $\Phi = \Phi^*$. Identifiability for mixture distributions is defined in a slightly different way. Suppose that $f(\boldsymbol{x}_i;\Phi)$ has two component densities, say $f_a(\boldsymbol{x}_i;\Phi_\mathrm{a})$ and $f_b(\boldsymbol{x}_i;\Phi_\mathrm{b})$ that belong to the same parametric family. Then Equation 27 will still hold when the component labels a and b are interchanged in Φ. That is, although this class of mixtures may be identifiable, Φ is not [347]. This lack of identifiability is not of concern in the normal course of events in the fitting of mixture models by maximum likelihood, say, via the EM algorithm. However, it does cause problems in a Bayesian framework where posterior simulation is used to make inferences from the mixture model [347].

Hasselblad [340, 341], Wolfe [343-345] and Day [346] all suggested an iterative scheme for solving the likelihood equations. This involved finding initial estimates of the posterior probabilities from given initial values of the parameters of the mixture, and then evaluating the right-hand sides of Equation 24, Equation 25 and Equation 26 to give revised estimates of the parameter values. From these, new estimates of the posterior probabilities were derived and the procedure was repeated until a suitable convergence criterion was satisfied. This procedure is a particular example of the EM algorithm [342], described by Dempster, Laird and Rubin [306] in the context of likelihood estimation for incomplete data problems.

One of the problems of the EM algorithm, noted by McLachlan and Basford [348], is its generally slow convergence rate. Others have highlighted the fact that the algorithm need not converge at the global maximum. Alternative algorithms, such as the Quasi-Newton algorithm, have also been shown to have convergence problems. A more general problem of likelihood estimation for mixture models is that examples can be found where the likelihood is unbounded, so that the maximum likelihood estimate does not exist [342].

10.2.1.2 Detecting Finite Mixture Densities

Finite mixture densities are most applicable in situations where separate groups of observations, with different distributions, seem to exist as a result of the nature of the application. In many real-life situations however, the evidence for modelling data with a mixture distribution will need to be obtained empirically. It is in these situations that the number of components in the mixture, c, is important having determined if a mixture density is an appropriate model for the data. The procedures that have been suggested for answering such questions involve a combination of graphical techniques and quasi-significance tests [342].

In the case of univariate data the simplest way to determine whether a mixture distribution is a suitable model for a sample of observations is to draw a sample histogram; clear multimodality is assumed to provide strong evidence that a mixture distribution is appropriate [342]. However, only in particular circumstances will the mixing of at least two unimodal densities lead to a mixed density with more than a single mode [349]. Therefore, it is important to look for alternatives to the histogram for evidence of a mixture. One possibility, particularly for mixtures with normal components, is to use a Q-Q plot of the data, or a variation of such a plot proposed by Fowlkes [350], which he claims is particularly sensitive to the presence of mixtures.

Fowlkes' suggested procedure is to plot $(x_{(i)} - \bar{x})/s$ against $\Phi\left(\frac{x_{(i)}-\bar{x}}{s}\right) - p_i$ where $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$ represent the ordered sample values, which have mean $\bar{x}$ and standard deviation s, $p_i = (i - 0.5)/n$ and Φ is the standard normal distribution function. When the observations arise from a single normal density, Fowlkes' proposed plot results in an approximately horizontal line at $y = 0$. Mixture densities lead to plots having a characteristic cyclical pattern about zero, which differs from the pattern given by other non-normal distributions [350].

To detect multivariate normal mixtures an option is a chi-squared probability plot of the Mahalanobis distance of each observation from the sample mean vector [342]. If the data are from a single multivariate normal density these distances have an approximately chi-squared distribution with d degrees of freedom, where d is the number of variables. Therefore a chi-squared probability plot of the ordered distances will result in an approximately straight line through the origin. Mixtures of multivariate normals will, however, tend to give plots that are 'S' shaped.

10.2.1.3 Determining the Number of Components in a Mixture

Despite the widespread use of finite mixture models, the decision of how many components are required to adequately represent the data is, according to many authors [351-354] an important problem, but without a satisfactory statistical solution.

A natural candidate for testing the hypothesis $c = c_0$ against $c = c_1$ $(c_1 > c_0)$, where c is the number of components in a mixture density, is the likelihood ratio statistic, λ. However, this does not lead to a suitable significance test, because, for mixture densities, regularity conditions do not hold for $-2\ln\lambda$ to have its usual asymptotic null distribution which is a chi-squared distribution with degrees of freedom equal to the difference in the number of parameters under the competing hypotheses [342]. This is because the null

distribution is on the edge of the parameter space. Despite there being no completely satisfactory method of using the likelihood ratio statistic as a formal significance test for number of components, McLachlan and Basford [348] suggest that Wolfe's modified likelihood ratio test [355] can be used as an informal guide to the appropriate number of components. Wolfe, on the basis of a limited simulation study, suggested that the null distribution of the likelihood ratio statistic, for testing $c+1$ against c components, is chi-squared with $2v-2$ degrees of freedom, where v is the number of extra parameters in the $c+1$ component mixture.

In the context of survival data, Hunsberger *et al* [356] propose a maximum-likelihood approach to estimate the number of mixture components. For Hunsberger's [356] approach to estimating the number of mixture components let y_i be the true failure time for individual i, let c_i be an individual's potential censoring time and d_i be the corresponding censoring indicator. Observations then consist of $t_i = \min(y_i, c_i)$ with $d_i = 1$ if $y_i < c_i$ (uncensored) and $d_i = 0$ if $y_i > c_i$ (censored).

Alternative Bayesian approaches are also possible such as reversible-jump Monte Carlo methods [357] and a Markov-chain Monte Carlo method that treats the parameters of the model as a marked Poisson process [358]. These require bootstrapping to obtain an assessment of the p-value.

Carrying out a hypothesis test based on bootstrapping can be computationally demanding. Therefore the estimation of the order of a mixture model has been considered mainly by using a penalised form of the log-likelihood function – as the likelihood increases with the addition of a component to a mixture model, information criteria such as AIC [163] and BIC [166] attempt to balance the increase in fit obtained against the larger number of parameters estimated for models with more components.

In addition, to ensure that over-parameterisation is accounted for, it is important to ensure that the segments are sufficiently separated. An entropy statistic can be used to evaluate the degree of separation and hence assess the ability of a mixture model to provide well-separated components.

10.2.1.4 Including Covariates in a Finite Mixture Model

Introduce a finite mixture model with m groups where the probability of being in a particular group $j = 1,2,\dots,m$ can be allowed to depend on subject-specific covariates where x_{ik} denotes covariates $k = 1,2,\dots,K$ measured on subject i. Polychotomous logistic regression parameterisation is used as per Equation 28 with $\theta_{i1} = 1 - \sum_{j=2}^{m} \theta_{ij}$, $\boldsymbol{\beta}_j = (\beta_{j0}, \beta_{j1}, \dots, \beta_{jK})'$ and $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \dots, x_{iK})'$. The vector $\boldsymbol{\beta}_j$ contains parameters that characterise the effect of covariates $\boldsymbol{x}_i$ on the probability of being in the j^{th} group.

Equation 28

$$\theta_{ij} = \theta_j(\boldsymbol{\beta}_j, \boldsymbol{x}_j) = \frac{\exp(\beta_{j0} + \sum_{k=1}^{K} x_{ik}\beta_{jk})}{(1 + \sum_{j=2}^{m}(\exp(\beta_{j0} + \sum_{k=1}^{K} x_{ik}\beta_{jk})))}$$

10.2.2 Finite Mixture Models for Survival

Finite mixtures have frequently been used in a variety of fields to model heterogeneous data. The potential of finite mixture models has been extended to survival analysis as they can be used to analyse failure-time data in a variety of situations including in the case of competing risks. In the case of survival analysis, the observations are the times to failure and since failure can be due to a variety of causes, each with its own particular survival time distribution, the overall survival time distribution will be a mixture. Historical attempts to fit negative exponential or Weibull distributions etc. have been described by Mendenhall and Hader [359] and by Kao [360] while, more recently, applications of mixture distributions to survival times have been described by McGiffin *et al* [361] and Lui, Darrow and Rutherford [362].

In general, to specify a survivor function by a mixture model, it is assumed that the probability density function, $f(t)$, of the survival time, T, has the finite mixture form [363]

Equation 29

$$f(t) = \sum_{i=1}^{g} \pi_i f_i(t)$$

where $f_1(t), \ldots, f_g(t)$ denote g component densities occurring in proportions $\pi_1, \ldots, \pi_g$ where $0 \leq \pi_i \leq 1$ $(i = 1, \ldots, g)$. The survival function $S(t)$ corresponding to Equation 29 has the functional form

$$S(t) = \sum_{i=1}^{g} \pi_i S_i(t)$$

where

$$S_i(t) = \int_t^{\infty} f_i(u)du \ \ (i = 1, \ldots, g).$$

For the mixture model specification, Equation 29, of the survivor function, $S(t)$, the corresponding hazard function $h(t)$ is

$$h(t) = -\frac{\partial \log S(t)}{\partial t} = \frac{\sum_{i=1}^{g} \pi_i h_i(t) S_i(t)}{S(t)}$$

where $h_i(t)$ is the hazard function corresponding to the i^{th} component survivor function $S_i(t)$ in the model. $h(t)$ can be written as

Equation 30

$$h(t) = \sum_{i=1}^{g} \pi_i \lambda_i(t)$$

where

$$\lambda_i(t) = \frac{h_i(t)S_i(t)}{S(t)} = \frac{h_i(t)\exp\left\{-\int_0^t h_i(u)du\right\}}{\sum_{j=1}^{g}\pi_j \exp\left\{-\int_0^t h_i(u)du\right\}} \ (i = 1, \ldots, g).$$

Thus, the imposition of a mixture model on the survivor function implies a mixture model for the hazard function of the form in Equation 30 [363].

Suppose that $h_i(t)$ has the proportional hazards form

$$h_i(t; \boldsymbol{x}) = \exp(\boldsymbol{\gamma}_i'\boldsymbol{x})h_{oi}(t)$$

where $\boldsymbol{x}$ is the vector of covariates and $\lambda_{0i}(t)$ is the baseline hazard function that does not depend on $\boldsymbol{x}(i = 1, \ldots, g)$. Then

Equation 31

$$\lambda_i(t; \boldsymbol{x}) = \frac{\exp(\boldsymbol{\gamma}', \boldsymbol{x})h_{oi}(t)\{S_{oi}(t)\}^{d_i}}{\sum_{j=1}^{g}\pi_j\{S_{oi}(t)\}^{d_i}}$$

where

$$S_{0i}(t) = \exp\left\{-\int_0^t h_{0i}(u)du\right\}$$

and

$$d_i = \exp(\boldsymbol{\gamma}', \boldsymbol{x})$$

for $i = 1, \ldots, g$. It can be seen from Equation 31 that although $h_i(t; \boldsymbol{x})$ has a proportional hazards form, $\lambda_i(t; \boldsymbol{x})$ does not in general [363].

In the context of cure mixture models with survival data Greenhouse and Silliman [364] use a parametric cure model to analyse a dataset where the patients who are not cured are modelled according to a Weibull distribution. The approach of Hunsberger [356] is similar to that of Greenhouse and Silliman but it potentially allows for more than two mixture

components. Both use a Weibull distribution to model the survival in the non-cured groups.

In Hunsberger [356] the survivor function for the i^{th} patient is

Equation 32

$$S(t|\boldsymbol{x}_i, \boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{p}) = \pi_{i1}F_1(t|\lambda_1, p_1) + \pi_{i2}F_2(t|\lambda_2, p_2) + \cdots + \pi_{im}F_m(t|\lambda_m, p_m)$$

where $F_j(t|\lambda_j, p_j)$ is a Weibull survival distribution indexed by parameters λ_j, p_j also $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_m)'$ and $\boldsymbol{p} = (p_1, p_2, \ldots, p_m)'$. Alternatively, a cure component can replace a Weibull distribution in this formulation. For example a cure component can replace the Weibull distribution for group 1 by replacing $F_1(t|\lambda_1, p_1)$ with $F_1(t) = 1$ in Equation 32. Models are identifiable up to a permutation of the group labels. Therefore, the groups are ordered in terms of the median survival in each group.

Hunsberger [356] uses maximum likelihood to estimate the model parameters. As an example, the likelihood, $\boldsymbol{L}(\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{p})$, for the model with three groups, two Weibull and a cure component is

Equation 33

$$\boldsymbol{L}(\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{p}) = \prod_{i=1}^{n} [\pi_{i2}f_2(t_i|\lambda_2, p_2) + \pi_{i3}f_3(t_i|\lambda_3, p_3)]^{d_i} \times [\pi_{i1} + \pi_{i2}F_2(t_i|\lambda_2, p_2) + \pi_{i3}F_3(t_i|\lambda_3, p_3)]^{(1-d_i)}$$

where $f_2(t|\lambda_2, p_2)$ and $f_3(t|\lambda_3, p_3)$ are different Weibull distributions and $F_2(t|\lambda_2, p_2)$ and $F_3(t|\lambda_3, p_3)$ are the respective cumulative distribution functions. The maximum likelihood estimates are found by maximising $\log(\boldsymbol{L}(\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{p}))$ with respect to the set of unknown parameters. The parameter estimates are obtained using a quasi Newton-Raphson algorithm, which is available in the software package 'gauss' [365]. Hypothesis testing for

assessing the effect of covariates on the probability of being in a particular group is conducted with likelihood ratio tests.

10.2.2.1 Long-Term Survivor Model (Proportion Cured Model)

In some situations where the aim is to estimate the survival distribution for a particular type of failure, a certain fraction of the population, say π_1, may never experience this type of failure. It is characterised by the overall survival curve being levelled at a non-zero probability. In some applications, the surviving fractions are said to be cured [347] and the estimation of π_1 is of interest. Although it can be difficult to distinguish between a cure group and survival functions with long tails [366], Yu *et al* [367] showed that the cure fraction could be estimated well if follow-up was longer than the median survival.

Assume that an individual has probability $\pi_2 = 1 - \pi_1$ of failing from the cause of interest and probability π_1 of never experiencing failure from this cause. Further, assume that the individual cannot fail from any other cause during the course of the study. Let T be a random variable denoting the time to failure, where $T = \infty$ denotes the event that the individual will not fail from the cause of interest. The probability of this latter event is π_1 [347]. The unconditional survival function of T can then be expressed as

Equation 34

$$S(t) = \pi_1 + \pi_2 S_2(t)$$

where $S_2(t)$ denotes the conditional survival function for failure from the cause of interest. The mixture model described in Equation 34, with the first component having mass one at $T = \infty$, can be regarded as a nonstandard mixture model [347] and is used in a variety of applications in survival analysis. This model is sometimes referred to as the long-term survival mixture model as individuals who will never fail can be viewed as being long-term survivors. This approach enables the proportion, π_1, of patients suffering no relapse

of the disease to be estimated without needing to specify a survivor function for those patients who die without evidence of a relapse of the disease under study. This mixture approach, which was adopted by Farewell [368] and by McLachlan *et al* [369], among others, provides an alternative to the use of the Kaplan-Meier estimate.

In more recent work involving this class of mixture models, Kuk and Chen [370] proposed a semi-parametric generalisation of the parametric model of Farewell. Their model also assumes a proportional hazards model for the component hazard function $h_2(t)$,

$$h_2(t, \boldsymbol{x}) = \exp(\boldsymbol{\theta}_2' \boldsymbol{x}) h_{02}(t),$$

but the baseline hazard function can be any arbitrary unspecified function not necessarily in the Weibull or Gompertz families. Estimates of the regression parameters are obtained by maximising a Monte Carlo approximation of a marginal likelihood and the EM algorithm is used to estimate the baseline survivor function.

Usually, for the long-term survival mixture model to be applicable, the observed data are of the form $\left(t_j, \delta_j\right)^T$ for the j^{th} entity where $\delta_j = 2$ implies that the j^{th} individual was observed to fail from the cause of interest at time t_j during the following up period, and $\delta_j = 0$ implies that the failure time was censored at time t_j [347]. In this context, Equation 34 can be fitted by maximum likelihood after the specification of a parametric model $S_2(t; \theta_2)$ for $S_2(t)$.

An individual who fails from the cause of interest at time t_j $(j = 1, \dots, n)$ contributes a likelihood factor $\pi_2 f_2(t_j; \theta_2)$ where $f_2(t; \theta_2)$ is the density of T corresponding to $S_2(t; \theta_2)$. An individual who has been followed to time t_j without failure contributes a likelihood factor $\pi_1 + \pi_2 S_2(t_j; \theta_2)$ which is the probability that an individual never experiences failure. The log likelihood, $\log L(\psi)$, for $\psi = (\pi_1, \theta_2^T)^T$ is then given by

$$\log L(\psi) = \sum_{j=1}^{n} \left[I_{[2]}(\delta_j) \log\{\pi_2 f_2(t_j; \theta_2)\} + I_{[0]}(\delta_j) \log\{\pi_1 + \pi_2 S_2(t_j; \theta_2)\} \right]$$

10.2.2.2 Fitting by Maximum Likelihood

In order to fit the mixture model, defined by Equation 29, to the observed failure-time data maximum likelihood can be used such that the form of each component density is specified up to a manageable number of unknown parameters. These parameters are then inferred by considering the likelihood function that can be formed under these specified parametric forms [363].

Let $f_i(t; \boldsymbol{x}, \boldsymbol{\theta}_i), S_i(t; \boldsymbol{x}, \boldsymbol{\theta}_i)$ and $h_i(t; \boldsymbol{x}, \boldsymbol{\theta}_i)$ denote the parametric forms of the density, survivor and hazard functions respectively with respect to the i^{th} component in the mixture representation, Equation 29. Here, $\boldsymbol{\theta}_i$ is a vector containing the unknown parameters in the parametric representation of these quantities for the i^{th} component of the mixture. The vector of all unknown parameters is denoted by $\boldsymbol{\psi}$. When parametric forms are adopted for the mixing proportions $\pi_i, \dots, \pi_g$, $\boldsymbol{\psi}$ is given by $\boldsymbol{\psi} = (\pi_1, \dots, \pi_{g-1}, \boldsymbol{\theta}')'$ where $\boldsymbol{\theta} = (\boldsymbol{\theta}_1', \dots, \boldsymbol{\theta}_g')'$.

In some situations, the mixing proportions π_i are modelled as logistic functions of the covariate vector $\boldsymbol{x}$. In this case, the vector $\boldsymbol{\psi}$ of all unknown parameters is given by $\boldsymbol{\psi} = (\boldsymbol{\alpha}', \boldsymbol{\theta}')'$. With $\boldsymbol{\psi}$ defined in either way, the log-likelihood function for $\boldsymbol{\psi}$ is given by

$$\log L(\boldsymbol{\psi}) = \sum_{j=1}^{n} \delta_j \log f(t_j; \boldsymbol{x}, \boldsymbol{\psi}) + \sum_{j=1}^{n} (1 - \delta_j) \log S(t_j; \boldsymbol{x}_j, \boldsymbol{\psi})$$

where $f(t; \boldsymbol{x}, \boldsymbol{\psi})$ and $S(t; \boldsymbol{x}, \boldsymbol{\psi})$ denote the mixture density and survivor functions with components modelled parametrically [363]. The maximum likelihood estimator $\widehat{\boldsymbol{\psi}}$ of $\boldsymbol{\psi}$ is then obtained as an appropriate root of the likelihood equation

Equation 35

$$(d \log L(\boldsymbol{\psi}))/d\boldsymbol{\psi} = 0.$$

10.2.2.3 Application of the EM Algorithm

Rather than solving the likelihood equation, Equation 35, directly, solutions can be found by applying Dempster *et al.*'s EM algorithm [306]. This algorithm can be applied in the same manner in the context of survival analysis as for mixture models in general, with modifications to allow for the presence of censored data [363]. On the $(k+1)$th iteration of the EM algorithm, $\boldsymbol{\psi}^{(k+1)}$ is given by the value of $\boldsymbol{\psi}$ that globally maximises

$$\mathbb{Q}\left(\boldsymbol{\psi}, \boldsymbol{\psi}^{(k)}\right) = E_{\boldsymbol{\psi}^{(k)}}\{\log L_c(\boldsymbol{\psi}) | \boldsymbol{y}\},$$

which is the conditional expectation of the complete-data log-likelihood, $\log L_c(\boldsymbol{\psi})$, given the observed data $\boldsymbol{y}$ and the current fit $\boldsymbol{\psi}^{(k)}$ for $\boldsymbol{\psi}$. The calculation of this equation corresponds to the *E*-step while its maximisation with respect to $\boldsymbol{\psi}$ corresponds to the *M* step of the EM algorithm [363], as described in section 8.3.2.6.

10.3 Modelling Time to 12 Month Remission

The SANAD time to 12 month remission data was investigated in two ways. In section 10.3.1 a non-mixture modelling approach was used while in section 10.3.2 a three component mixture model was derived.

10.3.1 Non-Mixture Modelling

The outcome immediate 12 month remission was assessed by logistic regression. For each prognostic factor for this outcome, the log-odds ratio and 95% confidence interval was calculated. Model fits were tested by chi-square methods. Of the 1620 patients suitable for analysis in arm A, 32 had no follow-up data and 74 (5%) had a follow-up of less than 365 days, 12 months. 47 of these patients had a seizure during follow-up and were therefore included in the denominator

The outcome time to 12 month remission conditional on not achieving it immediately at 365 days was assessed by life table analysis according to Kaplan-Meier. The prognostic values of factors for these outcomes were tested by log-rank methods. For each factor the hazard ratio and associated 95% confidence interval were calculated. Best-fitting, parsimonious, Cox multivariable models were produced with backwards elimination by AIC – selection starts with the full model and eliminates predictors one at a time, at each step considering whether the criterion will be improved, smaller AIC, by adding back in a variable removed at a previous step [178]. Of the 1514 patients with a follow-up of at least 365 days, 471 had a remission time of exactly 365 days and were consequently removed from the dataset. Therefore 1043 patients were analysed to determine prognostic factors.

10.3.1.1 Immediate 12 Month Remission

The outcome immediate 12 month remission is equivalent to a time to 12 month remission of 365 days. Univariate and multivariable odds ratios can be seen in Table 45.

In the univariate model the following factors were significantly associated with immediate 12 month remission: gender, treatment history, neurological insult, epilepsy type, seizure type, focal epilepsy site of onset, age, time from first seizure to randomisation, total number of seizures ever before randomisation and treatment. The resulting parsimonious multivariable model included variables for gender, treatment history, age, time from first seizure to randomisation, neurological insult, total number of seizures ever before randomisation and treatment, which was forced into the model.

Immediate 12 month remission was more likely in men than women (male vs. female: HR 1.06 95% CI (1.01 to 1.11)) and less likely in patients who were taking non-SANAD AEDs at randomisation compared to those who were treatment naive (non-SANAD AEDs vs. naive: HR 0.90 95% CI (0.85 to 0.96)). Immediate 12 month remission was also less likely in patients with a neurological insult (present vs. absent: HR 0.92 95% CI (0.86 to 0.99)) and in patients taking gabapentin rather than carbamazepine (gabapentin vs. carbamazepine: HR 0.89 95% CI (0.83 to 0.95)). Immediate 12 month remission was increasingly likely with increasing age while it was less likely with an increase in the number of seizures ever before randomisation.

