Individual Patient Data Meta-analysis with Time-to-Event Outcomes

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

by

Catrin Tudur Smith

June 2004

To

Shaun and Dad

Acknowledgements

I am extremely grateful to my supervisors Paula Williamson and Tony Marson for their help, enthusiasm and encouragement throughout the course of this project. In particular, I would like to thank Paula for her excellent supervision, inspirational discussions and the commitment she has given me. I am particularly grateful to Tony for providing valuable clinical guidance and help with the epilepsy data which motivated many of the methodological questions. I am grateful to Professor David Chadwick and the Epilepsy Monotherapy Trialists' Group for making their individual patient data available. Finally, I would like to thank Professor Takuhiro Yamaguchi for providing his SAS IML program for fitting a Cox model with random trial and treatment effects which was extended in this thesis.

From a personal point of view, completing the project was made a great deal easier and more enjoyable with the support, love and encouragement off my family and husband Shaun. Thanks to you all.

Abstract

Catrin Tudur Smith

Individual Patient Data Meta-analysis with Time-to-Event Outcomes

Aims

This thesis concerns the meta-analysis of time-to-event data, investigating methodology for both aggregate and individual patient data, and comparing the two approaches.

Methods

Meta-analysis may be based on either aggregate data or individual patient data collected from the original authors of each trial. An extension to a current aggregate data based approach is reviewed and methods for assessing the proportional hazards assumption proposed. Comparisons of treatment effect estimates from aggregate and individual patient data based meta-analyses are summarised. Three approaches to meta-analysis with individual patient data are reviewed and contrasted with each other theoretically and with simulated data. Models for investigating heterogeneity with individual patient data on time-to-event outcomes are reviewed and extended to incorporate random effects. A comparison of methods for exploring heterogeneity with aggregate or individual patient data is undertaken. Methodology for indirect comparisons are explored and extended to facilitate the analysis of totality of evidence. Several individual patient data based systematic reviews comparing anti-epileptic drugs with respect to time-to-event outcomes are used throughout to illustrate methods and motivate further research questions.

Results

An aggregate data approach is found to be difficult to apply in practice when time-to-event data are considered. One particular method based on Kaplan-Meier curves is investigated empirically and found to be unreliable. Comparisons of treatment effect estimates from aggregate and individual patient data based meta-analyses indicate that results obtained from the two analyses differ, but in no consistent direction across reviews. For analyses with individual patient data, simulated data indicate that the stratified Cox model and inverse variance weighted average of trial level Cox model estimates perform favourably and are to be preferred to the stratified log-rank analysis when the underlying treatment effect is large, hazards are proportional and there is no heterogeneity in effects across trials. For smaller treatment effects, all three methods perform well but further investigation is required. The methodology and facility to fit random effects Cox regression models are presented. Simulated data indicate that the estimate of treatment effect is more likely to be biased when there is a greater degree of heterogeneity particularly for treatment effects close to the null. The stratified Cox model with random treatment effects is found to be the least computer intensive random effects model making this an attractive approach. Empirical results of investigating heterogeneity are compared between models based on aggregate data and individual patient data with a more thorough explanation of heterogeneity obtained from the latter model. Using a totality of evidence approach, important clinical results are obtained for comparisons of anti-epileptic drugs that have not previously been undertaken within an RCT setting, and precision improved for those that have.

Conclusions

Individual patient data should be used whenever possible to reliably study patient characteristics and investigate heterogeneity in meta-analysis with time-to-event outcomes. The approach presented for a totality of evidence analysis, incorporating covariate main and interaction effects, highlights a further advantage of individual patient data.

CONTENTS

CHAPTER 1 Introduction **1**

1.1 Systematic Reviews 1
1.2 Meta-analysis 2
1.2.1 Measures of treatment effect within individual trials and meta-analysis 3
1.2.2 Pooling results 5
1.3 Heterogeneity 7
1.3.1 Detecting and quantifying heterogeneity 8
1.3.2 Investigating heterogeneity 9
1.4 Publication Bias 10
1.5 Discussion and structure of the thesis 11

CHAPTER 2 Aggregate data meta-analysis with time-to-event outcomes **13**

2.1 Time-to-event data 13
2.1.1 Survivor function and hazard function 14
2.2 Methods of analysis of time-to-event data 15
2.3 Estimating the log hazard ratio and its variance using aggregate data 17
2.3.1 Direct method 18
2.3.2 Indirect method 1: Estimating var(log(HR_j)) from a confidence interval 18
2.3.3 Indirect method 2: Estimating log(HR_j) and its variance using the quoted p-value of the log-rank test 19

2.3.4 Indirect method 3a: Estimating log(HR_j) and its variance from survival curves 20
2.3.5 Indirect method 3b: Estimating log(HR_j) and its variance from survival curves and numbers at risk 21
2.4 Comparison of methods based on survival curves 24
2.4.1 Example 1 24
2.4.2 Example 2 28
2.5 Reliability of methods for aggregate data meta-analysis 31
2.5.1 Examples 32
2.5.2 Data extraction 33
2.5.3 Estimation of summary statistics 36
2.5.4 Meta-analysis 39
2.6 Assessing the assumption of proportional hazards 45
2.6.1 Overall log(HR) estimate only available for each study 46
2.6.2 Log cumulative hazard plot (log-log plot) 46
2.6.3 Estimate of log(HR) available for different time intervals 50
2.7 Discussion 54

CHAPTER 3 Individual patient data meta-analysis 57

3.1 Comparison of IPD and AD based meta-analysis 59
3.2 Log-rank analysis 62
3.3 The Cox proportional hazards model 64
3.3.1 Cox regression model with trial indicator variables and fixed treatment effect 66
3.3.2 Cox regression model stratified by trial with fixed treatment effect 67
3.4 Inverse variance weighted average 67
3.5 Comparison of methods 68
3.5.1 Log-rank analysis versus Cox regression model 68
3.5.2 Stratified log-rank analysis versus an IV weighted average 70
3.5.3 Stratified Cox regression model versus an IV weighted average 71
3.6 Simulation study 76
3.7 Discussion 85

CHAPTER 4 Monotherapy drugs for Epilepsy: Meta-analyses based on individual patient data 87

4.1 Introduction 87
4.2 Epilepsy data 89
4.3 Meta-analyses of epilepsy data 100

4.4 Comparison of stratified log-rank analysis with Cox model 112
4.5 Missing Data 114
4.6 Discussion 115

CHAPTER 5 Modelling heterogeneity 118

5.1 Exploring heterogeneity with individual patient data using the Cox model 121
5.1.1 Fixed trial and treatment effect (FE/FE) 121
5.1.2 Stratified by trial with fixed treatment effect (SFE/FE) 122
5.1.3 Fixed trial effect and random treatment effects (FE/RE) 122
5.1.4 Stratified by trial with random treatment effects (SFE/RE) 123
5.1.5 Random trial effects and random treatment effects (RE/RE) 124
5.2 Parameter estimation 125
5.2.1 Software Development 128
5.3 Treatment coding 130
5.4 Handling Ties 132
5.5 Example: CBZ-VPS monotherapy trials using IPD 134
5.6 Simulation study 139
5.7 Exploring heterogeneity with aggregate data 144
5.7.1 Weighted regression with no allowance for residual heterogeneity 144
5.7.2 Weighted regression incorporating residual heterogeneity 145
5.8 Example: CBZ-VPS monotherapy trials using AD 145
5.9 Exploring heterogeneity with IPD or AD 150
5.10 Discussion 156

CHAPTER 6 External evidence and indirect comparisons 160

6.1 Introduction 160
6.2 Overview of methods for indirect comparisons 162
6.3 Indirect aggregate data comparisons with time-to-event data 165
6.4 Indirect individual patient data comparisons with time-to-event data 168
6.5 Assumption of no interaction between treatment and covariates 170
6.6 Combining indirect and direct evidence 173
6.6.1 Aggregate data 174
6.6.2 Individual patient data 174
6.7 Incorporating trials with more than two treatments 176
6.7.1 Aggregate data 176

6.7.2 Individual patient data 177
6.8 Illustration of methods 177
6.8.1 Indirect evidence from trials with two treatments 177
6.8.2 Combining direct and indirect evidence 184
6.8.3 Including trials with three treatments 187
6.8.4 Summary of illustration 188
6.9 Totality of evidence in epilepsy trials 189
6.9.1 Motivation for exploring totality of evidence 192
6.9.2 Exploring treatment main effects using totality of evidence 194
6.9.3 Exploring the interaction between treatment and epilepsy type using totality of evidence for time to first seizure 203
6.10 Discussion 210

CHAPTER 7 Concluding remarks and further work 214

7.1 Methodological conclusions 214
7.2 Aggregate data compared to individual patient data 220
7.3 Areas for further research 222