Table 45: The SANAD Trial – arm A: Univariate and Multivariable Odds Ratios for Immediate 12 Month Remission

			Odds Ratio (95% Confidence Interval)	
Prognostic Factor	Comparison	P-value (univariate)	Univariate (n=1513)	Multivariable (n=1513)
Gender	Female Male	0.00	1.00 1.42 (1.14 to 1.77)	1.00 1.06 (1.01 to 1.11)
Febrile seizure history	Absent Present	0.13	1.00 0.69 (0.39 to 1.17)	NA
First degree relative with epilepsy	Absent Present	0.13	1.00 0.78 (0.54 to 1.13)	NA
CT/MRI scan result	Normal Abnormal Not Done	0.20	1.00 1.03 (0.79 to 1.33) 1.18 (0.87 to 1.61)	NA
Treatment history	Treatment naive Seizures after remission Taking non-SANAD AEDs	0.00	1.00 1.30 (0.63 to 2.62) 0.43 (0.29 to 0.60)	1.00 1.09 (0.94 to 1.27) 0.90 (0.85 to 0.96)
Age (years) [Linear]	≤10 (10 to 24) (25 to 36) (37 to 49) (50 to 70) >71	0.00	1.00 1.15 (1.08 to 1.22) 1.39 (1.21 to 1.61) 1.66 (1.34 to 2.07) 2.10 (1.53 to 2.88) 2.73 (1.78 to 4.18)	1.00 1.03 (1.01 to 1.04) 1.06 (1.03 to 1.09) 1.10 (1.05 to 1.15) 1.15 (1.08 to 1.22) 1.20 (1.10 to 1.31)
Time from 1st seizure (months) [Log]	≤2 (2 to 5) (6 to 17) (18 to 59) (60 to 239) >240	0.00	1.00 0.99 (0.99 to 1.00) 0.98 (0.96 to 0.99) 0.93 (0.89 to 0.97) 0.92 (0.87 to 0.97) 0.57 (0.40 to 0.81)	1.00 1.00 (1.00 to 1.00) 1.00 (1.00 to 1.01) 1.01 (1.00 to 1.02) 1.01 (1.00 to 1.02) 1.09 (1.01 to 1.18)
Neurological insult	Absent Present	0.00	1.00 0.56 (0.38 to 0.82)	1.00 0.92 (0.86 to 0.99)
Total number of seizures [Scale: divided by 100, power 0: log]	≤2 (2 to 3) (4 to 11) (12 to 50) (51 to 299) >300	0.00	1.00 0.92 (0.91 to 0.93) 0.76 (0.72 to 0.79) 0.57 (0.52 to 0.63) 0.40 (0.34 to 0.47) 0.22 (0.17 to 0.29)	1.00 0.99 (0.98 to 0.99) 0.95 (0.94 to 0.96) 0.90 (0.89 to 0.92) 0.85 (0.82 to 0.87) 0.76 (0.72 to 0.79)
EEG results	Normal Missing Non-specific abnormality Epileptiform abnormality	0.17	1.00 1.05 (0.70 to 1.55) 1.10 (0.80 to 1.50) 0.87 (0.67 to 1.13)	NA
Seizure type	Simple or complex focal only 2° Generalised tonic-clonic Uncertain	0.00	1.00 1.31 (1.03 to 1.69) 2.60 (1.78 to 3.79)	NA
Epilepsy type	Focal Unclassified	0.00	1.00 2.16 (1.53 to 3.04)	NA
Focal epilepsy site of onset	Temporal Not localised Frontal Other Unclassified	0.00	1.00 1.04 (0.81 to 1.35) 0.94 (0.58 to 1.50) 0.86 (0.51 to 1.40) 2.17 (1.49 to 3.16)	NA
Treatment	Carbamazepine Gabapentin Lamotrigine Oxcarbazepine Topiramate	0.01	1.00 0.56 (0.40 to 0.78) 0.76 (0.55 to 1.04) 0.94 (0.64 to 1.39) 0.83 (0.60 to 1.15)	1.00 0.89 (0.83 to 0.95) 0.94 (0.88 to 1.00) 0.99 (0.91 to 1.07) 0.96 (0.90 to 1.03)

10.3.1.2 Time to Delayed 12 Month Remission

The outcome time to delayed 12 month remission is equivalent to time to 12 month remission conditional on not having achieved it at 365 days. Univariate and multivariable hazard ratios can be seen in Table 46.

In the univariate model CT/MRI results, treatment history, seizure type, time from first seizure to randomisation and total number of seizures ever before randomisation were significantly associated with delayed 12 month remission. The resulting parsimonious multivariable model included variables for gender, CT/MRI result, treatment history, seizure type, total number of seizures ever before randomisation and treatment which was forced into the model.

Delayed 12 month remission was less likely in patients with an abnormal CT/MRI result compared to a normal result (abnormal vs. normal: HR 0.81 95% CI (0.66 to 0.99)) and in patients who were taking non-SANAD AEDs at randomisation compared to those who were treatment naive (non-SANAD AEDs vs. naive: HR 0.67 95% CI (0.53 to 0.86)). Delayed 12 month remission was decreasingly likely with an increase in the number of seizures ever before randomisation.

Table 46: The SANAD Trial – arm A: Univariate and Multivariable Hazard Ratios for Time to Delayed 12 Month Remission

			Hazard Ratio (95% Confidence Interval)	
Prognostic Factor	Comparison	P-value (univariate)	Univariate (n=1043)	Multivariable (n=1043)
Gender	Female Male	0.24	1.00 1.11 (0.94 to 1.31)	1.00 1.13 (0.96 to 1.35)
Febrile seizure history	Absent Present	0.20	1.00 1.25 (0.90 to 1.74)	NA
First degree relative with epilepsy	Absent Present	0.39	1.00 1.12 (0.87 to 1.46)	NA
CT/MRI scan result	Normal Abnormal Not Done	0.03	1.00 0.79 (0.65 to 0.97) 1.09 (0.85 to 1.39)	1.00 0.81 (0.66 to 0.99) 1.09 (0.85 to 1.40)
Treatment history	Treatment naive Seizures after remission Taking non-SANAD AEDs	0.00	1.00 0.95 (0.52 to 1.73) 0.60 (0.47 to 0.76)	1.00 0.98 (0.53 to 1.79) 0.67 (0.53 to 0.86)
Age (years) [Linear]	≤10 (10 to 24) (25 to 36) (37 to 49) (50 to 70) >71	0.12	1.00 0.96 (0.48 to 1.93) 0.94 (0.30 to 2.92) 0.93 (0.23 to 3.67) 0.92 (0.18 to 4.62) 0.91 (0.15 to 5.48)	NA
Time from 1st seizure (months) [Log]	≤2 (2 to 5) (6 to 17) (18 to 59) (60 to 239) >240	0.00	1.00 0.95 (0.93 to 0.98) 0.92 (0.87 to 0.97) 0.88 (0.81 to 0.95) 0.88 (0.81 to 0.95) 0.82 (0.72 to 0.93)	NA
Neurological insult	Absent Present	0.09	1.00 0.81 (0.62 to 1.04)	NA
Total number of seizures [Scale: 100, power: 0]	≤2 (2 to 3) (4 to 11) (12 to 50) (51 to 299) >300	0.00	1.00 0.98 (0.97 to 0.99) 0.93 (0.90 to 0.96) 0.86 (0.81 to 0.92) 0.79 (0.71 to 0.87) 0.67 (0.56 to 0.80)	1.00 0.98 (0.97 to 0.99) 0.94 (0.91 to 0.97) 0.88 (0.82 to 0.94) 0.81 (0.73 to 0.90) 0.70 (0.58 to 0.84)
EEG results	Normal Missing Non-specific abnormality Epileptiform abnormality	0.59	1.00 0.89 (0.63 to 1.25) 1.10 (0.86 to 1.41) 1.08 (0.89 to 1.31)	NA
Seizure type	Simple or complex focal only 2° Generalised tonic-clonic Uncertain	0.03	1.00 0.86 (0.72 to 1.03) 1.26 (0.92 to 1.73)	1.00 0.83 (0.69 to 1.00) 1.05 (0.75 to 1.46)
Epilepsy type	Focal Unclassified	0.07	1.00 1.34 (0.99 to 1.81)	NA
Focal epilepsy site of onset	Temporal Not localised Frontal Other Unclassified	0.14	1.00 0.93 (0.77 to 1.12) 1.07 (0.76 to 1.49) 1.27 (0.91 to 1.76) 1.32 (0.96 to 1.83)	NA
Treatment	Carbamazepine Gabapentin Lamotrigine Oxcarbazepine Topiramate	0.29	1.00 0.87 (0.68 to 1.12) 1.06 (0.83 to 1.35) 1.02 (0.73 to 1.42) 0.83 (0.64 to 1.08)	1.00 0.84 (0.66 to 1.08) 1.04 (0.82 to 1.33) 1.00 (0.72 to 1.38) 0.84 (0.65 to 1.09)

10.3.1.3 Discussion

A prognostic model for overall time to 12 month remission was derived in Chapter 4. It included variables for gender, CT/MRI scan result, treatment history, age, time from first seizure to randomisation, neurological insult, total number of seizures before randomisation, focal epilepsy site on onset and treatment. All these variables were found to be significant prognostic factors for either the model for immediate 12 month remission (section 10.3.1.1), or time to delayed 12 month remission (section 10.3.1.2), or both.

Gender, treatment history, total number of seizures, site of onset and treatment were included in both the model for immediate 12 month remission, and the model for delayed remission. In the immediate remission model, age, time from first seizure and neurological insult were significant factors while in the delayed model CT/MRI result and seizure type were significant.

As there were some differences in the prognostic factors for immediate 12 month remission and those for delayed remission it is possible that the model developed in Chapter 4 with these patient groups combined is not the most appropriate model. This may explain why the internal validation of the model developed in Chapter 4 was poor at one year – patients with immediate 12 month remission have different significant clinical factors to those with delayed 12 month remission and therefore need to be modelled independently.

It is difficult to provide clinical reasoning for the difference in the results when comparing variables identified for immediate and delayed 12 month remission. If anything, this work highlights the need for further work into the biology of epilepsy. This research may then explain why age is important for predicting immediate remission but not delayed remission while time from first seizure and total number of seizures are similar across both outcomes. One reason may be that the immediate and delayed analyses comprise different number of

patients. Therefore, the delayed analyses have less power to detect the effects of prognostic factors.

10.3.2 Mixture Modelling

As the SANAD time to 12 month remission data comprised three distinct patient groups, a three component mixture model for the outcome was investigated. There were several stages to the derivation process – existence of a cure fraction, distribution of the survival data, covariate selection and likelihood formulation. These are described in sections 10.3.2.1 to 10.3.2.4.

10.3.2.1 Stage 1: Is there a cure fraction?

In the case of the refractory group, assume that an individual has probability $\pi_2 = 1 - \pi_1$ of achieving 12 month remission and probability π_1 of never achieving 12 month remission. Additionally, assume that the individual cannot fail from any other cause during the course of the study.

In the case of SANAD, follow-up may be insufficient to determine a cure fraction. A crude way to determine if there is a cure fraction is to ignore patients who achieved 12 month remission immediately. Then, a normal cure rate model can be fitted to the delayed and potentially refractory patient group using standard cure rate methodology via the 'gfcure' programme within R [371].

Table 47: The SANAD Trial – arm A: Cure Fractions

	Cure Fraction
Exponential	0.08%
Weibull	29.8%
Log-logistic	30.4%

The distribution of the delayed remission group is unknown. Therefore, the test for a cure fraction was performed using exponential, Weibull and log-logistic baseline distributions. The cure fractions for each distribution can be seen in the Table 47.

The 'gfcure' package uses an extended generalised gamma distribution which, with appropriate reparameterisation, includes most commonly used distributions as special cases. The boundary conditions of the likelihood ratio test are not violated with this distribution and the standard asymptotic theory of the likelihood ratio test can be applied [372].

Table 48: The SANAD Trial – arm A: Log-likelihoods for potential model fits to the data

	Log-likelihood		
	Survival Model	Cure Rate Model	Cure Fraction Present?
Exponential	-4656.467	-1055.915	Yes
Weibull	-4544.611	-896.314	Yes
Log-logistic	-4494.850	-853.357	Yes

Therefore, to test whether these cure fractions were significant, a form of the log-likelihood was used for the inclusion of the cure rate. A model was fitted with all relevant covariates included, as per the multivariable analysis of time to 12 month remission for the SANAD data (Chapter 4), and a cure fraction. This was then compared to a standard model fitted to the survival data only, with all the same covariates. The results, shown in Table 48, suggest that there is a cure fraction, irrespective of the distribution. The log-logistic model appeared to be the better fit based on these results.

10.3.2.2 Stage 2: Distribution of Delayed Remission Group

To determine the distribution of the patients who achieve 12 month remission but not immediately at 12 months, the so-called delayed remission group, the exponential, Weibull and log-logistic distributions were considered. To begin with, the survivor functions were estimated and then graphs were plotted to assess whether the data appeared to follow an

underlying exponential, Weibull or log-logistic distribution. Next, a straight line was fitted to these graphs to estimate the model parameters. Maximum likelihood estimation of the parameters was then carried out and the results compared to the estimates from the straight line. The fitted distribution was then compared to the actual data.

The exponential distribution has density function $f(t) = \lambda \exp(-\lambda t)$, hazard function $h(t) = \lambda$ and survivor function $S(t) = \exp(-\lambda t)$ where λ is a positive constant. It therefore follows that $-\log[S(t)] = \lambda t$. To test if the exponential distribution fits the data, $-\log[\hat{S}(t)]$ can be plotted against t. If the lifetime distribution underlying the data is exponential, then the plot will, approximately, be a straight line through the origin, gradient λ.

The Weibull distribution has density function $f(t) = \lambda\gamma t^{\gamma-1} \exp(-\lambda t^{\gamma})$, hazard function $h(t) = \lambda\gamma t^{\gamma-1}$ and survivor function $S(t) = \exp(-\lambda t^{\gamma})$ where γ, λ are positive constants. It therefore follows that $\log(-\log[S(t)]) = \log\lambda + \log t$. $\log(-\log[\hat{S}(t)])$ can be plotted against $\log t$ and if the lifetime distribution is Weibull, then the plot will approximately be a straight line, gradient γ, intercept $\log\lambda$ [373].

The log-logistic distribution has the density function $f(t) = \frac{\gamma p(pt)^{\gamma-1}}{(1+(pt)^{\gamma})^2}$, hazard function $h(t) = \frac{\gamma p(pt)^{\gamma-1}}{1+(pt)^{\gamma}}$ and survivor function $S(t) = \frac{1}{1+(pt)^{\gamma}}$ where γ, p are positive constants. It therefore follows that $\log\left(\frac{1}{\hat{S}(t)} - 1\right) = \gamma\log p + \gamma\log t$. If the lifetime distribution is log-logistic, the plot of $\log\left(\frac{1}{\hat{S}(t)} - 1\right)$ against $\log t$ should approximately give a straight line with intercept $\gamma\log p$ and gradient γ [373].

The three plots can be seen in Figure 34. None of the plots are a perfect straight line.

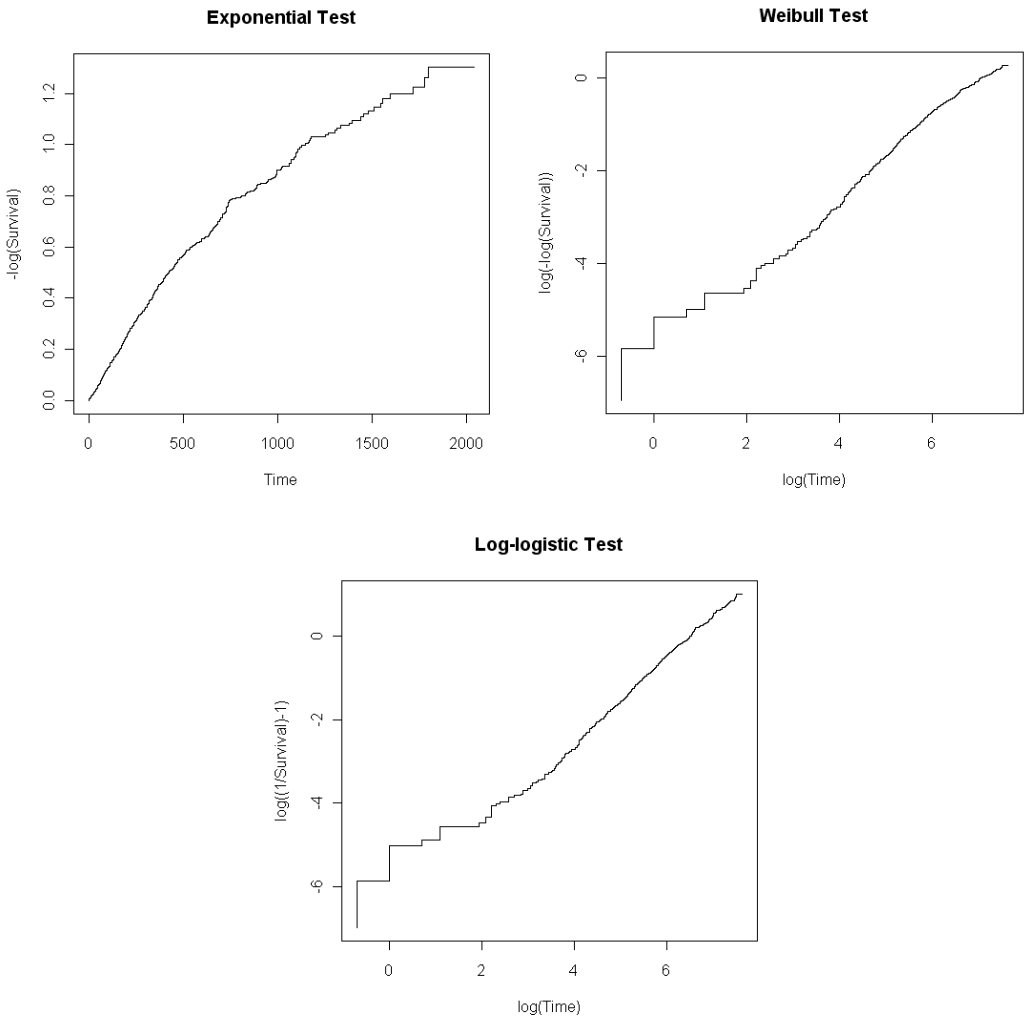


Figure 34: The SANAD Trial – arm A: Plots to test fit of various models to the data

Having fitted a straight line to each plot, the intercepts and gradients are as shown in Table 49. The parameters were additionally estimated by the maximum likelihood method and can also be seen in Table 49.

In the case of the straight line estimates, $\lambda = \exp(\text{intercept})$ and $\gamma = \text{gradient}$. In the case of the maximum likelihood estimates, the output from R for the estimated survivor function is in parametric form with the relationship between this and the estimates for a straight line model being $\text{Intercept} = \log\left(\frac{1}{\lambda}\right) = -\log(\lambda)$, and $\text{Scale} = {}^{1}/_{\gamma}$. Therefore,

$\lambda = \exp(-\text{Intercept})$ and $\gamma = {}^{1}/_{\text{Scale}}$ [373]. The parameters can now be compared, as per Table 49.

For an exponential fit, the gradient is required to be one. This is not the case according to our parameter estimates suggesting that the exponential is not a suitable model.

Table 49: The SANAD Trial – arm A: Parameter estimates for various model fits to the data

Distribution	Straight Line Estimates				Maximum Likelihood Estimates			
	Intercept	Gradient	λ	γ	Intercept	Scale	λ	γ
Exponential	0.1547	0.0007	1.1673	0.0007	7.4539	1.000	0.0006	1.0000
Weibull	-5.447	0.775	0.0043	0.775	7.2910	0.608	0.0007	1.6447
Log-logistic	-6.540	1.006	0.0014	1.006	6.9604	0.431	0.0009	2.3202

The fit was then evaluated by plotting the estimated survivor function along with the model estimates as shown in Figure 35. The unusual cure fraction of 0.08% together with the exponential fit plot in Figure 35 shows, convincingly, that a constant hazard was not a sensible assumption. This distribution is therefore not appropriate for the delayed remission group. The log-logistic model fits the data better than the Weibull. Therefore, the log-logistic distribution appears to be the best choice of model for the delayed refractory group.

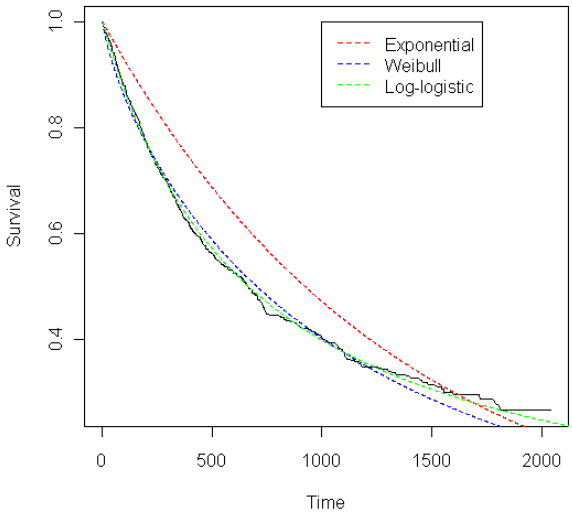


Figure 35: The SANAD Trial – arm A: Model Fit plots; Black line represents the data

Mixture modelling was therefore undertaken to model patients who achieved 12 month remission immediately (logistic regression), those who achieved 12 month remission but not immediately at 12 months (parametric survival model with log-logistic distribution) and those who did not achieve 12 month remission (refractory patient group i.e. cure fraction). The effects of covariates were considered to investigate whether predictions could be made about which remission group a patient fell into based on their baseline characteristics.

10.3.2.3 Step 3: Covariates

10.3.2.3.1 Immediate Remission

With 12 possible covariates there are 2^{12}=4096 possible combinations of variables to try in the model. Although there are pre-set routines available in standard software packages to try all these combinations with logistic regression and survival analysis, they are not available for cure rate models. Therefore Collett [167] recommends a general strategy for model selection.

In step one, each possible variable is fit, one at a time, and the resulting $-2\log\hat{L}$ for each model is compared with that for the null model. This determines which variables are significant independently. In step two, the variables that appear to be important in step 1 are then fitted together. The change in the value of $-2\log\hat{L}$ is computed when each variable on its own is omitted from the set. Only those that lead to a significant increase in the value of $-2\log\hat{L}$ are retained in the model. Once a variable has been dropped, the effect of omitting each of the remaining variables in turn is then examined.

In the third step variables that were not important on their own, and so were not under consideration in step 1 may become important in the presence of others. These variables are therefore added to the model from step 2, one at a time, and any that reduce $-2\log\hat{L}$

significantly are retained in the model. In the last step, a final check is made to ensure that no term in the model can be omitted without significantly increasing the value of $-2\log\hat{L}$, and that no term not included significantly reduces $-2\log\hat{L}$.

Before Collett's strategy was employed, the functional form for the continuous variables in SANAD was determined. In line with Chapter 4, our preferred transformations were linear, log or FP. However, FP transformations are not available with cure rate models currently and consequently were not considered here.

Table 50: The SANAD Trial – arm A: Coefficients and Standard Errors for Immediate Remission

Variable	Coefficient	Standard Error	Comparison	Odds Ratio (95% CI)
Intercept	0.39	0.05	-	N/A
Time from First Seizure	0.04	0.01	≤2 (2 to 5) (6 to 17) (18 to 59) (60 to 239) >240	1.00 1.02 (1.01 to 1.03) 1.04 (1.03 to 1.06) 1.06 (1.04 to 1.09) 1.09 (1.05 to 1.12) 1.16 (1.09 to 1.23)
Total Number of Seizures	-0.10	0.01	≤2 (2 to 3) (4 to 11) (12 to 50) (51 to 299) >300	1.00 0.95 (0.95 to 0.96) 0.91 (0.90 to 0.93) 0.86 (0.84 to 0.88) 0.80 (0.77 to 0.83) 0.71 (0.68 to 0.75)
Treatment: GBP	-0.12	0.03	GBP vs. CBZ	0.89 (0.84 to 0.95)
Treatment: LTG	-0.06	0.03	LTG vs. CBZ	0.94 (0.88 to 1.00)
Treatment: OXC	-0.01	0.04	OXC vs. CBZ	0.99 (0.91 to 1.07)
Treatment: TPM	-0.03	0.03	TPM vs. CBZ	0.97 (0.91 to 1.03)

Following Collett's general strategy for model selection, the final model included variables for time from first seizure to randomisation, total number of seizures ever before randomisation and treatment, which was forced into the model, with coefficient values, standard errors and associated odds ratios and confidence intervals as in Table 50. The chance of immediate remission increases with an increase in the time from first seizure to randomisation but decreases with an increase in total number of seizures before

randomisation. Immediate remission is less likely in patients treated with gabapentin than in those treated with carbamazepine.

10.3.2.3.2 Delayed Remission

Collett's strategy was used to determine the parsimonious multivariable model for delayed epilepsy. As before, every patient was treated, therefore treatment was forced into the model. The final model included variables for treatment history, time from first seizure to randomisation, total number of seizures ever before randomisation and treatment. The coefficients, standard errors, odds ratios and confidence intervals can be seen in Table 51.

The log-logistic model is both an accelerated failure time model and a proportional odds model [374]. Hence the regression coefficients can be interpreted as the log odds ratio (for surviving) with one unit increase in the variable.

Table 51: The SANAD Trial – arm A: Coefficients and Standard Errors for Delayed Remission

Variable	Coefficient	Standard Error	Comparison	Odds Ratio (95% CI)
Log(Scale)	-0.89	0.03	-	N/A
Intercept	6.66	0.08	-	N/A
Treatment History: Recent Seizures	0.05	0.18	Recent Seizures vs. Treatment Naive	1.06 (0.74 to 1.51)
Treatment History: Taking non-SANAD AEDs	0.25	0.07	Taking non-SANAD AEDs vs. Treatment Naive	1.28 (1.11 to 1.48)
Time from First Seizure	0.00	0.00	≤2 (2 to 7) (8 to 18) (19 to 60) (61 to 277) >277	1.00 1.00 (1.00 to 1.00) 1.00 (1.00 to 1.00) 0.99 (0.97 to 1.02) 0.97 (0.87 to 1.07) 0.92 (0.73 to 1.17)
Total Number of Seizures	0.07	0.02	≤2 (2 to 3) (4 to 11) (12 to 50) (51 to 299) >300	1.00 1.03 (1.02 to 1.05) 1.07 (1.03 to 1.10) 1.11 (1.06 to 1.17) 1.17 (1.09 to 1.26) 1.27 (1.13 to 1.43)
Treatment: GBP	0.09	0.07	GBP vs. CBZ	1.09 (0.94 to 1.26)
Treatment: LTG	-0.03	0.07	LTG vs. CBZ	0.97 (0.84 to 1.12)
Treatment: OXC	0.00	0.09	OXC vs. CBZ	1.00 (0.83 to 1.21)
Treatment: TPM	0.11	0.08	TPM vs. CBZ	1.12 (0.96 to 1.30)

Delayed remission was more likely in patients with recent seizures following previous remission on monotherapy than in treatment naive patients. The chance of delayed remission increased with an increase in total number of seizures before randomisation.

10.3.2.3.3 Refractory Epilepsy

The final model for the refractory epilepsy group just included treatment – no covariates were found to be significant for this group. Standard errors were not estimable from the cure part of the model but the coefficients and associated odds ratios can be seen in Table 52. It is unknown why the standard errors are not estimable – Dr Peng, the author of 'gfcure' was unable to suggest why this anomaly occurred.

Table 52: The SANAD Trial – arm A: Coefficient values for Refractory Epilepsy

Variable	Coefficient	Comparison	Odds Ratio
Intercept	2.54	-	NA
Treatment: GBP	0.18	GBP vs. CBZ	1.20
Treatment: LTG	0.47	LTG vs. CBZ	1.60
Treatment: OXC	0.26	OXC vs. CBZ	1.29
Treatment: TPM	0.40	TPM vs. CBZ	1.49

10.3.2.4 Step 4: Setting Up the Likelihood

There are several mixture models published in the literature such as [375] and [376]. For epilepsy, there are examples of seizure count analyses via mixture modelling [377, 378] but no models for remission from seizures. Now the component densities and covariates have been determined for each patient group (immediate remission, delayed remission and refractory) we wish to combine them to produce one overall model. To do this we need to specify the survivor function by a mixture model.

All times were transformed so that the adjusted time to 12 month remission for a patient with immediate 12 month remission was 0, i.e. $T_i = 0$.

- Let individual, i, have time to 12 month remission T_i where T_i = time to 12 month remission – 365 days.
- Let δ_{1i} be an indicator variable which is equal to 1 if time to 12 month remission is 0 and 0 otherwise, equivalent to an indicator variable for immediate 12 month remission.
- Let δ_{2i} be an indicator variable which is equal to 1 if T_i is observed (and greater than 0) and 0 if T_i is censored (and greater than 0). This is equivalent to an indicator variable for delayed 12 month remission.
- Let R_i be an (imaginary) indicator variable for the refractory group – it is imaginary as it is not possible to determine if a patient is refractory. We can only determine if they achieved 12 month remission.
- Let π_i be the probability that δ_{1i} equals 1, i.e. the probability of achieving 12 month remission immediately.
- Let p_i equal the probability that $T_i \to \infty$ which is equivalent to the probability that R_i is equal to one, i.e. the probability of having refractory epilepsy.

The function, $f_T(t_i, \delta_{1i}, \delta_{2i}, R_i)$ can be considered in three sections which correlate with the three shapes of the curve – the peak related to immediate 12 month remission, the curve related to achieving 12 month remission but not immediately at 12 months ('delayed'), and the flattened line related to patients who do not achieve 12 month remission ('refractory'). In Equation 36, $h_T(t_i)$ is the hazard function for time period t_i and $S_T(t_i)$ is the survival function. The middle two parts of the function relate to the observed time to delayed 12 month remission (2nd line) and censored time to delayed 12 month remission (3rd line).

Equation 36

$$f_T(t_i, \delta_{1i}, \delta_{2i}, R_i) = \begin{cases} \pi_i, & \text{if } t_i = 0, \delta_{1i} = 1 \\ (1-p_i)h_T(t_i)S_T(t_i), & \text{if } t_i > 0, \delta_{1i} = 0, \delta_{2i} = 1, R_i = 0 \\ (1-p_i)S_T(t_i), & \text{if } t_i > 0, \delta_{1i} = 0, \delta_{2i} = 0, R_i = 0 \\ p_i, & \text{if } t_i > 0, \delta_{1i} = 0, \delta_{2i} = 0, R_i = 1 \end{cases}$$

The likelihood function for i is

$$l_i(t_i, \delta_{1i}, \delta_{2i}, R_i) = \pi_i^{\delta_{1i}}(1-\pi_i)^{1-\delta_{1i}}\left\{\left[(1-p_i)h(t_i)^{\delta_{2i}}S(t_i)\right]^{1-R_i}\left[p^{1-\delta_{2i}}\right]^{R_i}\right\}^{1-\delta_{1i}}.$$

The log-likelihood is therefore

$$\log[l_i(t_i, \delta_{1i}, \delta_{2i}, R_i)]$$

$$= \delta_{1i}\log\pi_i + (1-\delta_{1i})\log(1-\pi_i)$$

$$+ (1-\delta_{1i})\log\left\{\left[(1-p_i)h(t_i)^{\delta_{2i}}S(t_i)\right]^{1-R_i}\left[p^{1-\delta_{2i}}\right]^{R_i}\right\}$$

$$= \delta_{1i}\log\pi_i + (1-\delta_{1i})\log(1-\pi_i) + (1-\delta_{1i})(1-R_i)\log\left[(1-p_i)h(t_i)^{\delta_{2i}}S(t_i)\right]$$

$$+ (1-\delta_{1i})R_i\log(p^{1-\delta_{2i}})$$

$$= \delta_{1i}\log\pi_i + (1-\delta_{1i})\log(1-\pi_i) + (1-\delta_{1i})(1-R_i)[\log(1-p_i) + \delta_{2i}\log h(t_i) +$$

$$\log S(t_i)] + (1-\delta_{1i})R_i\,(1-\delta_{2i})\log(p).$$

From this log-likelihood parameter estimates may be obtained. This requires use of an EM algorithm such as Newton-Raphson [379]. Implementation of these algorithms would need to be done in a software package such as Matlab [380] and is a complex stage of the analysis. Alternatively, software packages such as Latent GOLD [381] could be used to fit the mixture model but these require a new programming language to be learnt. Additionally, the relevant covariates need to be included into the model which is again, very complex, involving either a one-step full information maximum-likelihood approach, standard three-step approach or two adjusted three-step approaches [382]. Literature on

these methods is currently in press and was therefore unavailable for use as part of this thesis.

10.3.2.5 Step 5: Is the assumption valid?

In all the analyses above we have assumed that there were no withdrawals before 365 days. If some patients were censored with time to 12 month remission less than 365 days then further investigation is necessary.

There were 77 patients who withdrew before 365 days. The minimum time to withdrawal was 17 days while the maximum was 364 days with median 258 days (IQR 154 to 329). Of these 77 patients, 12 withdrew their consent, 22 died and 43 provided no reason for withdrawal.

To handle the potential invalidity of the assumption a sensitivity analysis could be performed considering the coefficient estimates at six months, nine months, 12 months and 15 months. If the coefficients were broadly similar then there would be some stability in the model so the invalidity would be less concerning. However, if the assumption was not valid then the negative binomial, Poisson or similar model would need to be invoked but this would require more research as the methodology needed is not currently available.

10.3.2.6 Discussion

Analyses in section 10.3.2.1 suggested the presence of a cure fraction (refractory patient group) for the time to 12 month remission outcome of the SANAD dataset. The log-logistic distribution was found to be the most appropriate distribution for the delayed remission group.

Significant covariates for immediate 12 month remission were time from first seizure to randomisation, total number of seizures before randomisation and treatment, which was forced into the model. For delayed 12 month remission, the significant covariates were

treatment history, time from first seizure to randomisation, total number seizures before randomisation and treatment, which was again forced into the model. No covariates were found to be significant for refractory epilepsy so only treatment was included in the model.

The cured population (refractory group) may be more like the delayed 12 month remission population so that the variables included in the model may be more like those identified in Table 46 for overall time to delayed 12 month remission. This is because patients who achieve immediate 12 month remission are unlikely to be similar to those who never achieve remission. However, those who achieve remission but at a later time point than 12 months may take 5 years to achieve remission, or longer, and hence are more likely to be similar to the refractory group who have not achieved remission whilst under observation. The impact of this on these results is unknown but could be explored in sensitivity analyses.

Fewer covariates were found to be significant via the mixture modelling approach (section 10.3.2) than both the non-mixture modelling approach (section 10.3.1) and the modelling all data together approach (Chapter 4). Given that the models in section 10.3.1 were derived via stepwise selection with AIC, and those in this section (10.3.2) were derived using Collett's strategy, it is unsurprising that the same significant covariates were not identified. Additionally, a different modelling approach has been used which may also account for the different covariate list. If the same strategies were employed to determine both the mixture and non-mixture models, the covariates may be the same. This is because the immediate remission group is the same in both approaches. Additionally, no significant covariates were identified for the refractory group via the mixture modelling approach. Therefore, despite the refractory patients being unidentifiable during the non-mixture modelling approach, this is unlikely to impact on the covariates.

Although a direct comparison of results across the mixture modelling and non-mixture modelling approaches is not appropriate, broad comparisons can be made. For patients

who achieve remission immediately at 12 months results are quite similar – time from first seizure to randomisation is moderately significant via the mixture modelling approach and slightly significant for the non-mixture modelling approach. The relationship between chance of immediate remission and total number of seizures before randomisation is the same across both approaches – as number of seizures increases, chance of immediate remission decreases. Immediate remission is less likely in patients on gabapentin than in those on carbamazepine, irrespective of the modelling method. Therefore, these prognostic factors appear to be robust to the choice of modelling method.

For time to delayed 12 month remission, remission was more likely in patients who were previously treated with non-SANAD AEDs but had a change in drug indicated than in treatment naive patients via the mixture modelling approach. In the non-mixture modelling results this conclusion was reversed – remission was less likely in patients who were previously treated with non-SANAD AEDs but had a change in drug indicated than in treatment naive patients. The effect of total number of seizures before randomisation was inconsistent across the mixture and non-mixture modelling methods. In the non-mixture modelling case, as number of seizures increased the chance of remission decreased while in the mixture modelling case the chance of remission increased.

The mixture modelling work is based on parametric modelling techniques while the non-mixture modelling work is based on semi-parametric modelling techniques. If parametric methods were used to model the immediate and delayed remission groups in the non-mixture approach the results may be more similar to those from the mixture modelling approach.

10.3.3 Future Work

Many issues remain in this area of my research. To begin with we need to determine why there is a difference between the maximum likelihood and straight line estimates (Table

49). It is most likely that the difference is caused by the influential curve up until log(time) three. To investigate whether this conjecture is correct, the fitted values could be plotted against actual times, taking account of the censoring. If the values do not match then this suggests our assumption is correct. Another option is to simulate data with both sets of λ and γ and then plot the two related lines and compare them with the actual fit.

Assuming that a mixture model for time to 12 month remission, adjusted for covariates, can be developed it will be necessary to validate it both internally and externally. No literature on how to validate mixture models exists. It may be possible to ensure that the covariates chosen for the immediate 12 month remission group are robust by performing a sensitivity analysis, checking the covariates by Cox's method at nine months and at 15 months. If the same covariates were found to be included in the multivariable model the choice of covariates may be robust. If not, further work may be required into covariate selection. For the delayed 12 month remission and the refractory groups, it may again be possible to perform a kind of internal validation via a sensitivity analysis. In this case the included covariates could be modelled together with any known clinically significant variables simultaneously, via Cox's method, for the delayed and refractory parts of the model.

10.4 Conclusion

Substantial progress has been made in the fitting of mixture models, especially by the method of maximum likelihood, since mixture models were first proposed in the 19th Century. Despite the arrival of high-speed computers, people have been generally unwilling to fit mixture models to data of more than one dimension [347]. This may be because of a lack of understanding of issues that arise with their fitting such as the presence of multiple maxima in the mixture likelihood function and the unboundedness of

the likelihood function in the case of normal components with unequal covariance matrices [347]. As the difficulties of these computational issues come to be properly understood and addressed there is hope for an increase in the use of mixture models in practice.

The first 50 years of using finite mixture densities, following Pearson's pioneering work in 1894, was focussed on the development of techniques designed to simplify the large amount of arithmetic involved in applying the method of moments in estimating parameters. During the last 20 years the wide availability of powerful computers has lessened these problems and maximum likelihood estimation can now be applied routinely to find parameter estimates for mixtures of normal and of other densities [342]. Mixtures continue, however, to provide a rich source of material for statisticians and recent work includes the development of models for data consisting of both continuous and categorical variables [383, 384], the linking of mixture models with hidden Markov chain models and with artificial neural networks [385, 386], the use of influence-based diagnostics for normal mixtures [387, 388] and further consideration of the number of components problem [389].

In section 10.3.1 prognostic models for immediate 12 month remission and delayed 12 month remission were developed. Variables found to be significant for these models were similar to those included in the overall model for time to 12 month remission in Chapter 4. Given that the immediate and delayed models in section 10.3.1 were found to have fairly distinct significant variables it is unsurprising that the model in Chapter 4 did not validate well internally, especially at one year – modelling techniques need to take account of the different patient groups.

Section 10.3.2 extended the work of section 10.3.1 by considering a mixture model for the time to 12 month remission outcome which considered the three patient groups independently, but together in one model. There were very few covariates found to be

significant for this model however a different variable selection technique was employed than used in Chapter 4 and section 10.3.1. Initial work suggests that mixture modelling may provide a reliable way of modelling the three distinct 'types' of remission – immediate, delayed and refractory, however, to date this is ongoing work due to the complexity of the modelling. The next stage is to derive the likelihood inclusive of the chosen covariates. From this estimates of risk may be obtained.

The tools required to derive a mixture model are complex and the literature on how to use mixture models is not always available. However, as we have shown in this thesis, if distinct groups are known to be present in the data they should be modelled independently. We have investigated two methods of doing this, one with mixture modelling methods and one without. The advantage of the non-mixture method is that it is computationally fairly simple, however in the case of epilepsy, the refractory patient group cannot easily be identified – a cure model could be fitted to the model that did not achieve immediate remission, however this is a form of mixture modelling. Therefore, mixture modelling techniques are required if prognostic models for time to 12 month remission from seizures is to be appropriately modelled.

Chapter 11: Conclusions and Further Work

11.1 Introduction

Predicting outcome in epilepsy is complex but necessary to inform patient counselling about prognosis and treatment choices. This thesis has examined prognostic factors for patients with a first seizure, with newly diagnosed epilepsy, and with established epilepsy or frequent unremitting seizures despite optimum treatment. It has been shown that very few prognostic models exist for epilepsy and within this thesis prognostic factors have been modelled for these patient groups for outcomes including seizure recurrence, seizure remission and treatment failure.

Prognostic models for time to treatment failure and time to 12 months of remission, developed using The SANAD Trial (Chapter 4), have been published in Lancet Neurology, which is owned by The Lancet. I am first author on this paper. The work within the publication has been actively discussed at several epilepsy meetings and conferences.

Patients with epilepsy are subject to driving restrictions if their risk of a seizure in the next 12 months is in excess of 20%. We have produced prognostic models from which we can estimate the length of time until the risk of a seizure is under this risk threshold for patients with a first ever seizure, and also for patients who withdraw treatment after a period of remission.

The prognostic model for patients with a first seizure (Chapter 5) has been published in the British Medical Journal (BMJ) and I am first author. In addition, the editors of the BMJ felt that the paper was sufficiently important to warrant an editorial, especially as it led the DVLA to change their guidelines for time off driving for patients with a first seizure.

The prognostic models for risk of recurrence after treatment withdrawal following a period of remission (Chapter 6) have been published in the Journal of Neurology, Neurosurgery and Psychiatry, which is owned by the BMJ. Again, I am first author of the paper and the DVLA are currently considering changing their guidelines following this publication.

Validation is an integral part of the modelling process and therefore internal validation methods have been applied to selected models in this thesis. A search of the literature revealed that very few prognostic models are externally validated, possibly due to a lack of guidance on suitable methods, or possibly due to the lack of appropriate datasets on which to perform external validation. A simulation study was therefore undertaken to assess methods of externally validating a prognostic model and also to assess methods for handling covariates missing from the validation dataset. The investigated methods were also demonstrated in a motivating example.

The prognostic model for time to 12 month remission for patients with newly diagnosed epilepsy was found to be internally valid via calibration plots at two and three years, but not at one year. It is possible that this is because of the distinct patient groups present for the outcome – some patients achieved remission immediately at 365 days, some achieved remission at a later time point, and some did not achieve remission. Mixture modelling techniques were thought to be appropriate for this scenario; they were therefore employed to attempt to develop a combined prognostic model which included all patient groups simultaneously, adjusted for covariates.

11.2 Recommendations for Practice

In this thesis strategies for model development and presentation, and strategies for model validation have been considered. In addition, prognostic models for epilepsy have been

developed which were shown to be fairly well validated both internally and externally. They, consequently, may be recommended for practice.

11.2.1 Model Development

Many methods need to be considered when developing a prognostic model. For example, variables should be centred to reduce multicollinearity and also to increase the relevance of the estimated regression coefficient [152]. Similarly, when model building, variable selection techniques such as AIC [163] must be carefully decided upon to balance the need for simplicity with a minimal loss of information. When only one of several different types of event can occur, the probability of these events must be modelled via competing risk techniques such as cumulative incidence analysis [177] with covariate testing via Gray's method [174].

When developing prognostic models, continuous covariates should not usually be dichotomised [390] and instead modelled via spline [254] or FP [144] transformations. Splines can be poorly behaved in the tails [141] and are very complex to write down [145]. However, FPs offer increased flexibility over polynomial transformations while still preserving the simplicity of the final model [184]. FPs were first proposed by Royston and Sauerbrei in 1994 [144] and are consequently relatively new. Therefore there may be elements to these methods which are not understood properly yet, especially with regard to treatment-covariate interactions. Consequently, although we recommend that continuous covariates be modelled via FP transformations, some caution is still required when using them.

11.2.2 Model Presentation

Prognostic models are developed to inform patient counselling and treatment choices. Therefore they must be easy to interpret by clinicians and patients alike. It is thus

inadvisable to present statistical models for clinical interpretation. Instead, we recommend presenting forest-style plots for combinations of risk factors as shown in this thesis. These can be used to illustrate the proportion of patients with the event of interest at various time points, for any combination of risk factors together with confidence intervals.

11.2.3 Model Validation

For a prognostic model to be implemented in general practice it needs to be validated. Internal validation assesses model validity for the setting where the development data originated from, and external validation assesses validity in other samples which are fully independent from the development data and originate from different but plausibly related settings [8].

Methods of internal validation are readily available and include an assessment of model performance such as R^2 [226] together with discrimination [248] and calibration [251, 252] methods. They must be applied to each prognostic model developed.

Literature on methods of external validation is lacking, as shown by our review. However, in our simulation study we discovered that Harrell's c-statistic [184], a measure of concordance, is a reliable method of highlighting agreement between the development and validation datasets when it exists. The simulation study has also shown that a covariate missing from the validation dataset may be imputed via several adaptations of methods for handling missing data within a covariate. The best performing methods were hot deck imputation and multiple imputation. Therefore external validation should be implemented for each prognostic model developed, assuming there is a plausibly related dataset available. If a suitable dataset is not available, attempts should be made to recruit a relevant cohort.

11.3 Recommendations for Research

This thesis has highlighted areas where further research is recommended particularly with regards to model development and validation.

11.3.1 Model Development

Investigation of possible treatment-covariate interactions for the time to treatment failure model using the SANAD dataset suggested two such interactions - focal epilepsy site of onset with treatment, and total number of seizures ever before randomisation with treatment. Although these terms could easily have been included in the multivariable model, interpretation of the results was difficult due to the 150 possible combinations of interaction terms. Current literature for interpreting regression coefficients for multiple inter-linked interaction terms is lacking. Further research into how to interpret hazard ratios and confidence intervals for large numbers of interaction terms is required.

Mixture modelling techniques, enabling all patient groups to be combined in a single prognostic model, are becoming increasingly popular. Our initial work suggests that mixture modelling may provide a way of modelling patients with immediate remission, delayed remission and refractory epilepsy together in one prognostic model. However, due to the complexity of the modelling, further research is required particularly with regards to deriving the likelihood and obtaining risk estimates. Additionally, further research is required into the estimation of standard errors for the group of refractory patients (Chapter 10) using the 'gfcure' package. It is currently unknown why the errors are unestimable.

Assuming a suitable mixture model for time to 12 month remission can be developed, it will be necessary to validate it both internally and externally. Literature on how to validate a mixture model is currently lacking – a brief search revealed no articles that mentioned

validation of a mixture model. Further research into internal and external validation methods of mixture models is required.

11.3.2 Model Validation

In our simulation study, which assessed methods of external validation and methods for handling covariates missing from the validation dataset, eight scenarios were considered in which the size of the validation dataset and the type of missing data were varied. Clearly there are many other possible scenarios including varying the number and combination of covariates present and missing, the types of covariates present and missing, the size of the development and validation datasets, how closely the super-population assumption is adhered to, and the length of follow-up etc. During the case study extension to the simulation study, length of follow-up of the NGPSE dataset was truncated to match the length of follow-up of the MESS study. In this single case, truncating NGPSE did not affect the level of external validation. This suggests that length of follow-up may not have an impact on the external validity of a model. However, the conditional seizure risks were found to be much higher in the truncated dataset. Further research is required to build upon the initial findings presented in this thesis and to investigate alternative scenarios.

In the simulation study discrimination was investigated via Harrell's *c*-statistic, calibration via deviance, and separation via Royston's measure of prognostic separation. However, alternative methods exist to investigate these areas such as calibration plots [391, 392], ROC curves [272] and the Hosmer-Lemeshow statistic [262]. These alternative methods need to be investigated to see if they show agreement between the validation and development datasets when it is known to be present. If so, then a battery of tools for external validation could be produced. However, interpretation of the methods of external validation is not always easy – the methods can give conflicting results and there are different degrees of agreement between the development and validation datasets rather

than clear cut “agreement” or “no agreement” conclusions. This makes it difficult to draw firm conclusions.

In theory, the dataset used to externally validate a prognostic model will be plausibly related to the development data. Therefore both datasets will effectively be samples taken from the same super-population. There are, however, no guidelines for choosing plausibly related datasets. We proposed that plausibly related datasets should conform to the Cochrane guidelines for combining studies in a meta-analysis [214]. Namely, that a group of studies is sufficiently homogeneous in terms of participants, interventions and outcomes to provide a meaningful summary. However, further simulation studies are required to determine if such guidelines are relevant and if extra conditions or amended conditions are required.

11.4 Future Work

Multivariable prognostic models for time to treatment failure and time to 12 month remission have been developed using the SANAD data [315]. In this thesis the models showed fairly good internal validity. External validation has not been undertaken due to the lack of a plausibly related dataset. The best match is a set of individual participant data collected by Tudur Smith *et al* [192]. This data is missing important covariates though and the treatments patients were randomised to do not always coincide with SANAD drugs.

It is important that the SANAD models developed in this thesis are externally validated so that the models may be implemented in general practice. One option would be to recruit a cohort of patients, with the same inclusion criteria as SANAD, and treated with carbamazepine, gabapentin, topirimate, oxcarbazepine and lamotrigine, as per SANAD. Baseline demographics, gender, treatment history, age, time from first seizure to randomisation, neurological insult, total number of seizures, EEG result, CT/MRI scan

result, seizure type, focal epilepsy site of onset and treatment, should then be obtained so that external validation methods can be employed. This however, is a very costly and time-consuming process.

Another option would be to use the individual patient dataset [192] to validate the SANAD models. Before this could happen however, methods for handling multiple missing covariates need to be found. Additionally, methods for handling non matching treatments need to be found which may involve clinical methods such as matching drugs by indication or chemical structure, or statistical methods such as mixed treatment comparisons [192]. Therefore, further research is required to develop methods for handling multiple missing covariates and also handling treatments that do not coincide with those used in the development dataset.

The methods of external validation investigated in the simulation study were demonstrated in a case study which validated a model developed using MESS via NGPSE, WA and FIRST. The MESS model was shown to generalise well to the NGPSE and WA datasets and fairly well to the FIRST dataset. Given that the model generalised fairly well to three alternative datasets it may be possible to use the combined information from the four studies to produce more precise conditional risk estimates in a similar fashion to a meta-analysis [214]. Given the heterogeneity between the datasets, a random-effects model is most likely to be appropriate if the meta-analysis style analysis is implemented [214].

This aim could be achieved by analysing all the data together in a Cox model stratified by dataset [167, 393]. Following on from modelling the combined data, more accurate, and potentially more representative, estimates of months from index seizure until annual seizure risk falls below 20% may be obtainable. These estimates could be used to better inform driving regulations both within the United Kingdom and European Union about length of time before the risk of a seizure in the next 12 months falls below the 20%

threshold level for specific patient subgroups. Further work is also needed to inform regulators as to whether a 20% threshold is optimum [202].