Appendix A SAS programs for fitting random effect Cox models using the Breslow method for handling ties 224
A.1 Cox model with fixed trial indicator variables and random treatment effects (FE/RE) using Breslow method for handling ties 224
A.2 Cox model stratified by trial with random treatment effects (SFE/RE) using Breslow method for handling ties 228
A.3 Cox model with random trial effects and random treatment effects (RE/RE) using Breslow method for handling ties 236

Appendix B SAS programs for fitting random effect Cox models using the Efron method for handling ties 240
B.1 Cox model with fixed trial indicator variables and random treatment effects (FE/RE) using Efron method for handling ties 240
B.2 Cox model stratified by trial with random treatment effects (SFE/RE) using Efron method for handling ties 246
B.3 Cox model with random trial effects and random treatment effects (RE/RE) using Efron method for handling ties 260

Appendix C Kaplan-Meier survival curves for CBZ-VPS analyses examined in Chapter 5 267
C.1 Time to 12 month remission 267
C.2 Time to first seizure 270

Appendix D **Calculations for indirect comparisons relating to Chapter 6** **273**
D.1 Calculating indirect comparison for PHT:CBZ 273
D.2 Calculating indirect comparison for VPS:CBZ 275

Appendix E **Parameter estimates for totality of evidence relating to Chapter 6** **277**
E.1. Totality of evidence model exploring treatment main effects 277
E.2. Totality of evidence model exploring effect of epilepsy type 279

BIBLIOGRAPHY **284**

Figures

Figure 2.1. Example 1: Kaplan–Meier estimates of time to first seizure of 466 patients with epilepsy treated with sodium valproate (VPS) or carbamazepine (CBZ) in a single randomized controlled trial. 26

Figure 2.2. Example 1: Comparing actual and estimated numbers at risk for CBZ group (similar pattern for VPS group). 28

Figure 2.3. Example 2: Comparing actual and estimated numbers at risk for males (similar pattern for females). 31

Figure 2.4. Example 3: Meta-analysis of unadjusted results 40

Figure 2.5. Example 3: Funnel plot using unadjusted estimates 41

Figure 2.6. Example 4: Meta-analysis using AD estimates only 42

Figure 2.7. Example 4: Meta-analysis using IPD estimates only 43

Figure 2.8. Example 4: Meta-analysis using IPD and AD estimates 43

Figure 2.9. Example 4: Funnel plot using IPD and AD unadjusted estimates 44

Figure 2.10. Log cumulative hazard plots using individual patient data (IPD) for five trials included in CBZ/VPS systematic review. 48

Figure 2.11. Log cumulative hazard plots using aggregate data (AD) for five trials included in CBZ/VPS systematic review. 49

Figure 2.12. Hazard Ratio pooled across trials within each interval (90-day) plotted against time for 5 trials in CBZ/VPS systematic review 53

Figure 3.1. Mean overall log hazard ratio estimated in 100 repetitions of 5 trial meta-analyses with underlying heterogeneity parameter $\tau^2 = 0$. 82

Figure 3.2. Mean overall log hazard ratio estimated in 100 repetitions of 5 trial meta-analyses with underlying heterogeneity parameter $\tau^2 = 0.1$. 82

Figure 3.3. Mean overall log hazard ratio estimated in 100 repetitions of 5 trial meta-analyses with underlying heterogeneity parameter $\tau^2 = 0.5$. 82

Figure 3.4. Mean overall log hazard ratio estimated in 100 repetitions of 5 trial meta-analyses with underlying heterogeneity parameter $\tau^2 = 0.9$. 83

Figure 3.5. Coverage (percentage of 95% confidence intervals for the log hazard ratio that contain the true value) over 100 repetitions of 5 trial meta-analyses with underlying heterogeneity parameter $\tau^2 = 0$. 83

Figure 3.6. Coverage (percentage of 95% confidence intervals for the log hazard ratio that contain the true value) over 100 repetitions of 5 trial meta-analyses with underlying heterogeneity parameter τ^2=0.1. 83

Figure 3.7. Coverage (percentage of 95% confidence intervals for the log hazard ratio that contain the true value) over 100 repetitions of 5 trial meta-analyses with underlying heterogeneity parameter τ^2=0.5. 84

Figure 3.8. Coverage (percentage of 95% confidence intervals for the log hazard ratio that contain the true value) over 100 repetitions of 5 trial meta-analyses with underlying heterogeneity parameter τ^2=0.9. 84