Bibliography

1. Information, U.M.; Available from: http://www.ukmi.nhs.uk/Research/common_abreviations.asp.
2. Hofer, S., et al., *Determinants of health-related quality of life in coronary artery disease patients: a prospective study generating a structural equation model.* Psychosomatics, 2005. **46**(3): p. 212-23.
3. Hofer, S., et al., *Determinants of health-related quality of life in patients with coronary artery disease.* Eur J Cardiovasc Prev Rehabil, 2006. **13**(3): p. 398-406.
4. Wang, P.P., E.M. Badley, and M. Gignac, *Exploring the role of contextual factors in disability models.* Disabil Rehabil, 2006. **28**(2): p. 135-40.
5. Breslow, A., *Prognostic factors in the treatment of cutaneous melanoma.* J Cutan Pathol, 1979. **6**(3): p. 208-12.
6. van der Steen, P.H., et al., *Prognostic factors in the treatment of alopecia areata with diphenylcyclopropenone.* J Am Acad Dermatol, 1991. **24**(2 Pt 1): p. 227-30.
7. *Prognostic factors in the treatment of mild hypertension. The Management Committee of the Australian National Blood Pressure Study.* Circulation, 1984. **69**(4): p. 668-76.
8. Steyerberg, E.W., *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. Statistics for Biology and Health, ed. M. Gail, et al. 2009, New York: Springer Science+Business Media, LLC.
9. Altman, D.G. and G.H. Lyman, *Methodological challenges in the evaluation of prognostic factors in breast cancer.* Breast Cancer Res Treat, 1998. **52**(1-3): p. 289-303.
10. Altman, D.G., *Systematic reviews of evaluations of prognostic variables.* BMJ, 2001. **323**(7306): p. 224-8.
11. Hayden, J.A., et al., *Identifying phases of investigation helps planning, appraising, and applying the results of explanatory prognosis studies.* J Clin Epidemiol, 2008. **61**(6): p. 552-60.
12. Windeler, J., *Prognosis - what does the clinician associate with this notion?* Stat Med, 2000. **19**(4): p. 425-30.
13. Hudak, P.L., D.C. Cole, and A.T. Haines, *Understanding prognosis to improve rehabilitation: the example of lateral elbow pain.* Arch Phys Med Rehabil, 1996. **77**(6): p. 586-93.
14. Simon, R. and D.G. Altman, *Statistical aspects of prognostic factor studies in oncology.* Br J Cancer, 1994. **69**(6): p. 979-85.
15. Lijmer, J.G., et al., *Empirical evidence of design-related bias in studies of diagnostic tests.* JAMA, 1999. **282**(11): p. 1061-6.
16. van der Schouw, Y.T., A.L. Verbeek, and S.H. Ruijs, *Guidelines for the assessment of new diagnostic tests.* Invest Radiol, 1995. **30**(6): p. 334-40.
17. Goldthorpe, J.H., *Causation, statistics, and sociology.* European Sociological Review, 2001. **17**(1): p. 1-20.
18. Lund, B., et al., *Comparison of the predictive power of different prognostic indices for overall survival in patients with advanced ovarian carcinoma.* Cancer Res, 1990. **50**(15): p. 4626-9.
19. Klein-Geltink, J.E., et al., *Readers should systematically assess methods used to identify, measure and analyze confounding in observational cohort studies.* J Clin Epidemiol, 2007. **60**(8): p. 766-72.

20. Goldman, L., *Cardiac risk in noncardiac surgery: an update.* Anesth Analg, 1995. **80**(4): p. 810-20.
21. Maly, R.C., B. Leake, and R.A. Silliman, *Health care disparities in older patients with breast carcinoma: informational support from physicians.* Cancer, 2003. **97**(6): p. 1517-27.
22. Ellis, F.H., Jr., W.A. Williamson, and G.J. Heatley, *Cancer of the esophagus and cardia: does age influence treatment selection and surgical outcomes?* J Am Coll Surg, 1998. **187**(4): p. 345-51.
23. Gospodarowicz, M., et al., *Prognostic factors in clinical decision making: the future.* Cancer, 2001. **91**(8 Suppl): p. 1688-95.
24. McShane, L.M., et al., *Reporting recommendations for tumor marker prognostic studies (REMARK).* J Natl Cancer Inst, 2005. **97**(16): p. 1180-4.
25. Ioannidis, J.P., et al., *A road map for efficient and reliable human genome epidemiology.* Nat Genet, 2006. **38**(1): p. 3-5.
26. Bogardus, S.T., Jr., J. Concato, and A.R. Feinstein, *Clinical epidemiological quality in molecular genetic research: the need for methodological standards.* JAMA, 1999. **281**(20): p. 1919-26.
27. Ioannidis, J.P., *Molecular evidence-based medicine: evolution and integration of information in the genomic era.* Eur J Clin Invest, 2007. **37**(5): p. 340-9.
28. Armitage, P., G. Berry, and J.N.S. Matthews, *Statistical Methods in Medical Research*. 4th ed 2002, Oxford: Blackwell.
29. Gompertz, B., *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.* Philosophical Transactions of the Royal Society of London, 1825. **115**: p. 513-583.
30. Kalbfleisch, J.D. and R.L. Prentice, *The Statistical Analysis of Failure Time Data* 1980, New York: Wiley.
31. Cox, D.R., *Regression Models and Life-Tables.* Journal of the Royal Statistical Society Series B-Statistical Methodology, 1972. **34**(2): p. 187-&.
32. Andersen, P.K., *Survival analysis 1982-1991: the second decade of the proportional hazards regression model.* Stat Med, 1991. **10**(12): p. 1931-41.
33. Faraggi, D. and R. Simon, *A neural network model for survival data.* Stat Med, 1995. **14**(1): p. 73-82.
34. Biganzoli, E., et al., *Feed forward neural networks for the analysis of censored survival data: a partial logistic regression approach.* Stat Med, 1998. **17**(10): p. 1169-86.
35. Ripley, B.D., *Neural networks as statstical methods in survival analysis*, in *Clinical Application of Artificial Neural Networks*, R. Dybowski and V. Gant, Editors. 2001, Cambridge University Press: New York. p. 237-255.
36. Biganzoli, E., P. Boracchi, and E. Marubini, *A general framework for neural network models on censored survival data.* Neural Netw, 2002. **15**(2): p. 209-18.
37. Ripley, R.M., A.L. Harris, and L. Tarassenko, *Non-linear survival analysis using neural networks.* Stat Med, 2004. **23**(5): p. 825-42.
38. Hollander, N. and W. Sauerbrei, *On statistical approaches for the multivariable analysis of prognostic factor studies*, in *Advances in Statistical Methods for the Health Sciences*, J.-L. Auget, et al., Editors. 2006, Birkenhauser: Boston.
39. Hornik, K., M. Stinchcombe, and H. White, *Multilayer Feedforward Networks Are Universal Approximators.* Neural Networks, 1989. **2**(5): p. 359-366.
40. Ripley, B.D., *Pattern Recognition and Neural Networks* 1996, Cambridge: Cambridge University Press.
41. Ripley, R.M., *Neural network models for breast cancer prognosis*, in *Engineering Science* 1998, Oxford: Oxford.

42. De Laurentiis, M. and P.M. Ravdin, *Survival analysis of censored data: neural network analysis detection of complex interactions between variables.* Breast Cancer Res Treat, 1994. **32**(1): p. 113-8.
43. Schumacher, M., et al., *Prognostic Factor Studies*, in *Handbook of Statistics in Clinical Oncology*, J. Crowley and D.P. Ankerst, Editors. 2006, Chapman & Hall/CRC Press: Boca Raton, FL. p. 289-333.
44. Burke, H.B., *Artificial neural networks for cancer research: outcome prediction.* Semin Surg Oncol, 1994. **10**(1): p. 73-9.
45. Gordon, L. and R.A. Olshen, *Tree-structured survival analysis.* Cancer Treat Rep, 1985. **69**(10): p. 1065-9.
46. LeBlanc, M. and J. Crowley, *Relative risk trees for censored survival data.* Biometrics, 1992. **48**(2): p. 411-25.
47. LeBlanc, M. and J. Crowley, *Survival Trees by Goodness of Split.* Journal of the American Statistical Association, 1993. **88**(422): p. 457-467.
48. Segal, M.R., *Regression Trees for Censored Data.* Biometrics, 1988. **44**(1): p. 35-47.
49. Segal, M.R., *Tree-structured survival analysis in medical research*, in *Statistical Analysis of Medical Data: New Developments*, B.S. Everitt and G. Dunn, Editors. 1998, Arnold: London. p. 101-125.
50. Breiman, L., *Bagging predictors.* Machine Learning, 1996. **24**(2): p. 123-140.
51. Friedman, J., T. Hastie, and R. Tibshirani, *Additive logistic regression: A statistical view of boosting.* Annals of Statistics, 2000. **28**(2): p. 337-374.
52. Breiman, L., *Random forests.* Machine Learning, 2001. **45**(1): p. 5-32.
53. Hothorn, T. and B. Lausen, *Bagging tree classifiers for laser scanning images: a data- and simulation-based strategy.* Artif Intell Med, 2003. **27**(1): p. 65-79.
54. Dietterich, T.G., *An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization.* Machine Learning, 2000. **40**(2): p. 139-157.
55. Kattan, M.W., *Comparison of Cox regression with other methods for determining prediction models and nomograms.* J Urol, 2003. **170**(6 Pt 2): p. S6-9; discussion S10.
56. Lisboa, P.J.G. and H. Wong. *Are neural networks best used to help logistic regression? An example from breast cancer survival analysis*. in *International Joint Conference on Neural Networks*. 2001. Washington, DC, USA.
57. Martinussen, T. and T.H. Scheike, *Dynamic Regression Models for Survival Data*, M. Gail, et al., Editors. 2006, Springer: New York.
58. Engel, J. and T.A. Pedley, *Epilepsy: a comprehensive textbook*. 2nd ed, ed. J. Aicardi. Vol. 1. 2008, Philadelphia: Lippincott Williams & Wilkins, A Wolters Kluwer Business.
59. Murray, C.J.L., A.D. Lopez, and D.T. Jamison, *The Global Burden of Disease in 1990 - Summary Results, Sensitivity Analysis and Future-Directions.* Bulletin of the World Health Organization, 1994. **72**(3): p. 495-509.
60. Deacon, K. and S. Wigglesworth, *Epilepsy Prevalence, Incidence and Other Statistics*, in http://www.jointepilepsycouncil.org.uk/downloads/Epilepsy%20Prevalence,%20Incidence%20and%20Other%20Statistics.pdf, J.E. Council, Editor 2004.
61. *Guidelines for Epidemiologic Studies on Epilepsy.* Epilepsia, 1993. **34**(4): p. 592-596.
62. Fisher, R.S., et al., *Epileptic seizures and epilepsy: Definitions proposed by the International League against Epilepsy (ILAE) and the International Bureau for Epilepsy (IBE).* Epilepsia, 2005. **46**(4): p. 470-472.
63. Sander, J.W.A.S. and S.D. Shorvon, *Epidemiology of the epilepsies.* Journal of Neurology Neurosurgery and Psychiatry, 1996. **61**(5): p. 433-443.

64. Sander, J.W., *The use of antiepileptic drugs - Principles and practice.* Epilepsia, 2003. **44**: p. 200-200.
65. Hauser, W.A., J.F. Annegers, and L.T. Kurland, *Incidence of Epilepsy and Unprovoked Seizures in Rochester, Minnesota - 1935-1984.* Epilepsia, 1993. **34**(3): p. 453-468.
66. Hesdorffer, D.C., et al., *Estimating risk for developing epilepsy.* Neurology. **76**(1): p. 23-27.
67. Action, E., *Epilepsy and Everyone*, 2008, Epilepsy Action.
68. Hauser, W.A. and J.F. Annegers, *Epidemiology of epilepsy*, in *Text Book of Epilepsy*, J.P. Laidlaw, A. Richens, and D. Chadwick, Editors. 1992, Churchill Livingstone: New York. p. 23-45.
69. Cockerell, O.C., et al., *Epilepsy in a population of 6000 re-examined: secular trends in first attendance rates, prevalence, and prognosis.* J Neurol Neurosurg Psychiatry, 1995. **58**(5): p. 570-6.
70. Cockerell, O.C., J.W.A.S. Sander, and S.D. Shorvon, *Remission of Epilepsy.* Lancet, 1995. **346**(8984): p. 1228-1228.
71. Moran, N.F., et al., *Epilepsy in the United Kingdom: seizure frequency and severity, anti-epileptic drug utilization and impact on life in 1652 people with epilepsy.* Seizure-European Journal of Epilepsy, 2004. **13**(6): p. 425-433.
72. Sander, J.W., *The use of antiepileptic drugs - Principles and practice.* Epilepsia, 2004. **45**: p. 28-34.
73. Action, E., *Epileptic Seizures Explained*, 2008, Epilepsy Action.
74. Engel, J., *A Proposed Diagnostic Scheme for People with Epileptic Seizures and with Epilepsy: Report of the ILAE Task Force on Classification and Terminology.* Epilepsia, 2001. **42**(6): p. 796-803.
75. Bancaud, J., et al., *Proposal for Revised Clinical and Electroencephalographic Classification of Epileptic Seizures.* Epilepsia, 1981. **22**(4): p. 489-501.
76. Berg, A.T., et al., *Revised terminology and concepts for organization of seizures and epilepsies: Report of the ILAE Commission on Classification and Terminology, 2005-2009.* Epilepsia, 2010. **51**(4): p. 676-685.
77. Browne, T.R. and G.L. Holmes, *Handbook of epilepsy* 2008, Philadelphia: Lippincott Williams & Wilkins.
78. Weishaupt, D., V.D. Kochli, and B. Marincek, *How does MRI work? : an introduction to the physics and function of magnetic resonance imaging* 2003, Berlin Heidelberg New York: Springer-Verlag.
79. Action, E., *Epilepsy: diagnosis, treatment and healthcare*, 2008, Epilepsy Action.
80. GlaxoSmithKlein. 2010; Available from: http://www.medicines.org.uk/EMC/medicine/4228/SPC/Lamictal/.
81. Group, U.P.P. 2001; Available from: http://www.candi.nhs.uk/_uploads/documents/medicines/medicines-VALPROATE.PDF.
82. BNF. *British National Formulary*. 2012; Available from: http://www.bnf.org/bnf/index.htm.
83. Polkey, C.E., *Alternative surgical procedures to help drug-resistant epilepsy - a review.* Epileptic Disord, 2003. **5**(2): p. 63-75.
84. Freeman, J.M., E.H. Kossoff, and A.L. Hartman, *The Ketogenic Diet: One Decade Later.* Pediatrics, 2007. **119**(3): p. 535-543.
85. Kossoff, E.H., B.A. Zupec-Kania, and J.M. Rho, *Ketogenic Diets: An Update for Child Neurologists.* Journal of Child Neurology, 2009. **24**(8): p. 979-988.
86. Sirven, J., et al., *The Ketogenic Diet for Intractable Epilepsy in Adults: Preliminary Results.* Epilepsia, 1999. **40**(12): p. 1721-1726.

87. Ramos-Lizana, J., et al., *Recurrence risk after withdrawal of antiepileptic drugs in children with epilepsy: a prospective study.* Eur J Paediatr Neurol. **14**(2): p. 116-24.
88. Peters, A.C., et al., *Randomized prospective study of early discontinuation of antiepileptic drugs in children with epilepsy.* Neurology, 1998. **50**(3): p. 724-30.
89. Braathen, G. and H. Melander, *Early discontinuation of treatment in children with uncomplicated epilepsy: a prospective study with a model for prediction of outcome.* Epilepsia, 1997. **38**(5): p. 561-9.
90. Mastropaolo, C., et al., *Prognosis after Therapy Discontinuation in Children with Epilepsy.* European Neurology, 1992. **32**(3): p. 141-145.
91. Matricardi, M., M. Brinciotti, and P. Benedetti, *Outcome after Discontinuation of Antiepileptic Drug-Therapy in Children with Epilepsy.* Epilepsia, 1989. **30**(5): p. 582-589.
92. Bouma, P.A., et al., *Discontinuation of antiepileptic therapy: a prospective study in children.* J Neurol Neurosurg Psychiatry, 1987. **50**(12): p. 1579-83.
93. Todt, H., *The late prognosis of epilepsy in childhood: results of a prospective follow-up study.* Epilepsia, 1984. **25**(2): p. 137-44.
94. Arts, W.F.M., et al., *Follow-up of 146 Children with Epilepsy after Withdrawal of Antiepileptic Therapy.* Epilepsia, 1988. **29**(3): p. 244-250.
95. Sillanpaa, M. and D. Schmidt, *Delayed time to first remission identifies poor long-term drug response of childhood-onset epilepsy: a prospective population-based study.* Epilepsy Behav, 2009. **16**(3): p. 507-11.
96. Wang, Z., L. Qi, and X. Song, *Prognosis and predictive factors of partial seizures in children.* Pediatr Neurol, 2007. **37**(1): p. 16-20.
97. Oskoui, M., et al., *Factors predictive of outcome in childhood epilepsy.* J Child Neurol, 2005. **20**(11): p. 898-904.
98. Grosso, S., et al., *Childhood absence epilepsy: evolution and prognostic factors.* Epilepsia, 2005. **46**(11): p. 1796-801.
99. Trinka, E., et al., *Long-term prognosis for childhood and juvenile absence epilepsy.* J Neurol, 2004. **251**(10): p. 1235-41.
100. Ramos-Lizana, J., et al., *[Long term remission of epilepsy in childhood: a prospective study].* Rev Neurol, 2002. **34**(9): p. 824-9.
101. Arts, W.F., et al., *The early prognosis of epilepsy in childhood: the prediction of a poor outcome. The Dutch study of epilepsy in childhood.* Epilepsia, 1999. **40**(6): p. 726-34.
102. Brorson, L.O. and L. Wranne, *Long-term prognosis in childhood epilepsy: survival and seizure prognosis.* Epilepsia, 1987. **28**(4): p. 324-30.
103. Banachowska, F., Z. Majewska, and M. Zuchowicz, *[Various prognostic elements in childhood epilepsy in the light of selected materials of the Department of Developmental Neurology, Medical Academy in Gdansk].* Neurol Neurochir Pol, 1978. **12**(1): p. 9-14.
104. Hauser, E., et al., *Prognosis of childhood epilepsy in newly referred patients.* J Child Neurol, 1996. **11**(3): p. 201-4.
105. Onsurbe, I., J. Marco, and A. Sanchez, *[First spontaneous epileptic crisis in childhood: risk of relapse and prognostic factors].* Rev Neurol, 1999. **28**(8): p. 761-5.
106. Winckler, M.I. and N.T. Rotta, *Prognostic factors for recurrence of a first seizure during childhood.* Arq Neuropsiquiatr, 1997. **55**(4): p. 749-56.
107. Berg, A.T., et al., *Remission of epilepsy after two drug failures in children: a prospective study.* Ann Neurol, 2009. **65**(5): p. 510-9.
108. Ohtsuka, Y., et al., *Predictors and underlying causes of medically intractable localization-related epilepsy in childhood.* Pediatr Neurol, 2001. **24**(3): p. 209-13.

109. Ohtsuka, Y., H. Yoshinaga, and K. Kobayashi, *Refractory childhood epilepsy and factors related to refractoriness.* Epilepsia, 2000. **41 Suppl 9**: p. 14-7.
110. Wolf, P. and Y. Inoue, *Therapeutic response of absence seizures in patients of an epilepsy clinic for adolescents and adults.* J Neurol, 1984. **231**(4): p. 225-9.
111. Sato, S., et al., *Long-term follow-up of absence seizures.* Neurology, 1983. **33**(12): p. 1590-5.
112. Sato, S., F.E. Dreifuss, and J.K. Penry, *Prognostic factors in absence seizures.* Neurology, 1976. **26**(8): p. 788-96.
113. Bodenstein-Sachar, H., et al., *Outcome of lamotrigine treatment in juvenile myoclonic epilepsy.* Acta Neurol Scand. **124**(1): p. 22-7.
114. Motamedi, M., S. Mohammad Ali, and M. Rahmat, *Comparison of valporic acid efficacy in familial versus sporadic cases of juvenile myoclonic epilepsy.* Neurol India, 2006. **54**(2): p. 186-9; discussion 189.
115. Pittau, F., et al., *Prognostic factors in patients with mesial temporal lobe epilepsy.* Epilepsia, 2009. **50 Suppl 1**: p. 41-4.
116. Willert, C., U. Runge, and C. Kessler, *[Prognostic factors and clinical outcome in symptomatic focal epilepsies].* Fortschr Neurol Psychiatr, 1996. **64**(12): p. 509-14.
117. Verrotti, A., et al., *Factors associated with poor control in partial complex epilepsy.* J Child Neurol, 2004. **19**(4): p. 262-4.
118. Schiller, Y. and Y. Najjar, *Quantifying the response to antiepileptic drugs: effect of past treatment history.* Neurology, 2008. **70**(1): p. 54-65.
119. Kim, L.G., et al., *Prediction of Risk of Seizure Recurrence after a Single Seizure and Early Epilepsy: Further Results from the MESS Trial.* The Lancet Neurology, 2006. **5**(4): p. 317-322.
120. MacDonald, B.K., et al., *Factors predicting prognosis of epilepsy after presentation with seizures.* Annals of Neurology, 2000. **48**(6): p. 833-841.
121. Specchio, L.M., et al., *Discontinuing antiepileptic drugs in patients who are seizure free on monotherapy.* J Neurol Neurosurg Psychiatry, 2002. **72**(1): p. 22-5.
122. Chadwick, D., J. Taylor, and T. Johnson, *Outcomes after seizure recurrence in people with well-controlled epilepsy and the factors that influence it. The MRC Antiepileptic Drug Withdrawal Group.* Epilepsia, 1996. **37**(11): p. 1043-50.
123. *Prognostic index for recurrence of seizures after remission of epilepsy. Medical Research Council Antiepileptic Drug Withdrawal Study Group.* BMJ, 1993. **306**(6889): p. 1374-8.
124. Sillanpaa, M., *Remission of seizures and predictors of intractability in long-term follow-up.* Epilepsia, 1993. **34**(5): p. 930-6.
125. Wong, I.C., et al., *A pharmacoepidemiologic study of factors influencing the outcome of treatment with lamotrigine in chronic epilepsy.* Epilepsia, 2001. **42**(10): p. 1354-8.
126. Lhatoo, S.D., I.C. Wong, and J.W. Sander, *Prognostic factors affecting long-term retention of topiramate in patients with chronic epilepsy.* Epilepsia, 2000. **41**(3): p. 338-41.
127. Turnbull, D.M., et al., *Which drug for the adult epileptic patient: phenytoin or valproate?* Br Med J (Clin Res Ed), 1985. **290**(6471): p. 815-9.
128. Kaliaperumal, V.G., N. Sundararaj, and K.S. Mani, *Seizure prognosis for partial epilepsies in India.* Epilepsy Res, 1989. **3**(1): p. 86-91.
129. Porro, G., et al., *Prognosis of partial epilepsy.* Arch Dis Child, 1988. **63**(10): p. 1192-7.
130. Wu, L.W., et al., *Prognostic factors of primary generalized epilepsy: a reappraisal of 96 cases in terminal remission.* Folia Psychiatr Neurol Jpn, 1985. **39**(2): p. 139-45.

131. Cockerell, O.C., et al., *Remission of Epilepsy - Results from the National General-Practice Study of Epilepsy.* Lancet, 1995. **346**(8968): p. 140-144.
132. *Randomised study of antiepileptic drug withdrawal in patients in remission. Medical Research Council Antiepileptic Drug Withdrawal Study Group.* Lancet, 1991. **337**(8751): p. 1175-80.
133. Marson, A., et al., *Immediate versus deferred antiepileptic drug treatment for early epilepsy and single seizures: a randomised controlled trial.* Lancet, 2005. **365**(9476): p. 2007-13.
134. Schumacher, M., N. Hollander, and W. Sauerbrei. *Reduction of bias caused by model building*. in *Statistical Computing Section*. 1996. American Statistical Association.
135. Pierre J. M. Verweij, H.C.V.H., *Cross-validation in survival analysis.* Statistics in Medicine, 1993. **12**(24): p. 2305-2314.
136. Schumacher, M., N. Hollander, and W. Sauerbrei, *Resampling and cross-validation techniques: a tool to reduce bias caused by model building?* Stat Med, 1997. **16**(24): p. 2813-27.
137. Van Houwelingen, J.C. and S. Le Cessie, *Predictive value of statistical models.* Stat Med, 1990. **9**(11): p. 1303-25.
138. Sauerbrei, W., P. Royston, and H. Binder, *Selection of important variables and determination of functional form for continuous predictors in multivariable model building.* Stat Med, 2007. **26**(30): p. 5512-28.
139. Harrell Jr., F.E., *Regression Modeling Stategies: with applications to linear models, logistic regression, and survival analysis*. Springer Series in Statistics 2001, New York: Springer.
140. Wood, S.N., *Generalised Additive Models: An introduction with R* 2006, Boca Raton, FL: Chapman and Hall/CRC.
141. Stone, C.J., *Additive Regression and Other Nonparametric Models.* Annals of Statistics, 1985. **13**(2): p. 689-705.
142. Devlin, T.F. and B.J. Weeks. *Spline functions for logistic regression modeling*. in *Eleventh Annual SAS Users Group International Conference*. 1986. SAS Institute, Inc.
143. Stone, C.J., *Comment: Generalized Additive Models.* Statistical Science, 1986. **1**: p. 3.
144. Royston, P. and D.G. Altman, *Regression Using Fractional Polynomials of Continuous Covariates - Parsimonious Parametric Modeling.* Applied Statistics-Journal of the Royal Statistical Society Series C, 1994. **43**(3): p. 429-467.
145. Royston, P. and W. Sauerbrei, *Multivariable Model-Building - A pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables* 2008: Wiley.
146. Hollander, N. and M. Schumacher, *Estimating the functional form of a continuous covariate's effect on survival time.* Computational Statistics & Data Analysis, 2006. **50**(4): p. 1131-1151.
147. Sauerbrei, W. and P. Royston, *Building multivariable prognostic and diagnostic models: transformation of the predictors by using fractional polynomials.* Journal of the Royal Statistical Society Series a-Statistics in Society, 1999. **162**: p. 71-94.
148. Royston, P., W. Sauerbrei, and H. Becher, *Modelling continuous exposures with a 'spike' at zero: A new procedure based on fractional polynomials.* Statistics in Medicine, 2010. **29**(11): p. 1219-1227.
149. Lambert, D., *Zero-Inflated Poisson Regression, with an Application to Defects in Manufacturing.* Technometrics, 1992. **34**(1): p. 1-14.

150. Robertson, C., et al., *Some Statistical Considerations in the Analysis of Case-Control Studies When the Exposure Variables Are Continuous Measurements.* Epidemiology, 1994. **5**(2): p. 164-170.
151. Pregibon, D., *Data analytic methods for matched case-control studies.* Biometrics, 1984. **40**(3): p. 639-51.
152. Cronbach, L.J., *Statistical Tests for Moderator Variables - Flaws in Analyses Recently Proposed.* Psychological Bulletin, 1987. **102**(3): p. 414-417.
153. Aiken, L.S. and S.G. West, *Multiple Regression: Testing and Interpreting Interactions* 1991, Newbury Park CA: Sage Publications.
154. Cohen, J., et al., *Applied Multiple Regression/correlation Analysis for the Behavioral Sciences* 2003, Hillsdale NJ: Lawrence Erlbaum Associates.
155. Glantz, S.A. and B.K. Slinker, *Primer of Applied Regression and Analysis of Variance* 2001, New York: McGraw-Hill.
156. Kromrey, J.D. and L. Foster-Johnson, *Mean centering in moderated multiple regression: Much ado about nothing.* Educational and Psychological Measurement, 1998. **58**(1): p. 42-67.
157. Marquardt, D.W., *A Critique of Some Ridge-Regression Methods - Comment.* Journal of the American Statistical Association, 1980. **75**(369): p. 87-91.
158. Kraemer, H.C. and C.M. Blasey, *Centring in regression analyses: a strategy to prevent errors in statistical inference.* International Journal of Methods in Psychiatric Research, 2004. **13**(3): p. 141-151.
159. Claeskens, G. and N.L. Hjort, *Model Selection and Model Averaging*. Cambridge Series in Statistical and Probabilistic Mathematics, ed. R. Gill, et al. 2010, Cambridge: Cambridge University Press.
160. Mantel, N., *Why Stepdown Procedures in Variable Selection.* Technometrics, 1970. **12**(3): p. 621-&.
161. Willi Sauerbrei, *The Use of Resampling Methods to Simplify Regression Models in Medical Statistics.* Journal of the Royal Statistical Society: Series C (Applied Statistics), 1999. **48**(3): p. 313-329.
162. Sauerbrei, W., *Comparison of Variable Selection Procedures in Regression Models: a simulation study and practical examples*, in *Europaische Perspektiven der Medizinishen Informatik, Biometri und Epidemiologie*, J. Michaelis, G. Hommel, and S. Wellek, Editors. 1993, MMV: Munchen. p. 108-113.
163. Akaike, H. *Information Theory as an Extension of the Maximum Likelihood Principle*. in *Second International Symposium on Information Theory*. 1973. Akademiai Kiado, Budapest.
164. Schwarz, G., *Estimating Dimension of a Model.* Annals of Statistics, 1978. **6**(2): p. 461-464.
165. Everitt, B.S., *The Cambridge Dictionary of Statistics* 1998, Cambridge: Cambridge University Press.
166. Volinsky, C.T. and A.E. Raftery, *Bayesian information criterion for censored survival models.* Biometrics, 2000. **56**(1): p. 256-262.
167. Collett, D., *Modelling Survival Data in Medical Research* 2003, Boca Raton, Fla.: Chapman & Hall/CRC.
168. Royston, P. and W. Sauerbrei, *A new approach to modelling interactions between treatment and continuous covariates in clinical trials by using fractional polynomials.* Stat Med, 2004. **23**(16): p. 2509-25.
169. Kleinbaum, D.G. and M. Klein, *Competing Risks Survival Analysis*, in *Survival Analysis: A self-learning text* 2005, Springer: New York.
170. Grambsch, P.M. and T.M. Therneau, *Proportional Hazards Tests and Diagnostics Based on Weighted Residuals.* Biometrika, 1994. **81**(3): p. 515-526.

171. Cox, D.R. and E.J. Snell, *A General Definition of Residuals.* Journal of the Royal Statistical Society. Series B (Methodological), 1968. **30**(2): p. 248-275.
172. Therneau, T.M., P.M. Grambsch, and T.R. Fleming, *Martingale-Based Residuals for Survival Models.* Biometrika, 1990. **77**(1): p. 147-160.
173. Schoenfeld, D., *Partial Residuals for the Proportional Hazards Regression-Model.* Biometrika, 1982. **69**(1): p. 239-241.
174. Gray, R.J., *A Class of K-Sample Tests for Comparing the Cumulative Incidence of a Competing Risk.* Annals of Statistics, 1988. **16**(3): p. 1141-1154.
175. Therneau, T.M. and P.M. Grambsch, *Modeling Survival Data: Extending the Cox Model* 2000, New York: Springer-Verlag.
176. Fine, J.P. and R.J. Gray, *A proportional hazards model for the subdistribution of a competing risk.* Journal of the American Statistical Association, 1999. **94**(446): p. 496-509.
177. Pintilie, M., *Competing Risks: A Practical Perspective* 2006, Chichester: John Wiley & Sons Ltd.
178. Akaike, H., *New Look at Statistical-Model Identification.* Ieee Transactions on Automatic Control, 1974. **Ac19**(6): p. 716-723.
179. Engel, J.J., *ILAE classification of epilepsy syndromes.* Epilepsy Research, 2006. **70**(Supplement 1): p. 5-10.
180. Marson, A.G., et al., *The SANAD study of effectiveness of carbamazepine, gabapentin, lamotrigine, oxcarbazepine, or topiramate for treatment of partial epilepsy: an unblinded randomised controlled trial.* Lancet, 2007. **369**(9566): p. 1000-15.
181. Marson, A.G., et al., *The SANAD study of effectiveness of valproate, lamotrigine, or topiramate for generalised and unclassifiable epilepsy: an unblinded randomised controlled trial.* Lancet, 2007. **369**(9566): p. 1016-26.
182. Team, R.C.D., *R Foundation for Statistical Computing*, 2009: Vienna, Austria.
183. Institute, S.A.S., *SAS/IML software: changes and enhancements, through release 6.11*1995: Sas Inst.
184. Harrell, F.E., K.L. Lee, and D.B. Mark, *Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors.* Statistics in Medicine, 1996. **15**(4): p. 361-387.
185. Royston, P., G. Ambler, and W. Sauerbrei, *The use of fractional polynomials to model continuous risk variables in epidemiology.* Int J Epidemiol, 1999. **28**(5): p. 964-74.
186. Royston, P. and W. Sauerbrei, *Building multivariable regression models with continuous covariates in clinical epidemiology--with an emphasis on fractional polynomials.* Methods Inf Med, 2005. **44**(4): p. 561-71.
187. Royston, P. and W. Sauerbrei, *Improving the robustness of fractional polynomial models by preliminary covariate transformation: A pragmatic approach.* Computational Statistics & Data Analysis, 2007. **51**(9): p. 4240-4253.
188. Panayiotopoulos, C.P., *Old versus new antiepileptic drugs: the SANAD study.* Lancet, 2007. **370**(9584): p. 313-4; author reply 315-6.
189. Panayiotopoulos, C.P., *Evidence-based epileptology, randomized controlled trials, and SANAD: a critical clinical view.* Epilepsia, 2007. **48**(7): p. 1268-74.
190. Heller, A.J., et al., *Phenobarbitone, Phenytoin, Carbamazepine, or Sodium Valproate for Newly-Diagnosed Adult Epilepsy - a Randomized Comparative Monotherapy Trial.* Journal of Neurology Neurosurgery and Psychiatry, 1995. **58**(1): p. 44-50.
191. deSilva, M., et al., *Randomised comparative monotherapy trial of phenobarbitone, phenytoin, carbamazepine, or sodium valproate for newly diagnosed childhood epilepsy.* Lancet, 1996. **347**(9003): p. 709-713.

192. Tudur Smith, C., et al., *Multiple treatment comparisons in epilepsy monotherapy trials.* Trials, 2007. **8**: p. 34.
193. Bonnett, L.J., et al., *Risk of recurrence after a first seizure and implications for driving: further analysis of the Multicentre study of early Epilepsy and Single Seizures.* BMJ, 2010. **341**: p. c6477.
194. Bonnett, L.J., et al., *Seizure recurrence after antiepileptic drug withdrawal and the implications for driving: further results from the MRC Antiepileptic Drug Withdrawal Study and a systematic review.* Journal of Neurology Neurosurgery and Psychiatry, 2011. **82**(12): p. 1328-1333.
195. Hart, Y.M., J.W.A.S. Sander, and S.D. Shorvon, *National General-Practice Study of Epilepsy and Epileptic Seizures - Objectives and Study Methodology of the Largest Reported Prospective Cohort Study of Epilepsy - National General-Practice Study of Epilepsy and Epileptic Seizures (Ngpse).* Neuroepidemiology, 1989. **8**(5): p. 221-227.
196. Fisher, R.S., et al., *Epilepsy and Driving - An International Perspective.* Epilepsia, 1994. **35**(3): p. 675-684.
197. Ooi, W.W. and J.A. Gutrecht, *International regulations for automobile driving and epilepsy.* J Travel Med, 2000. **7**(1): p. 1-4.
198. Carter, T., *Fitness to Drive: a Guide for Health Professionals* 2006: Royal Society of Medicine Press.
199. *Second European Working Group on Epilepsy and Driving - Epilepsy and Driving in Europe*, 2005.
200. *European Commission for Transport*. Available from: http://ec.europa.eu/transport/index_en.htm.
201. Schmedding, E., *Epilepsy and Driving in Belgium: Proposals and Justification.* Acta Neurologica Belgica, 2004. **104**(2): p. 68-79.
202. DVLA, *At a Glance Guide to the Current Medical Standards of Fitness to Drive*, 2010, Drivers Medical Group: Swansea.
203. Krauss, G.L., L. Ampaw, and A. Krumholz, *Individual state driving restrictions for people with epilepsy in the US.* Neurology, 2001. **57**(10): p. 1780-1785.
204. *Prognosis of Epilepsy in Newly Referred Patients: A Multicenter Prospective Study of the Effects of Monotherapy on the Long-Term Course of Epilepsy. Collaborative Group for the Study of Epilepsy.* Epilepsia, 1992. **33**(1): p. 45-51.
205. Hospital, M.G. *Growing Up With Epilepsy*. 2006; Available from: http://www2.massgeneral.org/childhoodepilepsy/medical/index.htm.
206. Burneo, J.G., et al., *Video-EEG study in an adult and a child with eyelid myoclonia with absences.* Epileptic Disord, 2004. **6**(4): p. 287-91.
207. Davis, F.G., et al., *The conditional probability of survival of patients with primary malignant brain tumors - Surveillance, epidemiology, and end results (SEER) data.* Cancer, 1999. **85**(2): p. 485-491.
208. Cox, D.R. and D. Oakes, *Analysis of Survival Data*. Monographs on Statistics and Applied Probability, ed. D.R. Cox and D.V. Hinkley 1984, London: Chapman and Hall Ltg.
209. Hart, Y.M., et al., *National General-Practice Study of Epilepsy - Recurrence After a First Seizure.* Lancet, 1990. **336**(8726): p. 1271-1274.
210. *Road Traffic Act*, 1988.
211. Lin, C.L., et al., *The conditional probabilities of survival in patients with anaplastic astrocytoma or glioblastoma multiforme.* Surgical Neurology, 2003. **60**(5): p. 402-406.
212. Wells, G.A., et al. *The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses*. [cited 2010; Available from: http://www.ohri.ca/programs/clinical_epidemiology/oxford.htm.

213. Juni, P., D.G. Altman, and M. Egger, *Systematic reviews in health care: Assessing the quality of controlled clinical trials.* BMJ, 2001. **323**(7303): p. 42-6.
214. Collaboration, C. *Cochrane Handbook for Systematic Reviews of Interventions*. 2001; Available from: http://www.cochrane-handbook.org/.
215. Lossius, M.I., et al., *Consequences of antiepileptic drug withdrawal: A randomized, double-blind study (Akershus Study).* Epilepsia, 2008. **49**(3): p. 455-463.
216. Aktekin, B., et al., *Withdrawal of antiepileptic drugs in adult patients free of seizures for 4 years: A prospective study.* Epilepsy & Behavior, 2006. **8**(3): p. 616-619.
217. Callaghan, N., A. Garrett, and T. Goggin, *Withdrawal of anticonvulsant drugs in patients free of seizures for two years. A prospective study.* N Engl J Med, 1988. **318**(15): p. 942-6.
218. Overweg, J., et al., *Clinical and EEG prediction of seizure recurrence following antiepileptic drug withdrawal.* Epilepsy Res, 1987. **1**(5): p. 272-83.
219. Ohta, H., et al., *Prognosis after withdrawal of antiepileptic drugs in childhood-onset cryptogenic localization-related epilepsies.* Brain Dev, 2004. **26**(1): p. 19-25.
220. Shinnar, S., et al., *Discontinuing antiepileptic drugs in children with epilepsy: a prospective study.* Ann Neurol, 1994. **35**(5): p. 534-45.
221. Aldenkamp, A.P., et al., *Withdrawal of antiepileptic medication in children--effects on cognitive function: The Multicenter Holmfrid Study.* Neurology, 1993. **43**(1): p. 41-50.
222. Ehrhardt, P. and W.I. Forsythe, *Prognosis after Grand Mal Seizures - a Study of 187 Children with 3-Year Remissions.* Developmental Medicine and Child Neurology, 1989. **31**(5): p. 633-635.
223. Callaghan, N., et al., *A Prospective-Study between Carbamazepine, Phenytoin and Sodium Valproate as Monotherapy in Previously Untreated and Recently Diagnosed Patients with Epilepsy.* Journal of Neurology Neurosurgery and Psychiatry, 1985. **48**(7): p. 639-644.
224. Parmar, M.K., V. Torri, and L. Stewart, *Extracting summary statistics to perform meta-analyses of the published literature for survival endpoints.* Stat Med, 1998. **17**(24): p. 2815-34.
225. Chadwick, D., *Does withdrawal of different antiepileptic drugs have different effects on seizure recurrence? Further results from the MRC Antiepileptic Drug Withdrawal Study.* Brain, 1999. **122 (Pt 3)**: p. 441-8.
226. Nagelkerke, N.J.D., *A Coefficient of Determination for Cox's Proportional Hazards Model and Conditional Logistic Regression*, in *28th Annual Conference of the International Society for Clinical Biostatistics* 2007: Alexandroupolis, Greece.
227. Schemper, M. and J. Stare, *Explained variation in survival analysis.* Stat Med, 1996. **15**(19): p. 1999-2012.
228. Stare, J., *Measures of Explained Variation in Survival Analysis*, 1994, University of Ljubljana: Slovenia.
229. Kvalseth, T.O., *Cautionary Note About R2.* American Statistician, 1985. **39**(4): p. 279-285.
230. Kendall, M.G. and J.D. Gibbons, *Rank Correlation Methods*. Fifth ed 1990, London: Arnold.
231. Altman, D.G., *Practical statistics for medical research* 1991, London: Chapman & Hall.
232. Colwell, D.J. and J.R. Gillett, *66.49 Spearman versus Kendall.* The Mathematical Gazette, 1982. **66**(438): p. 307-309.
233. Somers, R.H., *A New Asymmetric Measure of Association for Ordinal Variables.* American Sociological Review, 1962. **27**(6): p. 799-811.

234. LeRoy, M.K., *Research Methods in Political Science*. 7th ed 2008, Boston: Michael Rosenberg.
235. Goktas, A. and I. Oznur, *A Comparison of the Most Commonly Used Measures of Association for Doubly Ordered Square Contingency Tables via Simulation.* Metodolski zvezki, 2011. **8**(1): p. 17-37.
236. Korn, E.L. and R. Simon, *Explained Residual Variation, Explained Risk, and Goodness of Fit.* American Statistician, 1991. **45**(3): p. 201-206.
237. Brown, B.W.J., M. Hollander, and R.M. Korwar, *Nonparametric tests of independence for censored data, with applications to heart transplant studies*, in *Reliability and Biometry*, F. Proshan and R.J. Serfling, Editors. 1974, SIAM: Philadelphia. p. 327-354.
238. Schemper, M., *Predictive accuracy and explained variation in Cox regression.* Biometrics, 2000. **56**(1): p. 249-255.
239. Schemper, M., *The Explained Variation in Proportional Hazards Regression.* Biometrika, 1990. **77**(1): p. 216-218.
240. Schemper, M. and A. Kaider, *A new approach to estimate correlation coefficients in the presence of censoring and proportional hazards.* Computational Statistics & Data Analysis, 1997. **23**(4): p. 467-476.
241. Graf, E. and M. Schumacher, *An Investigation on Measures of Explained Variation in Survival Analysis.* The Statistician, 1995. **44**(4): p. 497-507.
242. Kent, J.T. and J. Oquigley, *Measures of Dependence for Censored Survival-Data.* Biometrika, 1988. **75**(3): p. 525-534.
243. O'Quigley, J., *Explained Variation*, in *Proportional Hazards Regression* 2008, Springer. p. 366-370.
244. Kent, J.T., *Information Gain and a General Measure of Correlation.* Biometrika, 1983. **70**(1): p. 163-173.
245. Maddala, G.S., *Limited-dependent and Qualitative Variables in Econometrics* 1983, Cambridge: Cambridge University Press.
246. Harrell, F.E., *The PHGLM procedure*, in *SUGI Supplemental Library User's Guide* 1985, SAS Institute Inc.: Cary, NC. p. 437-448.
247. Harrell, F.E., et al., *Evaluating the Yield of Medical Tests.* Jama-Journal of the American Medical Association, 1982. **247**(18): p. 2543-2546.
248. D'Agostino, R.B. and B.H. Nam, *Evaluation of the Performance of Survival Analysis Models: Discrimination and Calibration Measures*, in *Advances in Survival Analysis*, N. Balakrishnan and C.R. Rao, Editors. 2004, Elsevier: Amsterdam.
249. Hanley, J.A. and B.J. Mcneil, *The Meaning and Use of the Area under a Receiver Operating Characteristic (Roc) Curve.* Radiology, 1982. **143**(1): p. 29-36.
250. Lemeshow, S. and Jean-Roger Le, *Modeling the Severity of Illness of ICU Patients: A Systems Update.* JAMA, 1994. **272**(13): p. 1049-1055.
251. Altman, D.G., et al., *Prognosis and prognostic research: validating a prognostic model.* BMJ, 2009. **338**: p. b605.
252. Moons, K.G., et al., *Prognosis and prognostic research: application and impact of prognostic models in clinical practice.* BMJ, 2009. **338**: p. b606.
253. Hosmer, D.W. and S. Lemeshow, *Goodness of Fit Tests for the Multiple Logistic Regression-Model.* Communications in Statistics Part a-Theory and Methods, 1980. **9**(10): p. 1043-1069.
254. Harrell, F.E., et al., *Regression Modeling Strategies for Improved Prognostic Prediction.* Statistics in Medicine, 1984. **3**(2): p. 143-152.
255. Molinaro, A.M., R. Simon, and R.M. Pfeiffer, *Prediction error estimation: a comparison of resampling methods.* Bioinformatics, 2005. **21**(15): p. 3301-3307.

256. Schumacher, M., H. Binder, and T. Gerds, *Assessment of survival prediction models based on microarray data.* Bioinformatics, 2007. **23**(14): p. 1768-1774.
257. Steyerberg, E.W., et al., *Internal validation of predictive models: Efficiency of some procedures for logistic regression analysis.* Journal of Clinical Epidemiology, 2001. **54**(8): p. 774-781.
258. Efron, B. and R.J. Tibishirani, *An Introduction to the Bootstrap* 1993: CRC Press.
259. Steyerberg, E.W., et al., *Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small data sets.* Statistics in Medicine, 2000. **19**(8): p. 1059-1079.
260. Steyerberg, E.W., et al., *Internal and external validation of predictive models: A simulation study of bias and precision in small samples.* Journal of Clinical Epidemiology, 2003. **56**(5): p. 441-447.
261. Garson, G.D. *Logistic Regression*. 2009; Available from: http://faculty.chass.ncsu.edu/garson/pa765/statnote.htm.
262. Hosmer, D.W. and S. Lemeshow, *Applied Logistic Regression*. Second ed 2000, New York: John Wiley & Sons.
263. Jackson, D., et al., *Systematically missing confounders in individual participant data meta-analysis of observational cohort studies.* Stat Med, 2009. **28**(8): p. 1218-37.
264. Altman, D.G. and P. Royston, *What do we mean by validating a prognostic model?* Stat Med, 2000. **19**(4): p. 453-73.
265. Justice, A.C., K.E. Covinsky, and J.A. Berlin, *Assessing the Generalizability of Prognostic Information.* Ann Intern Med, 1999. **130**(6): p. 515-524.
266. D'Agostino, R.B., Sr, et al., *General Cardiovascular Risk Profile for Use in Primary Care: The Framingham Heart Study.* Circulation, 2008. **117**(6): p. 743-753.
267. Haybittle, J.L., et al., *A prognostic index in primary breast cancer.* Br J Cancer, 1982. **45**(3): p. 361-6.
268. Wilson, P.W.F., et al., *Prediction of coronary heart disease using risk factor categories.* Circulation, 1998. **97**(18): p. 1837-1847.
269. Altman, D.G. and J.M. Bland, *Diagnostic tests. 1: Sensitivity and specificity.* BMJ, 1994. **308**(6943): p. 1552.
270. Altman, D.G. and J.M. Bland, *Diagnostic tests 2: Predictive values.* BMJ, 1994. **309**(6947): p. 102.
271. Newson, R., *Comparing the predictive power of survival models using Harrell's c or Somers' D*, 2010, National Heart and Lung Institute, Imperial College London: London. p. 18.
272. Altman, D.G. and J.M. Bland, *Diagnostic tests 3: receiver operating characteristic plots.* BMJ, 1994. **309**(6948): p. 188.
273. Mekhail, T.M., et al., *Validation and extension of the Memorial Sloan-Kettering prognostic factors model for survival in patients with previously untreated metastatic renal cell carcinoma.* J Clin Oncol, 2005. **23**(4): p. 832-41.
274. Royston, P., M.K.B. Parmar, and R. Sylvester, *Construction and validation of a prognostic model across several studies, with an application in superficial bladder cancer.* Statistics in Medicine, 2004. **23**(6): p. 907-926.
275. Jorgensen, A.L. and M. Pirmohamed, *Rsik modeling strategies for pharmacogenetic studies.* Future Medicine, 2011. **12**(3).
276. Graf, E., et al., *Assessment and comparison of prognostic classification schemes for survival data.* Stat Med, 1999. **18**(17-18): p. 2529-45.
277. Newson, R., *Confidence intervals for rank statistics: Somers' D and extensions.* Stata Journal, 2006. **6**(3): p. 309-334.
278. Newson, R., *Parameters behind "nonparametric" statistics: Kendall's tau, Somers' D and median differences.* The Stata Journal, 2002. **2**(1): p. 45-64.