Figure 4.1. Graphical display of meta-analysis comparing CBZ to VPS 104

Figure 4.2. Graphical display of meta-analysis comparing PHT to VPS 105

Figure 4.3. Graphical display of meta-analysis comparing CBZ to PHB 106

Figure 4.4. Graphical display of meta-analysis comparing CBZ to PHT 107

Figure 4.5. Graphical display of meta-analysis comparing PHB to PHT 108

Figure 4.6. Graphical display of meta-analysis comparing PHB to VPS 109

Figure 4.7. Graphical display of meta-analysis comparing CBZ to LTG 110

Figure 4.8. Graphical display of meta-analysis comparing PHT to OXC 111

Figure 6.1. Example 1: Illustration of treatment-covariate interaction with indirect comparison where trial settings involve different covariate values 172

Figure 6.2. Example 2: Illustration of treatment-covariate interaction with indirect comparison where trial settings involve similar covariate values 173

Figure 6.3. Relationship between AEDs from each direct comparison (maximum number of trials and patients available for analysis for each comparison) 189

Tables

Table 2.1. Example 1: Comparison of estimates of overall log(HR) and SE(log(HR)) (pooled across intervals) 27

Table 2.2. Example 2: Comparison of estimates of overall log(HR) and SE(log(HR)) (pooled across intervals) 30

Table 2.3. Example 3: Summary of information available in each trial. 34

Table 2.4. Example 4: Summary of information available in each trial. 35

Table 2.5. Example 3: estimates of log(HR_j) and corresponding standard error 37

Table 2.6. Example 4: estimates of log(HR_j) and corresponding standard error 38

Table 3.1. Empirical evidence of the comparison between meta-analysis methods. Table entries relate to the pooled treatment effect and 95% confidence interval. HR=Hazard Ratio, sdiff=difference in survival probabilities at 30 months, OR=odds ratio, RR=relative risk 60

Table 3.2. Mean log hazard ratio (standard deviation) and coverage for parameter combinations in 100 simulated meta-analyses of 5 trials. 81

Table 4.1. Systematic reviews and availability of outcome data for monotherapy comparisons VPS: Valproate, CBZ: Carbamazepine, PHT: phenytoin, PHB: phenobarbitone, LTG: Lamotrigine, OXC: Oxcarbazepine. 94

Table 4.2. Characteristics of trials included in eight IPD systematic reviews of anti-epileptic drugs 96

Table 4.3. Availability of patient covariate data across trials 98

Table 4.4. Pooled hazard ratio and 95% CI from stratified log-rank analysis, and test of homogeneity of treatment effect 103

Table 4.5. Pooled hazard ratio and 95% CI from Cox proportional hazards model stratified by trial (using Efron method for handling ties) 113

Table 5.1. Parameter estimates (standard error) for time to 12 month remission and time to first seizure obtained using alternative hierarchical formulations of the Cox regression model (5 trials, 1225 individuals) 136

Table 5.2. Parameter estimates (standard errors) for 'time to 12 month remission' using alternative hierarchical formulations of the Cox regression model (using Efron's approximation) allowing for patient-level covariates (5 trials, 1183 individuals) 137

Table 5.3. Parameter estimates (standard error) for time to 12 month remission fitted to a subset of data (5 trials, 750 events, 1183 individuals) 139

Table 5.4. Mean (standard deviation) of parameter estimates from 100 simulated meta-analyses of 5 trials (40 individuals per group) 142

Table 5.5. CBZ-VPS example: Aggregate data generated from IPD for each trial 146

Table 5.6. Parameter estimates (SE) from univariate meta-regression models using AD 147

Table 5.7. Parameter estimates (SE) from Cox proportional hazards models with main effect of treatment, covariate and corresponding interaction term using IPD 153

Table 5.8. Sensitivity analysis: Parameter estimates (SE) from Cox proportional hazards models with main effect of treatment, covariate and corresponding interaction term using IPD using subset of data for 1183 individuals 155

Table 6.1. Pair-wise comparisons from separate stratified Cox models with fixed or random treatment effects 178

Table 6.2. Direct pair-wise comparison between VPS and PHT from a Cox model stratified by trial with fixed or random treatment effects 181

Table 6.3. Estimates for each indirect comparison using AD or IPD with fixed or random effects 183

Table 6.4. Combining direct and indirect evidence from 2-arm trials 186

Table 6.5. Results from a single fixed effect stratified Cox regression model including all data from 3-arm trials and 2-arm trials 188

Table 6.6 Number of patients randomised to each drug, pair-wise comparisons examined and IPD availability for each outcome across each trial 190

Table 6.7. Time to first seizure: Analysis of totality of evidence and direct evidence for comparison 197

Table 6.8. Time to 12 month remission: Analysis of totality of evidence and direct evidence for comparison 198