279. Newson, R., *SOMERSD: Stata module to calculate Kendall's tau-a, Somers' D and median differences*, 2008, Stata.
280. May, M., et al., *Development and validation of a prognostic model for survival time data: application to prognosis of HIV positive patients treated with antiretroviral therapy.* Statistics in Medicine, 2004. **23**(15): p. 2375-2398.
281. Spiegelhalter, D.J., et al., *Bayesian measures of model complexity and fit.* Journal of the Royal Statistical Society Series B-Statistical Methodology, 2002. **64**: p. 583-616.
282. Sauerbrei, W., et al., *Validation of existing and development of new prognostic classification schemes in node negative breast cancer.* Breast Cancer Research and Treatment, 1997. **42**(2): p. 149-163.
283. Royston, P. and W. Sauerbrei, *A new measure of prognostic separation in survival data.* Stat Med, 2004. **23**(5): p. 723-48.
284. Machin, D. and M.J. Campbell, *Design of Studies for Medical Research* 2005, Chichester: John Wiley & Sons Ltd.
285. Blom, G., *Statistical Estimates and Transformed Beta-Variables* 1958, New York: Wiley.
286. Newson, R., *Efficient calculation of jackknife confidence intervals for rank statistics.* Journal of Statistical Software, 2006. **15**(1): p. -.
287. Burton, A. and D.G. Altman, *Missing covariate data within cancer prognostic studies: a review of current reporting and proposed guidelines.* Br J Cancer, 2004. **91**(1): p. 4-8.
288. Royston, P. and M.K. Parmar, *Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects.* Stat Med, 2002. **21**(15): p. 2175-97.
289. van Houwelingen, H.C., *Validation, calibration, revision and combination of prognostic survival models.* Stat Med, 2000. **19**(24): p. 3401-15.
290. Royston, P. and M.K. Parmar, *External Validation and Updating of a Prognostic Survival Model*, 2010, University College London: London. p. 24.
291. Miller, M.E., S.L. Hui, and W.M. Tierney, *Validation techniques for logistic regression models.* Stat Med, 1991. **10**(8): p. 1213-26.
292. Schafer, J.L. and J.W. Graham, *Missing data: Our view of the state of the art.* Psychological Methods, 2002. **7**(2): p. 147-177.
293. Little, R.J. and D.B. Rubin, *Statistical Analysis with Missing Data*. Wiley Series in Probability and Mathematical Statistics 1987: John Wiley & Sons, Inc.
294. Vach, W., *Some issues in estimating the effect of prognostic factors from incomplete covariate data.* Statistics in Medicine, 1997. **16**(1-3): p. 57-72.
295. Greenland, S. and W.D. Finkle, *A critical look at methods for handling missing covariates in epidemiologic regression analyses.* Am J Epidemiol, 1995. **142**(12): p. 1255-64.
296. Rubin, D.B., *Multiple Imputation for Nonresponse in Surveys* 1987, New York: John Wiley and Sons.
297. Schafer, J.L., *Analysis of Incomplete Multivariate Data* 1997, New York: Chapman & Hall.
298. Allison, P.D., *Missing Data*. Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-136, ed. M.S. Lewis-Beck 2002, Thousand Oaks, CA: Sage.
299. Koopman, L., et al., *Comparison of methods of handling missing data in individual patient data meta-analyses: An empirical example on antibiotics in children with acute otitis media.* American Journal of Epidemiology, 2008. **167**(5): p. 540-545.
300. Rubin, D.B., *Inference and Missing Data.* Biometrika, 1976. **63**(3): p. 581-590.

301. Vach, M. and M. Blettner, *Missing data in epidemiological studies*, in *Encyclopedia of Biostatistics*, P. Armitage and T. Colton, Editors. 1998, Wiley and Sons: New York. p. 2641-2654.
302. Fitzmaurice, G., *Missing data: Implications for analysis.* Nutrition, 2008. **24**(2): p. 200-202.
303. Pugh, M., et al., *Inference in the Cox proportional hazards model with missing covariates*, 1992, Department of Biostatistics, Harvard School of Public Health: Boston.
304. Little, R.J.A., *Missing-Data Adjustments in Large Surveys.* Journal of Business & Economic Statistics, 1988. **6**(3): p. 287-296.
305. Jones, M.P., *Indicator and stratification methods for missing explanatory variables in multiple linear regression.* Journal of the American Statistical Association, 1996. **91**(433): p. 222-230.
306. Dempster, A.P., N.M. Laird, and D.B. Rubin, *Maximum Likelihood from Incomplete Data Via Em Algorithm.* Journal of the Royal Statistical Society Series B-Methodological, 1977. **39**(1): p. 1-38.
307. Buck, S.F., *A Method of Estimation of Missing Values in Multivariate Data Suitable for Use with an Electronic-Computer.* Journal of the Royal Statistical Society Series B-Statistical Methodology, 1960. **22**(2): p. 302-306.
308. Rubin, D.B. and N. Schenker, *Multiple Imputation for Interval Estimation from Simple Random Samples with Ignorable Nonresponse.* Journal of the American Statistical Association, 1986. **81**(394): p. 366-374.
309. Perez, A., et al., *Use of the mean, hot deck and multiple imputation techniques to predict outcome in intensive care unit patients in Colombia.* Stat Med, 2002. **21**(24): p. 3885-96.
310. Schoier, G. *On partial nonresponse situations: the hot deck imputation method*. in *Bulletin of the International Statistical Institute*. 1999. Finland.
311. Molenberghs, G. and M.G. Kenward, *Missing Data (in Clinical Studies)*. Statistics in Practice, ed. S. Senn and V. Barnett 2007, Chichester: John Wiley & Sons Ltd.
312. Anderson, T.W., *Maximum-Likelihood Estimates for a Multivariate Normal-Distribution When Some Observations Are Missing.* Journal of the American Statistical Association, 1957. **52**(278): p. 200-203.
313. Finkbeiner, C., *Estimation for the Multiple Factor Model When Data Are Missing.* Psychometrika, 1979. **44**(4): p. 409-420.
314. Enders, C.K. and D.L. Bandalos, *The Relative Performance of Full Information Maximum Likelihood for Missing Data in Structural Equation Models.* Structural Equation Modeling, 2001. **8**(3): p. 430-457.
315. Bonnett, L., et al., *Prognostic factors for time to treatment failure and time to 12 months of remission for patients with focal epilepsy: post-hoc, subgroup analyses of data from the SANAD trial.* Lancet Neurol, 2012. **11**(4): p. 331-40.
316. Claekens, G. and N.L. Hjort, *Model Selection and Model Averaging*. Cambridge Series in Statistical and Probabilistic Mathematics 2008, Cambridge, U.K.: Cambridge University Press.
317. Bender, R., T. Augustin, and M. Blettner, *Generating survival times to simulate Cox proportional hazards models.* Statistics in Medicine, 2005. **24**(11): p. 1713-1723.
318. Mood, A.M., F.A. Graybill, and D.C. Boes, *Introduction to the Theory of Statistics* 1974, New York: McGraw-Hill.
319. Burton, A., et al., *The design of simulation studies in medical statistics.* Statistics in Medicine, 2006. **25**(24): p. 4279-4292.
320. Keene, O.N., *The log transformation is special.* Stat Med, 1995. **14**(8): p. 811-9.

321. Marshall, A., et al., *Comparison of techniques for handling missing covariate data within prognostic modelling studies: a simulation study.* BMC Med Res Methodol, 2010. **10**: p. 7.
322. Diaz-Emparanza, I., *Is a small Monte Carlo analysis a good analysis? Checking the size, power and consistency of a simulation-based test.* Statistical Papers, 2002. **43**(4): p. 567-577.
323. Collins, L.M., J.L. Schafer, and C.M. Kam, *A comparison of inclusive and restrictive strategies in modern missing data procedures.* Psychological Methods, 2001. **6**(4): p. 330-351.
324. Lawn, N.D., et al., *Injuries due to seizures in persons with epilepsy: a population-based study.* Neurology, 2004. **63**(9): p. 1565-70.
325. Julious, S.A., *Using confidence intervals around individual means to assess statistical significance between two means.* Pharmaceutical Statistics, 2004. **3**(3): p. 217-222.
326. Tangri, N., et al., *A Predictive Model for Progression of Chronic Kidney Disease to Kidney Failure.* Jama-Journal of the American Medical Association, 2011. **305**(15): p. 1553-1559.
327. Pearson, K., *Contributions to the mathematical theory of evolution.* Philosophical Transactions of the Royal Society of London A, 1894. **185**: p. 71-110.
328. Weldon, W.F.R., *Certain correlatd variations in Crangon vulgaris.* Proceedings of the Royal Society of London, 1892. **51**: p. 2-21.
329. Weldon, W.F.R., *On certain correlated variations in Carcinus moenas.* Proceedings of the Royal Society of London, 1893. **54**: p. 318-329.
330. Charlier, C.V.L., *Researchers into the theory of probability*, 1906, Lunds Universities Arskrift, Ny foljd.
331. Charlier, C.V.L. and S.D. Wicksell, *On the dissection of frequency functions*, 1924, Arkiv for Mathematik Astronomi och Fysik.
332. Stromgren, B., *Tables and diagrams for dissecting a frequency curve into components by the half-invariant method*, 1934, Skandinavian Aktuarietidskr. p. 7-54.
333. Cohen, A.C., *Estimation in Mixtures of 2 Normal Distributions.* Technometrics, 1967. **9**(1): p. 15-&.
334. Tan, W.Y. and W.C. Chang, *Some comparisons of the method of moments and the method of maximum likelihood in estimating parameters of a mixture of two normal densities.* Journal of the American Statistical Association, 1972. **67**: p. 702-708.
335. Fryer, J.G. and Robertso.Ca, *Comparison of Some Methods for Estimating Mixed Normal Distributions.* Biometrika, 1972. **59**(3): p. 639-648.
336. Lindsay, B.G. and P. Basak, *Multivariate Normal Mixtures - a Fast Consistent Method of Moments.* Journal of the American Statistical Association, 1993. **88**(422): p. 468-476.
337. Rao, C.R., *The Utilization of Multiple Measurements in Problems of Biological Classification.* Journal of the Royal Statistical Society Series B-Statistical Methodology, 1948. **10**(2): p. 159-203.
338. Butler, R.W., *Predictive Likelihood Inference with Applications.* Journal of the Royal Statistical Society Series B-Methodological, 1986. **48**(1): p. 1-38.
339. Newcomb, S., *A generalized theory of the combination of observations so as to obtain the best result.* American Journal of Mathematics, 1886. **8**: p. 383-394.
340. Hasselbl.V, *Estimation of Parmaeters for a Mixture of Normal Distributions.* Technometrics, 1966. **8**(3): p. 431-&.

341. Hasselbl.V, *Estimation of Finite Mixtures of Distributions from Exponential Family.* Journal of the American Statistical Association, 1969. **64**(328): p. 1459-&.
342. Everitt, B.S., *An introduction to finite mixture distributions.* Stat Methods Med Res, 1996. **5**(2): p. 107-27.
343. Wolfe, J.H., *A computer program for the maximum likelihood analysis of types.*, in *Technical Bulletin* 1965, US Naval Personnel Research Activity: San Diego.
344. Wolfe, J.H., *NORMIX: computation methods for estimating the parameters of multivariate normal mixtures of distributions*, in *Research Memorandum* 1967, US Naval Personnel Research Activity: San Diego.
345. Wolfe, J.H., *Pattern Clustering by Multivariate Mixture Analysis.* Multivariate Behavioral Research, 1970. **5**(3): p. 329-350.
346. Day, N.E., *Estimating Components of a Mixture of Normal Distributions.* Biometrika, 1969. **56**(3): p. 463-&.
347. McLachlan, G. and D. Peel, *Finite Mixture Models*. Wiley series in probability and statistics. Applied probability and statistics section. 2000, New York: Wiley.
348. McLachlan, G.J. and K.E. Basford, *Mixture models; inference and applications to clustering* 1988, New York: Marcel Dekker, Inc.
349. Everitt, B.S. and D.J. Hand, *Finite mixture distributions* 1981, London: Chapman & Hall.
350. Fowlkes, E.B., *Some Methods for Studying the Mixture of 2 Normal (Lognormal) Distributions.* Journal of the American Statistical Association, 1979. **74**(367): p. 561-575.
351. Andrews, R.L. and I.S. Currim, *A comparison of segment retention criteria for finite mixture logit models.* Journal of Marketing Research, 2003. **40**(2): p. 235-243.
352. Andrews, R.L. and I.S. Currim, *Recovering and profiling the true segmentation structure in markets: an empirical investigation.* International Journal of Research in Marketing, 2003. **20**(2): p. 177-192.
353. Bozdogan, H., *Mixture-Model Cluster-Analysis Using Model Selection Criteria and a New Informational Measure of Complexity.* Proceedings of the First Us / Japan Conference on the Frontiers of Statistical Modeling: An Informational Approach, Vols 1-3, 1994: p. B69-B113.
354. Wedel, M. and W.S. Desarbo, *A Mixture Likelihood Approach for Generalized Linear-Models.* Journal of Classification, 1995. **12**(1): p. 21-55.
355. Wolfe, J.H., *A Monte Carlo study of the sampling distributions of the likelihood ratio for mixtures of multinormal distributions*, 1971, Naval Personnel and Training Research Laboratory: San Diego.
356. Hunsberger, S., P.S. Albert, and W.B. London, *A finite mixture survival model to characterize risk groups of neuroblastoma.* Statistics in Medicine, 2009. **28**(8): p. 1301-1314.
357. Richardson, S. and P.J. Green, *On Bayesian analysis of mixtures with an unknown number of components.* Journal of the Royal Statistical Society Series B-Methodological, 1997. **59**(4): p. 731-758.
358. Stephens, M., *Bayesian analysis of mixture models with an unknown number of components - An alternative to reversible jump methods.* Annals of Statistics, 2000. **28**(1): p. 40-74.
359. Mendenhall, W. and R.J. Hader, *Estimation of Parameters of Mixed Exponentially Distributed Failure Time Distributions from Censored Life Test Data.* Biometrika, 1958. **45**(3-4): p. 504-520.
360. Kao, J.H.K., *A Graphical Estimation of Mixed Weibull Parameters in Life-Testing of Electron Tubes.* Technometrics, 1959. **1**(4): p. 389-407.

361. McGiffin, D.C., et al., *Aortic valve infection. Risk factors for death and recurrent endocarditis after aortic valve replacement.* J Thorac Cardiovasc Surg, 1992. **104**(2): p. 511-20.
362. Lui, K.J., W.W. Darrow, and G.W. Rutherford, 3rd, *A model-based estimate of the mean incubation period for AIDS in homosexual men.* Science, 1988. **240**(4857): p. 1333-5.
363. McLachlan, G.J. and D.C. McGiffin, *On the role of finite mixture models in survival analysis.* Stat Methods Med Res, 1994. **3**(3): p. 211-26.
364. Greenhouse, J.B. and N.P. Silliman, *Applications of a mixture survival model with covariates to the analysis of a depression prevention trial.* Statistics in Medicine, 1996. **15**(19): p. 2077-2094.
365. Kent, A.S.V., *Gauss Systems*, 1995: Washington.
366. Li, C.S., J.M.G. Taylor, and J.P. Sy, *Identifiability of cure models.* Statistics & Probability Letters, 2001. **54**(4): p. 389-395.
367. Yu, B.B., et al., *Cure fraction estimation from the mixture cure models for grouped survival data.* Statistics in Medicine, 2004. **23**(11): p. 1733-1747.
368. Farewell, V.T., *The use of mixture models for the analysis of survival data with long-term survivors.* Biometrics, 1982. **38**(4): p. 1041-6.
369. McLachlan, G.J., et al., *Modelling via finite mixtures of time to reoperationg following aortic valve replacement*, in *Technical Report* 1993, The University of Queensland: Brisbane: Centre of Statistics.
370. Kuk, A.Y.C. and C.H. Chen, *A Mixture Model Combining Logistic-Regression with Proportional Hazards Regression.* Biometrika, 1992. **79**(3): p. 531-541.
371. Peng, Y. *GFCURE*. 2007; Available from: http://post.queensu.ca/~pengp/software.html.
372. Peng, Y.W., K.B.G. Dear, and J.W. Denham, *A generalized F mixture model for cure rate estimation.* Statistics in Medicine, 1998. **17**(8): p. 813-830.
373. Farrow, M. *Parametric Models*. 2009; Available from: www.mas.ncl.ac.uk/~nmf16/teaching/mas3311.
374. Zhang, D., *Modeling Survival Data with Parametric Regression Models*, in *Analysis of Survival Data* 2005: North Carolina State University.
375. Humphrey, J.D. and K.R. Rajagopal, *A constrained mixture model for growth and remodeling of soft tissues.* Mathematical Models & Methods in Applied Sciences, 2002. **12**(3): p. 407-430.
376. McLachlan, G.J., R.W. Bean, and D. Peel, *A mixture model-based approach to the clustering of microarray expression data.* Bioinformatics, 2002. **18**(3): p. 413-422.
377. Albert, P.S., *A 2-State Markov Mixture Model for a Time-Series of Epileptic Seizure Counts.* Biometrics, 1991. **47**(4): p. 1371-1381.
378. Le, N.D., B.G. Leroux, and M.L. Puterman, *Exact Likelihood Evaluation in a Markov Mixture Model for Time-Series of Seizure Counts.* Biometrics, 1992. **48**(1): p. 317-323.
379. Ypma, T.J., *Historical development of the Newton-Raphson method.* Siam Review, 1995. **37**(4): p. 531-551.
380. MathWorks, *MATLAB*, 2011.
381. Vermunt, J.K. and J. Magidson, *Latent GOLD 4.0 User's Guide*, 2005, Statistical Innovations Inc.: Belmont, MA.
382. Gudicha, D.W. and J.K. Vermunt, *Mixture model clustering with covariates using adjusted three-step approaches*, in *Proceedings GfKl/ICFS*, B. Lausen, D. van den Poel, and A. Ultsch, Editors. 2011, Springer.
383. Everitt, B.S., *A Finite Mixture Model for the Clustering of Mixed-Mode Data.* Statistics & Probability Letters, 1988. **6**(5): p. 305-309.

384. Lawrence, C.J. and W.J. Krzanowski, *Mixture separation for mixed-mode data.* Statistics and Computing, 1996. **6**(1): p. 85-92.
385. Titterington, D.M., *Mixture distributions (update)*, in *Encyclopedia of statistical science (update)*, S.M. Kotz, Editor 1996, Wiley: New York.
386. Ripley, B.D., *Neural networks and related models for classification (with discussion).* Journal of the Royal Statistical Society, Series B, 1994. **56**: p. 409-56.
387. Jorgensen, M.A., *Influence-Based Diagnostics for Finite Mixture Models.* Biometrics, 1990. **46**(4): p. 1047-1058.
388. Lindsay, B.G. and K. Roeder, *Residual Diagnostics for Mixture Models.* Journal of the American Statistical Association, 1992. **87**(419): p. 785-794.
389. Roeder, K., *A Graphical Technique for Determining the Number of Components in a Mixture of Normals.* Journal of the American Statistical Association, 1994. **89**(426): p. 487-495.
390. Royston, P., D.G. Altman, and W. Sauerbrei, *Dichotomizing continuous predictors in multiple regression: a bad idea.* Stat Med, 2006. **25**(1): p. 127-41.
391. Yates, J.F., *External Correspondence - Decompositions of the Mean Probability Score.* Organizational Behavior and Human Performance, 1982. **30**(1): p. 132-156.
392. Spiegelhalter, D.J., *Probabilistic Prediction in Patient-Management and Clinical-Trials.* Statistics in Medicine, 1986. **5**(5): p. 421-433.
393. Smith, C.T., P.R. Williamson, and A.G. Marson, *Investigating heterogeneity in an individual patient data meta-analysis of time to event outcomes.* Statistics in Medicine, 2005. **24**(9): p. 1307-1319.

Appendix A – Further Results from Chapter 5

Table 53: The MESS Study: Risk of seizure recurrence at 12 months for patients with various combinations of risk factors -estimated from the multivariable regression model

Variable						Immediate Treatment		Delayed Treatment	
Aetiology	First Degree Relative with Epilepsy	Seizures Only While Asleep	EEG Results	CT/MRI Scan Results	Months from index seizure	Risk of seizure in next 12 months (%, 95% CI)	Months from index seizure until annual risk falls <20%	Risk of seizure in next 12 months (%, 95% CI)	Months from index seizure until annual risk falls <20%
Not remote symptomatic	No	No	Normal	Normal	6 12	13 (10 to 16)* 7 (5 to 10)*	1.2	16 (12 to 19)* 9 (6 to 12)*	3.2
Not remote symptomatic	No	No	Normal	Abnormal	6 12	14 (11 to 17)* 8 (5 to 11)*	2.1	17 (14 to 20) 10 (7 to 12)*	3.6
Not remote symptomatic	No	No	Normal	Not Done	6 12	17 (13 to 20) 9 (7 to 12)*	3.5	20 (17 to 23) 11 (9 to 14)*	5.7
Not remote symptomatic	No	No	Abnormal	Normal	6 12	20 (16 to 23) 11 (9 to 14)*	5.5	23 (20 to 26) 13 (11 to 16)*	7.0
Not remote symptomatic	No	No	Abnormal	Abnormal	6 12	21 (18 to 24) 12 (9 to 15)*	6.1	25 (22 to 28)† 14 (12 to 17)*	8.1
Not remote symptomatic	No	No	Abnormal	Not Done	6 12	24 (22 to 27)† 14 (12 to 17 *	7.8	29 (26 to 32)† 17 (15 to 20)	10.0
Not remote symptomatic	No	No	Not Done	Normal	6 12	17 (13 to 20) 9 (7 to 12)*	3.5	20 (17 to 23) 11 (9 to 14)*	5.7
Not remote symptomatic	No	No	Not Done	Abnormal	6 12	18 (15 to 21) 10 (7 to 13)*	4.0	21 (18 to 24) 12 (10 to 15)*	6.2
Not remote symptomatic	No	No	Not Done	Not Done	6 12	21 (18 to 24) 12 (9 to 15)*	6.1	25 (22 to 28)† 14 (12 to 17)*	8.1
Not remote symptomatic	No	Yes	Normal	Normal	6 12	19 (15 to 22) 11 (8 to 13)*	4.8	22 (19 to 25) 13 (10 to 15)*	6.8
Not remote symptomatic	No	Yes	Normal	Abnormal	6 12	20 (19 to 23) 11 (9 to 14)*	5.7	24 (21 to 27)† 14 (11 to 16)*	7.6
Not remote symptomatic	No	Yes	Normal	Not Done	6 12	23 (20 to 26) 14 (11 to 16)*	7.0	28 (25 to 30)† 16 (14 to 19)*	9.8
Not remote symptomatic	No	Yes	Abnormal	Normal	6 12	27 (25 to 30)† 16 (14 to 19)*	9.2	32 (30 to 35)† 19 (17 to 22)	11.8

Not remote symptomatic	No	Yes	Abnormal	Abnormal	6 12	29 (26 to 32)† 17 (15 to 20)	10.0	34 (32 to 37)† 21 (18 to 23)	16.3
Not remote symptomatic	No	Yes	Abnormal	Not Done	6 12	34 (31 to 36)† 20 (18 to 23)	12.5	39 (37 to 42)† 24 (22 to 26)†	24.4
Not remote symptomatic	No	Yes	Not Done	Normal	6 12	23 (20 to 26) 14 (11 to 16)*	7.0	28 (25 to 30)† 16 (14 to 19)*	9.8
Not remote symptomatic	No	Yes	Not Done	Abnormal	6 12	25 (22 to 28)† 15 (12 to 17)*	8.2	29 (27 to 32)† 17 (15 to 20)	10.5
Not remote symptomatic	No	Yes	Not Done	Not Done	6 12	29 (26 to 32)† 17 (15 to 20)	10.0	34 (32 to 37)† 20 (18 to 23)	16.3
Not remote symptomatic	Yes	No	Normal	Normal	6 12	17 (14 to 20) 10 (7 to 12)*	3.6	20 (17 to 23) 12 (9 to 14)*	6.0
Not remote symptomatic	Yes	No	Normal	Abnormal	6 12	18 (15 to 21) 10 (8 to 13)*	4.8	22 (19 to 25) 13 (10 to 15)*	6.5
Not remote symptomatic	Yes	No	Normal	Not Done	6 12	21 (18 to 24) 12 (10 to 15)*	6.3	25 (23 to 28)† 15 (12 to 17)*	8.7
Not remote symptomatic	Yes	No	Abnormal	Normal	6 12	25 (22 to 28)† 15 (12 to 17)*	8.7	30 (27 to 32)† 18 (15 to 20)	10.7
Not remote symptomatic	Yes	No	Abnormal	Abnormal	6 12	27 (24 to 30)† 16 (13 to 18)*	9.2	32 (29 to 34)† 19 (16 to 21)	11.3
Not remote symptomatic	Yes	No	Abnormal	Not Done	6 12	31 (29 to 34)† 19 (16 to 21)	11.3	37 (34 to 39)† 22 (20 to 24)	23.6
Not remote symptomatic	Yes	No	Not Done	Normal	6 12	21 (18 to 24) 12 (10 to 15)*	6.3	25 (23 to 28)† 15 (12 to 17)*	8.7
Not remote symptomatic	Yes	No	Not Done	Abnormal	6 12	23 (20 to 26) 13 (11 to 16)*	7.0	27 (24 to 30)† 16 (13 to 18)*	9.2
Not remote symptomatic	Yes	No	Not Done	Not Done	6 12	27 (24 to 30)† 16 (13 to 18)*	9.2	32 (29 to 34)† 19 (16 to 21)	11.3
Not remote symptomatic	Yes	Yes	Normal	Normal	6 12	24 (21 to 27)† 14 (11 to 17)*	7.7	28 (26 to 31)† 17 (14 to 19)*	10.0
Not remote symptomatic	Yes	Yes	Normal	Abnormal	6 12	26 (23 to 28)† 15 (12 to 17)*	8.8	30 (28 to 33)† 18 (15 to 20)	10.8
Not remote symptomatic	Yes	Yes	Normal	Not Done	6 12	30 (27 to 33)† 18 (15 to 20)	10.7	35 (33 to 38)† 21 (19 to 23)	23.4
Not remote symptomatic	Yes	Yes	Abnormal	Normal	6 12	35 (32 to 37)† 21 (18 to 23)	16.3	40 (38 to 43)† 25 (22 to 27)†	24.6

Not remote symptomatic	Yes	Yes	Abnormal	Abnormal	6 12	37 (34 to 39)† 22 (20 to 25)	23.6	43 (41 to 45)† 26 (24 to 29)†	25.1
Not remote symptomatic	Yes	Yes	Abnormal	Not Done	6 12	42 (40 to 45)† 26 (24 to 28)†	25.1	49 (47 to 51)† 31 (29 to 33)†	27.0
Not remote symptomatic	Yes	Yes	Not Done	Normal	6 12	30 (27 to 32)† 18 (15 to 20)	10.7	35 (33 to 37)† 21 (19 to 23)	23.4
Not remote symptomatic	Yes	Yes	Not Done	Abnormal	6 12	32 (29 to 34)† 19 (16 to 21)	11.3	37 (35 to 40)† 22 (20 to 25)	24.1
Not remote symptomatic	Yes	Yes	Not Done	Not Done	6 12	37 (34 to 39)† 22 (20 to 25)	23.6	43 (41 to 45)† 26 (24 to 29)†	25.1
Remote symptomatic	No	No	Normal	Normal	6 12	17 (14 to 20) 10 (7 to 12)*	3.6	20 (17 to 23) 12 (9 to 14)*	6.0
Remote symptomatic	No	No	Normal	Abnormal	6 12	18 (15 to 21) 10 (8 to 13)*	4.8	22 (19 to 25) 13 (10 to 15)*	6.5
Remote symptomatic	No	No	Normal	Not Done	6 12	21 (18 to 24) 12 (10 to 15)*	6.3	25 (23 to 28)† 15 (12 to 17)*	8.7
Remote symptomatic	No	No	Abnormal	Normal	6 12	25 (22 to 28)† 15 (12 to 17)*	8.7	30 (27 to 32)† 18 (15 to 20)	10.7
Remote symptomatic	No	No	Abnormal	Abnormal	6 12	27 (24 to 40)† 16 (13 to 18)*	9.2	32 (29 to 34)† 19 (16 to 21)	11.3
Remote symptomatic	No	No	Abnormal	Not Done	6 12	31 (29 to 34)† 19 (16 to 21)	11.3	37 (34 to 39)† 22 (20 to 24)	23.6
Remote symptomatic	No	No	Not Done	Normal	6 12	21 (18 to 24) 12 (10 to 15)*	6.3	25 (23 to 28)† 15 (12 to 17)*	8.7
Remote symptomatic	No	No	Not Done	Abnormal	6 12	23 (20 to 26) 13 (11 to 16)*	7.0	27 (24 to 30)† 16 (13 to 18)*	9.2
Remote symptomatic	No	No	Not Done	Not Done	6 12	27 (24 to 30)† 16 (13 to 18)*	9.2	32 (29 to 34)† 19 (16 to 21)	11.3
Remote symptomatic	No	Yes	Normal	Normal	6 12	24 (21 to 27)† 14 (11 to 17)*	7.7	28 (26 to 31)† 17 (14 to 19)*	10.0
Remote symptomatic	No	Yes	Normal	Abnormal	6 12	26 (23 to 28)† 15 (12 to 17)*	8.8	30 (28 to 33)† 18 (15 to 20)	10.8
Remote symptomatic	No	Yes	Normal	Not Done	6 12	30 (27 to 33)† 18 (15 to 20)	10.7	35 (33 to 38)† 21 (19 to 23)	23.4
Remote symptomatic	No	Yes	Abnormal	Normal	6 12	35 (32 to 37)† 21 (18 to 23)	16.3	40 (38 to 43)† 25 (22 to 27)†	24.6

Remote symptomatic	No	Yes	Abnormal	Abnormal	6 12	37 (34 to 39)† 22 (20 to 25)	23.6	43 (41 to 45)† 26 (24 to 29)†	25.1
Remote symptomatic	No	Yes	Abnormal	Not Done	6 12	42 (40 to 45)† 26 (24 to 28)†	25.1	49 (47 to 51)† 31 (29 to 33)†	27.0
Remote symptomatic	No	Yes	Not Done	Normal	6 12	30 (27 to 32)† 18 (15 to 20)	11.7	35 (33 to 37)† 21 (19 to 23)	23.4
Remote symptomatic	No	Yes	Not Done	Abnormal	6 12	32 (29 to 34)† 19 (16 to 21)	11.3	37 (35 to 40)† 22 (20 to 25)	24.1
Remote symptomatic	No	Yes	Not Done	Not Done	6 12	37 (34 to 39)† 22 (20 to 25)	23.6	43 (41 to 45)† 26 (24 to 29)†	25.1
Remote symptomatic	Yes	No	Normal	Normal	6 12	22 (19 to 25) 13 (10 to 15)*	6.7	26 (23 to 29)† 15 (13 to 18)*	8.8
Remote symptomatic	Yes	No	Normal	Abnormal	6 12	24 (21 to 26)† 14 (11 to 16)*	7.6	28 (25 to 31)† 16 (14 to 19)*	9.8
Remote symptomatic	Yes	No	Normal	Not Done	6 12	27 (25 to 30)† 16 (14 to 19) *	9.8	32 (30 to 35)† 19 (17 to 22)	12.0
Remote symptomatic	Yes	No	Abnormal	Normal	6 12	32 (29 to 35)† 19 (17 to 21)	11.8	37 (35 to 40)† 23 (20 to 25)	24.1
Remote symptomatic	Yes	No	Abnormal	Abnormal	6 12	34 (32 to 37)† 20 (18 to 23)	16.3	40 (37 to 42)† 24 (22 to 27)†	24.6
Remote symptomatic	Yes	No	Abnormal	Not Done	6 12	39 (37 to 42)† 24 (22 to 26)†	24.4	46 (43 to 48)† 28 (26 to 30)†	25.6
Remote symptomatic	Yes	No	Not Done	Normal	6 12	27 (25 to 30)† 16 (14 to 19)*	9.8	32 (30 to 35)† 19 (17 to 22)	12.0
Remote symptomatic	Yes	No	Not Done	Abnormal	6 12	29 (27 to 32)† 17 (15 to 20)	10.0	34 (32 to 37)† 21 (18 to 23)	16.3
Remote symptomatic	Yes	No	Not Done	Not Done	6 12	34 (31 to 37)† 20 (18 to 23)	16.3	40 (37 to 42)† 24 (22 to 26)†	24.4
Remote symptomatic	Yes	Yes	Normal	Normal	6 12	31 (28 to 33)† 18 (16 to 21)	10.8	36 (33 to 38)† 22 (19 to 24)	23.5
Remote symptomatic	Yes	Yes	Normal	Abnormal	6 12	33 (30 to 35)† 19 (17 to 22)	12.0	38 (36 to 40)† 23 (21 to 25)†	24.2
Remote symptomatic	Yes	Yes	Normal	Not Done	6 12	38 (35 to 40)† 23 (21 to 25)†	24.2	44 (42 to 46)† 27 (25 to 29)†	25.5
Remote symptomatic	Yes	Yes	Abnormal	Normal	6 12	43 (41 to 45)† 27 (25 to 29)†	25.5	50 (48 to 52)† 32 (30 to 34)†	27.0

Remote symptomatic	Yes	Yes	Abnormal	Abnormal	6 12	46 (44 to 48)† 29 (26 to 31)†	25.6	52 (51 to 54)† 34 (32 to 35)†	28.8
Remote symptomatic	Yes	Yes	Abnormal	Not Done	6 12	52 (50 to 54)† 33 (31 to 35)†	27.0	59 (57 to 61)† 39 (37 to 41)†	55.6
Remote symptomatic	Yes	Yes	Not Done	Normal	6 12	38 (35 to 40)† 23 (21 to 25)†	24.2	44 (42 to 46)† 27 (25 to 29)†	25.5
Remote symptomatic	Yes	Yes	Not Done	Abnormal	6 12	40 (38 to 42)† 24 (22 to 27)†	24.6	46 (44 to 48)† 29 (27 to 31)†	25.7
Remote symptomatic	Yes	Yes	Not Done	Not Done	6 12	46 (44 to 48)† 28 (26 to 31)†	25.6	52 (51 to 54)† 33 (31 to 35)†	28.8

*Confidence interval lies completely below 20%. †Confidence interval lies completely above 20%.

Appendix B – Search Strategy used in Chapter 6

MEDLINE (Ovid) search strategy:

1. randomized controlled trial.pt.

2. controlled clinical trial.pt.

3. (randomized or randomised).ab.

4. placebo.ab.

5. clinical trials as topic.sh.

6. randomly.ab.

7. trial.ti.

8. epidemiologic studies/ or case-control studies/ or cohort studies/ or longitudinal studies/ or follow-up studies/ or prospective studies/

9. (case control or case-control).tw.

10. (cohort adj (study or studies)).tw.

11. cohort analy$.tw.

12. (("follow up" or "follow-up") adj (study or studies)).tw.

13. (observational adj (study or studies)).tw.

14. (longitudinal adj (study or studies)).tw.

15. or/1-14

16. (epilep$ or convulsion$ or seizure$).ti,ab.

17. exp Epilepsy/

18. exp Seizures/

19. convulsive disorder.tw.

20. or/16-19

21. 15 and 20

22. exp animals/ not humans.sh.

23. 21 not 22

24. withdraw$.ti,ab.

25. alcohol withdrawal.tw.

26. alcohol dependence.tw.

27. 25 or 26

28. 24 not 27

29. 23 and 28

30. limit 29 to english language

Appendix C – Further Results from Chapter 8

Numerical results for the simulation study. Values used as part of the multivariable normal distribution can be seen in section C.1 while numbers of failed simulations can be seen in section C.2. Results of the three measures of external validation can be seen in sections C.3 to C.5. Coefficient values are shown in section C.6 and measures of bias, mean square error and coverage can be seen in section C.7.

C.1 Multivariate Normal Distribution

Values obtained from SANAD data.

C.1.1 Means

Binary	Continuous	Categorical
0.55	38.30	1.21

C.1.2 Covariance Matrix

	Binary	Continuous	Categorical
Binary	0.25	-0.15	0.01
Continuous	-0.15	336.36	-0.26
Categorical	0.01	-0.26	0.35

C.2 Number of failed simulations

Scenario	Imputation Method				
	(1) – Variable Matching	(2) – Random Selection with Replacement	(3) – Imputation via Estimation	(4) – Hot Deck	(5) – Multiple Imputation
1a & 1b	0				
1a & 1c	0	0	0	0	0
2a & 2b	0	0	0	0	0
3a & 3b	0	0	0	0	0
4a & 4b	0				
4a & 4c	0	0	0	0	0
5a & 5b	0	0	0	0	0
6a & 6b	0	0	0	0	0

C.3 Concordance Statistic

C.3.1 No missing data

Scenario	Development Concordance	Validation Concordance	Mean difference	% p-values ≤ 0.05	% p-values ≤ 0.10
1a & 1b	0.56	0.56	0.00	99.8	100
4a & 4b	0.56	0.56	0.00	99.9	100

C.3.2 Variable Matching

Scenario	Development Deviance	Validation Deviance	Mean difference	% p-values ≤ 0.05	% p-values ≤ 0.10
1a & 1c	0.54	0.54	0.01	97.1	100
2a & 2b	0.55	0.55	0.00	99.9	100
3a & 3b	0.55	0.55	0.00	99.9	100
4a & 4c	0.54	0.54	0.00	99.2	100
5a & 5b	0.55	0.55	0.00	100	100
6a & 6b	0.55	0.55	0.00	100	100

C.3.3 Random Selection with Replacement

Scenario	Development Deviance	Validation Deviance	Mean difference	% p-values ≤ 0.05	% p-values ≤ 0.10
1a & 1c	0.56	0.56	0.00	99.8	100
2a & 2b	0.56	0.56	0.00	99.8	100
3a & 3b	0.56	0.56	0.00	99.6	100
4a & 4c	0.56	0.56	0.00	100	100
5a & 5b	0.56	0.56	0.00	99.9	100
6a & 6b	0.56	0.56	0.00	100	100

C.3.4 Single Imputation via Estimation

Scenario	Development Deviance	Validation Deviance	Mean difference	% p-values ≤ 0.05	% p-values ≤ 0.10
1a & 1c	0.56	0.56	0.00	99.8	100
2a & 2b	0.56	0.56	0.00	99.6	100
3a & 3b	0.56	0.56	0.00	99.4	100
4a & 4c	0.56	0.56	0.00	99.9	100
5a & 5b	0.56	0.56	0.00	99.9	100
6a & 6b	0.56	0.56	0.00	100	100

C.3.5 Hot Deck Imputation

Scenario	Development Deviance	Validation Deviance	Mean difference	% p-values ≤ 0.05	% p-values ≤ 0.10
1a & 1c	0.56	0.56	0.00	99.6	100
2a & 2b	0.56	0.56	0.00	99.7	100
3a & 3b	0.56	0.56	0.00	99.6	100
4a & 4c	0.56	0.56	0.00	99.8	100
5a & 5b	0.56	0.56	0.00	100	100
6a & 6b	0.56	0.56	0.00	100	100

C.3.6 Random Selection with Replacement Multiple Times

Scenario	Development Deviance	Validation Deviance	Mean difference	% p-values ≤ 0.05	% p-values ≤ 0.10
1a & 1c	0.56	0.56	0.00	100	100
2a & 2b	0.56	0.56	0.00	100	100
3a & 3b	0.56	0.56	0.00	100	100
4a & 4c	0.56	0.56	0.00	100	100
5a & 5b	0.56	0.56	0.00	100	100
6a & 6b	0.56	0.56	0.00	100	100

C.4 Deviance Statistic

Mean differences are mean differences in deviance statistics between the development and validation statistics over 1400 simulations. The p-value is the mean p-value for the chi-squared test of difference in deviances between the development and validation data.

C.4.1 No missing data

Scenario	Development Deviance	Validation Deviance	Mean difference	p-value for difference	% p-values ≤ 0.05	% p-values ≤ 0.10
1a & 1b	45.63	28.12	-17.51	0.11	73.9	78.6
4a & 4b	46.50	52.14	5.64	0.17	61.1	68.8

C.4.2 Variable Matching

Scenario	Development Deviance	Validation Deviance	Mean difference	p-value for difference	% p-values ≤ 0.05	% p-values ≤ 0.10
1a & 1c	15.19	4.72	-10.46	0.20	52.1	61.6
2a & 2b	37.74	22.59	-15.16	0.14	68.4	73.6
3a & 3b	37.56	22.51	-15.06	0.12	70.6	75.5
4a & 4c	15.30	7.30	-8.00	0.24	43.2	51.2
5a & 5b	37.42	42.96	5.54	0.18	58.6	65.4
6a & 6b	38.07	42.61	4.54	0.17	60.3	68.0

C.4.3 Random Selection with Replacement

Scenario	Development Deviance	Validation Deviance	Mean difference	p-value for difference	% p-values ≤ 0.05	% p-values ≤ 0.10
1a & 1c	45.74	27.93	-17.81	0.11	74.1	79.0
2a & 2b	45.71	27.51	-18.20	0.11	74.7	79.1
3a & 3b	45.67	27.99	-17.68	0.12	73.4	78.5
4a & 4c	45.62	52.63	7.02	0.16	64.4	70.1
5a & 5b	45.45	51.88	6.44	0.16	64.8	70.6
6a & 6b	46.10	52.83	6.74	0.14	66.5	72.4

C.4.4 Single Imputation via Estimation

Scenario	Development Deviance	Validation Deviance	Mean difference	p-value for difference	% p-values ≤ 0.05	% p-values ≤ 0.10
1a & 1c	45.74	27.74	-18.00	0.11	74.3	79.2
2a & 2b	45.71	27.70	-18.01	0.11	73.7	78.4
3a & 3b	45.67	28.15	-17.52	0.11	73.9	78.9
4a & 4c	45.62	52.69	7.07	0.16	62.4	68.4
5a & 5b	45.45	51.58	6.13	0.15	64.4	71.3
6a & 6b	46.10	52.75	6.65	0.15	65.9	72.1

C.4.5 Hot Deck Imputation

Scenario	Development Deviance	Validation Deviance	Mean difference	p-value for difference	% p-values ≤ 0.05	% p-values ≤ 0.10
1a & 1c	45.74	28.23	-17.52	0.12	73.2	77.9
2a & 2b	45.71	27.63	-18.08	0.11	72.9	78.1
3a & 3b	45.64	27.98	-17.69	0.11	75.1	79.1
4a & 4c	45.62	52.18	6.57	0.16	65.0	69.8
5a & 5b	45.45	51.69	6.24	0.15	63.1	69.8
6a & 6b	46.10	52.95	6.85	0.15	64.9	70.5

C.4.6 Random Selection with Replacement Multiple Times

Scenario	Development Deviance	Validation Deviance	Mean difference	p-value for difference	% p-values ≤ 0.05	% p-values ≤ 0.10
1a & 1c	45.71	27.93	-17.78	0.11	13.4	45.9
2a & 2b	45.63	27.54	-18.08	0.11	13.6	47.1
3a & 3b	45.60	28.15	-17.45	0.11	12.1	44.9
4a & 4c	45.59	52.24	6.64	0.16	2.3	18.9
5a & 5b	45.63	51.75	6.12	0.16	3.0	17.7
6a & 6b	45.71	52.82	7.11	0.15	2.8	18.4

C.5 Prognostic Separation

D values are the mean prognostic separation values over 1400 simulations. ΔD represents the change in prognostic separation values.

C.5.1 No missing data

Scenario	D_{dev}	D_{val}	ΔD	% ΔD (-0.05, 0.05)	% ΔD (-0.10, 0.10)
1a & 1b	0.36	0.37	0.01	41.6	72.6
4a & 4b	0.36	0.36	0.00	49.5	80.4

C.5.2 Variable Matching

Scenario	D_{dev}	D_{val}	ΔD	% ΔD (-0.05, 0.05)	% ΔD (-0.10, 0.10)
1a & 1c	0.21	0.24	0.03	30.4	57.4
2a & 2b	0.39	0.39	0.00	34.4	63.5
3a & 3b	0.31	0.31	0.00	43.6	72.9
4a & 4c	0.21	0.21	0.00	38.8	68.8
5a & 5b	0.39	0.39	0.00	41.7	72.6
6a & 6b	0.31	0.31	0.00	48.7	81.3

C.5.3 Random Selection with Replacement

	D_{dev}	D_{val}	ΔD	% ΔD (-0.05, 0.05)	% ΔD (-0.10, 0.10)
1a & 1c	0.36	0.37	0.01	41.7	73.3
2a & 2b	0.36	0.37	0.01	40.6	73.2
3a & 3b	0.36	0.37	0.01	41.5	72.4
4a & 4c	0.36	0.36	0.00	48.5	80.8
5a & 5b	0.36	0.36	0.00	50.1	82.8
6a & 6b	0.36	0.36	0.00	47.9	80.6

C.5.4 Single Imputation via Estimation

	D_{dev}	D_{val}	ΔD	% ΔD (-0.05, 0.05)	% ΔD (-0.10, 0.10)
1a & 1c	0.36	0.37	0.01	42.9	73.7
2a & 2b	0.36	0.37	0.01	40.0	72.4
3a & 3b	0.36	0.37	0.01	42.6	73.1
4a & 4c	0.36	0.36	0.00	49.7	80.5
5a & 5b	0.36	0.36	0.00	50.6	81.7
6a & 6b	0.36	0.36	0.00	47.7	81.3

C.5.5 Hot Deck Imputation

	D_{dev}	D_{val}	ΔD	% ΔD (-0.05, 0.05)	% ΔD (-0.10, 0.10)
1a & 1c	0.36	0.37	0.01	40.2	72.1
2a & 2b	0.36	0.37	0.01	41.1	72.3
3a & 3b	0.36	0.37	0.01	43.2	72.1
4a & 4c	0.36	0.36	0.00	48.9	80.7
5a & 5b	0.36	0.36	0.00	49.6	82.0
6a & 6b	0.36	0.36	0.00	48.0	81.8

C.5.6 Random Selection with Replacement Multiple Times

Scenario	D_{dev}	D_{val}	ΔD	% ΔD (-0.05, 0.05)	% ΔD (-0.10, 0.10)
1a & 1c	0.36	0.37	0.01	98.1	100
2a & 2b	0.36	0.37	0.01	98.2	100
3a & 3b	0.36	0.37	0.01	97.1	100
4a & 4c	0.36	0.36	0.00	99.8	100
5a & 5b	0.36	0.36	0.00	99.5	100
6a & 6b	0.36	0.36	0.00	99.8	100

C.6 Coefficient Values

The values presented are the mean values over 1400 simulations

C.6.1 No missing data

Scenario	Development								Validation							
	Binary		Continuous		Categorical (Dummy 1)		Categorical (Dummy 2)		Binary		Continuous		Categorical (Dummy 1)		Categorical (Dummy 2)	
	$\tilde{\beta}$	$se(\tilde{\beta})$	$\tilde{\beta}$	$se(\tilde{\beta})$	$\tilde{\beta}$	$se(\tilde{\beta})$	$\tilde{\beta}$	$se(\tilde{\beta})$	$\tilde{\beta}$	$se(\tilde{\beta})$	$\tilde{\beta}$	$se(\tilde{\beta})$	$\tilde{\beta}$	$se(\tilde{\beta})$	$\tilde{\beta}$	$se(\tilde{\beta})$
1a & 1b	-0.19	0.07	-0.18	0.07	-0.52	0.13	-0.50	0.15	-0.19	0.09	-0.18	0.09	-0.53	0.17	-0.51	0.21
4a & 4b	-0.20	0.07	-0.19	0.07	-0.52	0.13	-0.51	0.16	-0.19	0.07	-0.18	0.06	-0.52	0.12	-0.50	0.14

C.6.2 Variable Matching

Scenario	Development								Validation							
	Binary		Continuous		Categorical (Dummy 1)		Categorical (Dummy 2)		Binary		Continuous		Categorical (Dummy 1)		Categorical (Dummy 2)	
	$\tilde{\beta}$	$se(\tilde{\beta})$	$\tilde{\beta}$	$se(\tilde{\beta})$	$\tilde{\beta}$	$se(\tilde{\beta})$	$\tilde{\beta}$	$se(\tilde{\beta})$	$\tilde{\beta}$	$se(\tilde{\beta})$	$\tilde{\beta}$	$se(\tilde{\beta})$	$\tilde{\beta}$	$se(\tilde{\beta})$	$\tilde{\beta}$	$se(\tilde{\beta})$
1a & 1c	-0.18	0.07	-0.18	0.07	NA				-0.20	0.16	-0.15	0.15	NA			
2a & 2b	-0.19	0.07	NA		-0.52	0.13	-0.50	0.15	-0.19	0.09	NA		-0.52	0.17	-0.51	0.21
3a & 3b	NA		-0.18	0.07	-0.52	0.13	-0.49	0.15	NA		-0.18	0.09	-0.52	0.17	-0.49	0.21
4a & 4c	-0.18	0.07	-0.18	0.07	NA				-0.20	0.11	-0.15	0.11	NA			
5a & 5b	-0.19	0.07	NA		-0.52	0.13	-0.50	0.15	-0.19	0.07	NA		-0.52	0.12	-0.50	0.14
6a & 6b	NA		-0.18	0.07	-0.52	0.13	-0.50	0.15	NA		-0.18	0.06	-0.51	0.12	-0.49	0.14

C.6.3 Random Selection with Replacement

Scenario	Development								Validation							
	Binary		Continuous		Categorical (Dummy 1)		Categorical (Dummy 2)		Binary		Continuous		Categorical (Dummy 1)		Categorical (Dummy 2)	
	$\tilde{\beta}$	$se(\tilde{\beta})$	$\tilde{\beta}$	$se(\tilde{\beta})$	$\tilde{\beta}$	$se(\tilde{\beta})$	$\tilde{\beta}$	$se(\tilde{\beta})$	$\tilde{\beta}$	$se(\tilde{\beta})$	$\tilde{\beta}$	$se(\tilde{\beta})$	$\tilde{\beta}$	$se(\tilde{\beta})$	$\tilde{\beta}$	$se(\tilde{\beta})$
1a & 1c	-0.19	0.07	-0.18	0.07	-0.52	0.13	-0.50	0.15	-0.19	0.09	-0.19	0.09	-0.53	0.17	-0.51	0.21
2a & 2b	-0.19	0.07	-0.18	0.07	-0.52	0.13	-0.50	0.15	-0.19	0.09	-0.18	0.09	-0.52	0.17	-0.50	0.21
3a & 3b	-0.19	0.07	-0.18	0.07	-0.52	0.13	-0.50	0.15	-0.19	0.09	-0.18	0.09	-0.52	0.17	-0.50	0.21
4a & 4c	-0.19	0.07	-0.18	0.07	-0.52	0.13	-0.50	0.15	-0.19	0.07	-0.19	0.06	-0.52	0.12	-0.50	0.14
5a & 5b	-0.19	0.07	-0.19	0.07	-0.52	0.13	-0.50	0.15	-0.19	0.07	-0.19	0.06	-0.52	0.12	-0.50	0.14
6a & 6b	-0.19	0.07	-0.19	0.07	-0.52	0.13	-0.51	0.15	-0.19	0.07	-0.18	0.06	-0.52	0.12	-0.50	0.14

C.6.4 Single Imputation via Estimation

Scenario	Development								Validation							
	Binary		Continuous		Categorical (Dummy 1)		Categorical (Dummy 2)		Binary		Continuous		Categorical (Dummy 1)		Categorical (Dummy 2)	
	$\tilde{\beta}$	$se(\tilde{\beta})$	$\tilde{\beta}$	$se(\tilde{\beta})$	$\tilde{\beta}$	$se(\tilde{\beta})$	$\tilde{\beta}$	$se(\tilde{\beta})$	$\tilde{\beta}$	$se(\tilde{\beta})$	$\tilde{\beta}$	$se(\tilde{\beta})$	$\tilde{\beta}$	$se(\tilde{\beta})$	$\tilde{\beta}$	$se(\tilde{\beta})$
1a & 1c	-0.19	0.07	-0.18	0.07	-0.52	0.13	-0.50	0.15	-0.19	0.09	-0.18	0.09	-0.52	0.17	-0.51	0.21
2a & 2b	-0.19	0.07	-0.18	0.07	-0.52	0.13	-0.50	0.15	-0.19	0.09	-0.19	0.09	-0.52	0.17	-0.51	0.21
3a & 3b	-0.19	0.07	-0.18	0.07	-0.52	0.13	-0.50	0.15	-0.19	0.09	-0.18	0.09	-0.52	0.17	-0.50	0.21
4a & 4c	-0.19	0.07	-0.18	0.07	-0.52	0.13	-0.50	0.15	-0.19	0.07	-0.19	0.06	-0.53	0.12	-0.50	0.14
5a & 5b	-0.19	0.07	-0.19	0.07	-0.52	0.13	-0.50	0.15	-0.19	0.07	-0.18	0.07	-0.52	0.12	-0.50	0.14
6a & 6b	-0.19	0.07	-0.19	0.07	-0.52	0.13	-0.51	0.15	-0.19	0.07	-0.18	0.06	-0.52	0.12	-0.50	0.14

C.6.5 Hot Deck Imputation

Scenario	Development								Validation							
	Binary		Continuous		Categorical (Dummy 1)		Categorical (Dummy 2)		Binary		Continuous		Categorical (Dummy 1)		Categorical (Dummy 2)	
	$\tilde{\beta}$	$se(\tilde{\beta})$	$\tilde{\beta}$	$se(\tilde{\beta})$	$\tilde{\beta}$	$se(\tilde{\beta})$	$\tilde{\beta}$	$se(\tilde{\beta})$	$\tilde{\beta}$	$se(\tilde{\beta})$	$\tilde{\beta}$	$se(\tilde{\beta})$	$\tilde{\beta}$	$se(\tilde{\beta})$	$\tilde{\beta}$	$se(\tilde{\beta})$
1a & 1c	-0.19	0.07	-0.18	0.07	-0.52	0.13	-0.50	0.15	-0.19	0.09	-0.19	0.09	-0.53	0.17	-0.51	0.21
2a & 2b	-0.19	0.07	-0.18	0.07	-0.52	0.13	-0.50	0.15	-0.19	0.09	-0.19	0.09	-0.52	0.17	-0.50	0.21
3a & 3b	-0.19	0.07	-0.18	0.07	-0.52	0.13	-0.50	0.15	-0.19	0.09	-0.18	0.09	-0.52	0.17	-0.50	0.21
4a & 4c	-0.19	0.07	-0.18	0.07	-0.52	0.13	-0.50	0.15	-0.19	0.07	-0.19	0.06	-0.52	0.12	-0.50	0.15
5a & 5b	-0.19	0.07	-0.19	0.07	-0.52	0.13	-0.50	0.15	-0.19	0.07	-0.18	0.07	-0.52	0.12	-0.50	0.14
6a & 6b	-0.19	0.07	-0.19	0.07	-0.52	0.13	-0.51	0.15	-0.19	0.07	-0.18	0.06	-0.52	0.12	-0.50	0.14

C.6.6 Random Selection with Replacement Multiple Times

Scenario	Development								Validation							
	Binary		Continuous		Categorical (Dummy 1)		Categorical (Dummy 2)		Binary		Continuous		Categorical (Dummy 1)		Categorical (Dummy 2)	
	$\tilde{\beta}$	$se(\tilde{\beta})$	$\tilde{\beta}$	$se(\tilde{\beta})$	$\tilde{\beta}$	$se(\tilde{\beta})$	$\tilde{\beta}$	$se(\tilde{\beta})$	$\tilde{\beta}$	$se(\tilde{\beta})$	$\tilde{\beta}$	$se(\tilde{\beta})$	$\tilde{\beta}$	$se(\tilde{\beta})$	$\tilde{\beta}$	$se(\tilde{\beta})$
1a & 1c	-0.19	0.07	-0.18	0.07	-0.52	0.13	-0.50	0.15	-0.19	0.09	-0.18	0.09	-0.53	0.17	-0.50	0.21
2a & 2b	-0.19	0.07	-0.18	0.07	-0.52	0.13	-0.50	0.15	-0.19	0.09	-0.18	0.09	-0.53	0.17	-0.50	0.21
3a & 3b	-0.19	0.07	-0.18	0.07	-0.52	0.13	-0.50	0.15	-0.19	0.09	-0.18	0.09	-0.53	0.17	-0.50	0.21
4a & 4c	-0.19	0.07	-0.18	0.07	-0.52	0.13	-0.50	0.15	-0.19	0.07	-0.18	0.06	-0.53	0.12	-0.50	0.14
5a & 5b	-0.19	0.07	-0.18	0.07	-0.52	0.13	-0.50	0.15	-0.19	0.07	-0.18	0.06	-0.53	0.12	-0.50	0.14
6a & 6b	-0.19	0.07	-0.18	0.07	-0.52	0.13	-0.50	0.15	-0.19	0.07	-0.18	0.06	-0.52	0.12	-0.50	0.14

C.7 Accuracy

C.7.1 No missing data

Scenario	Binary			Continuous			Categorical (Dummy 1)			Categorical (Dummy 2)		
	Bias (%)	MSE	Coverage (%)	Bias (%)	MSE	Coverage (%)	Bias (%)	MSE	Coverage (%)	Bias (%)	MSE	Coverage (%)
1a & 1b	-0.89	0.00	86.2	-5.62	0.00	88.7	-8.24	0.01	88.2	-14.13	0.01	88.3
4a & 4b	6.24	0.01	80.8	3.08	0.01	80.1	-1.59	0.02	81.8	-2.42	0.03	83.4

C.7.2 Variable Matching

Scenario	Binary			Continuous			Categorical (Dummy 1)			Categorical (Dummy 2)		
	Bias (%)	MSE	Coverage (%)	Bias (%)	MSE	Coverage (%)	Bias (%)	MSE	Coverage (%)	Bias (%)	MSE	Coverage (%)
1a & 1c	-30.30	-0.02	91.6	68.91	0.04	92.1	NA					
2a & 2b	-2.36	0.00	88.1	NA			-3.68	0.02	88.5	-7.50	0.02	88.5
3a & 3b	NA			-1.40	0.00	89.8	-2.71	0.02	89.6	-2.00	0.03	87.8
4a & 4c	-33.85	-0.02	89.5	58.59	0.04	88.4	NA					
5a & 5b	-0.88	0.00	81.7	NA			-6.24	0.01	81.9	-5.89	0.02	82.6
6a & 6b	NA			6.77	0.01	82.3	-0.92	0.02	82.8	-0.83	0.03	83.6

C.7.3 Random Selection with Replacement

Scenario	Binary			Continuous			Categorical (Dummy 1)			Categorical (Dummy 2)		
	Bias (%)	MSE	Coverage (%)	Bias (%)	MSE	Coverage (%)	Bias (%)	MSE	Coverage (%)	Bias (%)	MSE	Coverage (%)
1a & 1c	2.05	0.01	86.1	-4.92	0.00	88.9	-7.72	0.01	89.7	-10.24	0.02	88.9
2a & 2b	-3.95	0.00	88.2	-0.23	0.00	87.4	-3.05	0.02	87.9	-7.32	0.02	88.4
3a & 3b	1.10	0.01	89.0	-1.67	0.00	89.8	-3.07	0.02	89.5	-2.79	0.03	87.8
4a & 4c	0.14	0.01	82.4	-0.85	0.00	81.9	-6.31	0.01	81.6	-7.09	0.02	80.1
5a & 5b	-1.41	0.00	82.0	1.11	0.00	82.3	-5.95	0.01	80.9	-5.82	0.02	82.4
6a & 6b	-6.91	0.00	80.9	7.07	0.01	83.0	-0.39	0.02	83.4	-0.15	0.03	83.9

C.7.4 Single Imputation via Estimation

Scenario	Binary			Continuous			Categorical (Dummy 1)			Categorical (Dummy 2)		
	Bias (%)	MSE	Coverage (%)	Bias (%)	MSE	Coverage (%)	Bias (%)	MSE	Coverage (%)	Bias (%)	MSE	Coverage (%)
1a & 1c	-0.41	0.00	86.9	-3.62	0.00	89.2	-2.35	0.02	88.1	-7.18	0.02	87.1
2a & 2b	-2.25	0.00	88.2	-8.09	0.00	88.7	-3.49	0.02	87.6	-8.00	0.02	87.9
3a & 3b	-5.15	0.00	88.4	-1.70	0.00	89.5	-2.23	0.02	89.3	-3.68	0.03	87.4
4a & 4c	0.05	0.01	82.6	-1.03	0.00	82.9	-8.31	0.01	81.1	-6.18	0.02	83.4
5a & 5b	-1.32	0.00	81.5	3.58	0.01	83.1	-6.28	0.01	81.4	-5.14	0.02	83.1
6a & 6b	-3.10	0.00	81.2	7.67	0.01	82.6	-1.29	0.02	83.2	-0.93	0.03	83.4

C.7.5 Hot Deck Imputation

Scenario	Binary			Continuous			Categorical (Dummy 1)			Categorical (Dummy 2)		
	Bias (%)	MSE	Coverage (%)	Bias (%)	MSE	Coverage (%)	Bias (%)	MSE	Coverage (%)	Bias (%)	MSE	Coverage (%)
1a & 1c	-0.23	0.00	87.0	-5.83	0.00	89.4	-10.75	0.01	87.4	-8.79	0.02	87.9
2a & 2b	-2.93	0.00	87.3	-6.90	0.00	89.2	-4.04	0.02	88.1	-7.56	0.02	88.5
3a & 3b	-1.65	0.00	88.2	-1.00	0.00	90.1	-2.24	0.02	89.6	-2.43	0.03	87.9
4a & 4c	-0.49	0.00	81.9	-0.24	0.00	81.8	-4.17	0.02	81.3	-4.44	0.02	81.6
5a & 5b	-1.70	0.00	82.0	3.10	0.01	83.1	-6.55	0.01	81.9	-5.05	0.02	83.0
6a & 6b	-6.41	0.00	81.7	6.38	0.01	83.2	-0.94	0.02	83.0	-0.28	0.03	83.6

C.7.6 Random Selection with Replacement Multiple Times

Scenario	Binary			Continuous			Categorical (Dummy 1)			Categorical (Dummy 2)		
	Bias (%)	MSE	Coverage (%)	Bias (%)	MSE	Coverage (%)	Bias (%)	MSE	Coverage (%)	Bias (%)	MSE	Coverage (%)
1a & 1c	-1.14	0.00	100	-0.49	0.00	100	-3.97	0.01	100	-3.00	0.02	100
2a & 2b	0.30	0.01	100	1.45	0.01	100	-2.25	0.01	100	-2.90	0.02	100
3a & 3b	-2.22	0.00	100	-1.05	0.00	100	-2.41	0.01	100	-3.08	0.02	100
4a & 4c	-0.31	0.00	100	0.11	0.00	100	-0.44	0.02	100	-0.59	0.02	100
5a & 5b	-0.10	0.01	100	0.17	0.00	100	-0.29	0.02	100	-0.58	0.02	100
6a & 6b	0.07	0.01	100	-0.25	0.00	100	0.71	0.02	100	-0.19	0.02	100

Appendix D – Further Results from Chapter 9

Table 54: Seizure recurrence risk in the next 12 months estimated from multivariable model at specific time points of seizure freedom. Estimates presented assume seizures not confined to sleep and no first degree relative with epilepsy. MESS (EEG/CT as 2 categories) & FIRST (random selection)

				MESS				FIRST			
Variable				Immediate Treatment		Delayed Treatment					
Aetiology	EEG	CT/MRI scan results	Months from index seizure	Risk of seizure within next 12 months (%, 95% CI)	Months from index seizure until annual risk falls below 20%	Risk of seizure within next 12 months (%, 95% CI)	Months from index seizure until annual risk falls below 20%	Risk of seizure within next 12 months (%, 95% CI)	Months from index seizure until annual risk falls below 20%	Risk of seizure within next 12 months (%, 95% CI)	Months from index seizure until annual risk falls below 20%
Not remote symptomatic	Normal	Normal	6 12	13 (10 to 16) 7 (4 to 10)	1.2	16 (13 to 19) 9 (6 to 12)	3.2	7 (2, 11) 4 (0, 9)	0.0	14 (9, 18) 10 (5, 14)	0.6
Not remote symptomatic	Abnormal	Normal	6 12	19 (16 to 22) 11 (8 to 13)	5.3	23 (20 to 26) 13 (10 to 16)	6.8	10 (5, 14) 7 (2, 11)	0.0	20 (16, 24) 14 (10, 18)	4.5
Not remote symptomatic	Normal	Abnormal	6 12	13 (10 to 17) 7 (5 to 10)	1.2	16 (13 to 19) 9 (6 to 12)	3.3	8 (3, 13) 5 (1, 10)	0.0	16 (12, 21) 11 (7, 15)	2.2
Not remote symptomatic	Abnormal	Abnormal	6 12	19 (16 to 22) 11 (8 to 14)	5.5	23 (20 to 26) 13 (11 to 16)	6.9	11 (7, 16) 8 (4, 12)	0.0	23 (20, 27) 17 (13, 20)	7.4
Remote symptomatic	Normal	Normal	6 12	18 (14 to 21) 10 (7 to 13)	4.0	21 (18 to 24) 12 (9 to 15)	6.1	10 (5, 15) 7 (3, 11)	0.0	21 (17, 25) 15 (11, 19)	6.1
Remote symptomatic	Abnormal	Normal	6 12	25 (22 to 28) 14 (12 to 17)	8.7	30 (27 to 33) 17 (15 to 20)	10.7	15 (10, 19) 10 (6, 14)	0.8	30 (26, 33) 21 (18, 25)	11.0
Remote symptomatic	Normal	Abnormal	6 12	18 (15 to 21) 10 (7 to 13)	4.0	21 (18 to 24) 12 (9 to 15)	6.3	12 (7, 17) 8 (4, 12)	0.0	25 (21, 29) 17 (14, 21)	9.9
Remote symptomatic	Abnormal	Abnormal	6 12	26 (23 to 28) 15 (12 to 17)	8.7	30 (28 to 33) 18 (15 to 20)	10.7	18 (13, 22) 12 (8, 16)	2.5	35 (31, 38) 25 (21, 28)	16.0

Table 55: Seizure recurrence risk in the next 12 months estimated from multivariable model at specific time points of seizure freedom. Estimates presented assume seizures not confined to sleep and no first degree relative with epilepsy. FIRST (single imputation)

Variable				Immediate Treatment		Delayed Treatment	
Aetiology	**EEG**	**CT/MRI**	**Months from index seizure**	**Risk of seizure within next 12 months (%, 95% CI)**	**Months from index seizure until annual risk falls below 20%**	**Risk of seizure within next 12 months (%, 95% CI)**	**Months from index seizure until annual risk falls below 20%**
Not remote symptomatic	Normal	Normal	6 12	6 (2, 11) 4 (0, 9)	0.0	14 (9, 18) 10 (5, 14)	0.6
Not remote symptomatic	Abnormal	Normal	6 12	10 (5, 14) 7 (2, 11)	0.0	20 (16, 24) 14 (10, 18)	4.5
Not remote symptomatic	Normal	Abnormal	6 12	8 (3, 13) 5 (1, 10)	0.0	16 (12, 21) 11 (7, 15)	2.2
Not remote symptomatic	Abnormal	Abnormal	6 12	11 (7, 16) 8 (4, 12)	0.0	23 (19, 27) 16 (13, 20)	7.4
Remote symptomatic	Normal	Normal	6 12	10 (6, 15) 7 (3, 11)	0.0	21 (17, 25) 15 (11, 19)	6.1
Remote symptomatic	Abnormal	Normal	6 12	15 (11, 19) 10 (6, 14)	0.8	30 (27, 34) 22 (18, 25)	12.4
Remote symptomatic	Normal	Abnormal	6 12	12 (8, 17) 8 (4, 13)	0.0	25 (21, 29) 18 (14, 21)	9.9
Remote symptomatic	Abnormal	Abnormal	6 12	18 (13, 22) 12 (8, 16)	2.5	35 (32, 38) 25 (22, 29)	16.0

Table 56: Seizure recurrence risk in the next 12 months estimated from multivariable model at specific time points of seizure freedom. Estimates presented assume seizures not confined to sleep and no first degree relative with epilepsy. FIRST (hot deck imputation)

Variable				Immediate Treatment		Delayed Treatment	
Aetiology	EEG	CT/MRI	Months from index seizure	Risk of seizure within next 12 months (%, 95% CI)	Months from index seizure until annual risk falls below 20%	Risk of seizure within next 12 months (%, 95% CI)	Months from index seizure until annual risk falls below 20%
Not remote symptomatic	Normal	Normal	6 12	7 (2, 12) 5 (1, 9)	0.0	14 (10, 19) 10 (6, 14)	0.6
Not remote symptomatic	Abnormal	Normal	6 12	11 (6, 15) 8 (4, 12)	0.0	22 (18, 26) 16 (12, 20)	7.1
Not remote symptomatic	Normal	Abnormal	6 12	7 (3, 12) 5 (1, 9)	0.0	15 (11, 20) 11 (7, 15)	1.6
Not remote symptomatic	Abnormal	Abnormal	6 12	12 (7, 16) 8 (4, 12)	0.0	24 (20, 28) 17 (14, 21)	9.9
Remote symptomatic	Normal	Normal	6 12	11 (6, 16) 8 (4, 12)	0.0	23 (19, 27) 16 (13, 20)	7.1
Remote symptomatic	Abnormal	Normal	6 12	18 (13, 22) 13 (9, 16)	2.5	34 (31, 38) 25 (22, 29)	16.8
Remote symptomatic	Normal	Abnormal	6 12	12 (8, 17) 9 (5, 13)	0.0	25 (21, 29) 18 (14, 22)	10.2
Remote symptomatic	Abnormal	Abnormal	6 12	19 (15, 23) 14 (10, 18)	4.0	37 (34, 40) 28 (24, 31)	16.8

Table 57: Seizure recurrence risk in the next 12 months estimated from multivariable model at specific time points of seizure freedom. Estimates presented assume seizures not confined to sleep and no first degree relative with epilepsy. FIRST (multiple imputation)

Variable				Immediate Treatment		Delayed Treatment	
Aetiology	**EEG**	**CT/MRI**	**Months from index seizure**	**Risk of seizure within next 12 months (%, 95% CI)**	**Months from index seizure until annual risk falls below 20%**	**Risk of seizure within next 12 months (%, 95% CI)**	**Months from index seizure until annual risk falls below 20%**
Not remote symptomatic	Normal	Normal	6 12	6 (2, 11) 4 (0, 9)	0.0	14 (9, 18) 10 (5, 14)	0.5
Not remote symptomatic	Abnormal	Normal	6 12	10 (5, 14) 7 (2, 11)	0.0	20 (16, 24) 14 (10, 18)	4.8
Not remote symptomatic	Normal	Abnormal	6 12	8 (3, 12) 5 (1, 10)	0.0	16 (12, 20) 11 (7, 15)	2.0
Not remote symptomatic	Abnormal	Abnormal	6 12	11 (7, 16) 8 (4, 12)	0.0	23 (19, 27) 16 (13, 20)	7.4
Remote symptomatic	Normal	Normal	6 12	10 (6, 15) 7 (3, 11)	0.0	21 (17, 25) 15 (11, 19)	6.6
Remote symptomatic	Abnormal	Normal	6 12	15 (11, 19) 10 (6, 14)	1.1	30 (27, 34) 22 (18, 25)	12.1
Remote symptomatic	Normal	Abnormal	6 12	12 (7, 17) 8 (4, 13)	0.0	25 (21, 29) 17 (14, 21)	9.9
Remote symptomatic	Abnormal	Abnormal	6 12	18 (13, 22) 12 (8, 16)	2.6	35 (31, 38) 25 (22, 29)	15.7

Appendix E – Published Papers

Work from Chapters 4, 5 and 6 has been published. Full references for the relevant articles are shown and copies included.

Chapter 4

Prognostic factors for time to treatment failure and time to 12 months of remission for patients with focal epilepsy: post-hoc, subgroup analyses of data from the SANAD trial

Laura Bonnett[1]; Catrin Tudur Smith[1]; David Smith[2]; Paula Williamson[1]; David Chadwick[2]; Anthony Marson[3]

[1]Department of Biostatistics, University of Liverpool, U.K.; [2]The Walton Centre Foundation NHS Trust, Liverpool, U.K.; [3]Department of Molecular and Clinical Pharmacology, University of Liverpool, U.K.

Lancet Neurology, 2012, 11(4), pp. 331-340

Chapter 5

Risk of recurrence after a first seizure and implications for driving: further analysis of the Multicentre study of early Epilepsy and Single Seizures

Laura Bonnett[1]; Catrin Tudur Smith[1]; Paula Williamson[1]; Anthony Marson[2]

[1]Department of Biostatistics, University of Liverpool, U.K.;
[2]Department of Molecular and Clinical Pharmacology, University of Liverpool, U.K.

British Medical Journal, 2010, 341, c6477

Chapter 6

Seizure recurrence after antiepileptic drug withdrawal and the implications for driving: further results from the MRC Antiepileptic Drug Withdrawal Study and a systematic review

Laura Bonnett[1]; Arif Shukralla[2]; Catrin Tudur Smith[1]; Paula Williamson[1]; Anthony Marson[3]

[1]Department of Biostatistics, University of Liverpool, U.K.; [2]Clinical Sciences Centre, University of Liverpool, U.K.; [3]Department of Molecular and Clinical Pharmacology, University of Liverpool, U.K.

Journal of Neurology, Neurosurgery and Psychiatry, 2011, 82(12), pp. 1328-1333

This text box is where the unabridged thesis included the following third party copyrighted material;

Bonnett LJ, Tudur-Smith C, Smith D, Williamson PR, Chadwick D, Marson AG. (2012) “Prognostic factors for time to treatment failure and time to 12 months of remission for patients with focal epilepsy: post-hoc, subgroup analyses of data from the SANAD trial.” Lancet Neurology **11**(4):331-40.

doi: http://dx.doi.org/10.1016/S1474-4422(12)70018-2

This text box is where the unabridged thesis included the following third party copyrighted material;

Bonnett LJ, Tudur-Smith C, Williamson PR, Marson AG. (2010) “Risk of recurrence after a first seizure and implications for driving: further analysis of the Multicentre study of early Epilepsy and Single Seizures.” British Medical Journal. 341:c6477.

doi: http://dx.doi.org/10.1136/bmj.c6477

This text box is where the unabridged thesis included the following third party copyrighted material;

Bonnett LJ, Shukralla A, Tudur-Smith C, Williamson PR, Marson AG. (2011) “Seizure recurrence after antiepileptic drug withdrawal and the implications for driving: further results from the MRC Antiepileptic Drug Withdrawal Study and a systematic review.” Journal of Neurology Neurosurgery and Psychiatry **82**(12):1328-33.

doi: http://dx.doi.org/ 10.1136/jnnp.2010.222885