Table 6.9. Time to withdrawal: Analysis of totality of evidence and direct evidence for comparison 199

Table 6.10. Analysis of totality of evidence for time to first seizure adjusted for epilepsy type and type by treatment interaction terms 207

CHAPTER 1

Introduction

1.1. Systematic reviews

Evidence-based medicine is the conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients [1]. Systematic reviews of research studies are a very useful tool increasingly used in health care related areas to improve the evidence base. In particular, a systematic review including two or more randomised controlled trials is considered by many to be the 'best possible' source of evidence about the effects of treatments. A systematic review is an overview of primary studies using explicit, systematic and reproducible methods which aim to limit bias and random error. The ability to condense large amounts of information into more manageable and informative parts is of particular benefit to health care professionals endeavouring to keep up to date with new developments in evidence-based medicine.

The general approach for preparing a systematic review is to formulate the question of interest, develop and follow a strategy for locating and selecting studies, assess the quality of each study believed to be suitable, collect or extract relevant data from each study, summarise the results from each study either qualitatively or quantitatively, interpret, summarise and disseminate findings. A protocol should be developed early in the review process to outline review objectives, questions of interest, proposed review

methods for identifying and selecting studies and steps for collecting and analysing the data. Development of a protocol is important to avoid bias being introduced by decisions that are influenced by the data [2].

The Cochrane Collaboration was established in 1993 to assist researchers and health care professionals in making well informed decisions in clinical practice by preparing, maintaining and promoting the accessibility of systematic reviews of the effects of healthcare interventions. The Cochrane Collaboration consists of approximately 50 'Collaborative Review Groups', each interested in a particular area of health care and each responsible for preparing and maintaining systematic reviews in that area. The Cochrane Library [3] is an electronic library produced and distributed quarterly including a database of systematic reviews prepared by the Cochrane Collaboration, other relevant databases and other sources of useful information. Systematic reviews undertaken by the Cochrane Collaboration are extremely valuable as they are regularly updated to incorporate new evidence.

A systematic review may be undertaken to provide a summary of the available evidence for any type of study design. This thesis will only address issues related to systematic reviews of randomised controlled trials (RCTs) as this individual study design will generally provide better evidence for intervention effects compared to other study designs since the process of randomisation ensures comparability of patients in terms of measurable and un-measurable factors.

1.2. Meta-analysis

Meta-analysis is the statistical procedure used to quantitatively summarise the results of several studies that have been designed to answer similar clinical questions. As the precision of a treatment effect estimate largely depends on the number of patients, a meta-analysis that draws on patients in many studies will have more power to detect small but clinically important results compared with any individual study identified as eligible. In general, a meta-analysis will consist of calculating a pooled treatment effect estimate, which is basically a weighted average of individual trial results, and appropriate confidence interval.

There are numerous ways in which bias can be introduced in reviews and meta-analyses of controlled clinical trials [2]. The quality of individual trials included in the meta-analysis is extremely important and an adequate assessment of different quality components should be undertaken prior to meta-analysis. Potential biases relate to systematic differences in patients characteristics at baseline (selection bias), unequal provision of care apart from the treatment under evaluation (performance bias), biased assessment of outcomes (detection bias), and bias due to exclusion of patients after they have been allocated to treatment groups (attrition bias) [2]. These biases threaten the validity of individual trials and hence meta-analysis. Other examples of types of bias relate to reporting biases such as publication bias and language bias. Egger *et al* [2] describe in-depth details of several such biases which reviewers and researchers should be aware of prior to undertaking or interpreting any systematic review and meta-analysis. The process should therefore involve an adequate investigation into potential sources of bias and an assessment of how robust the results may be to modifications in any assumptions made.

If a meta-analysis is considered appropriate within a systematic review, an estimate of treatment effect, for example an odds ratio for binary data, and corresponding measure of variation from each study are usually required. This information will sometimes be provided directly in the trial report, but more often, summary data for each treatment group will be extracted and used to calculate the relative treatment effect estimate and measure of variation for each study. This summary information relating to the relative treatment effect is termed aggregate data (AD) from here on. Although this level of information will be adequate in many situations, in others it may be necessary to collect individual patient data (IPD) from each study in order to calculate the estimates of interest. Further discussion about the advantages and disadvantages of using IPD will be given in Chapter 3.

1.2.1. Measures of treatment effect within individual trials and meta-analysis

For dichotomous data (i.e. data that can be categorised into a good or bad outcome for each patient e.g. dead or alive), the odds ratio, relative risk and risk difference are commonly used measures of treatment effect. Although the odds ratio is generally

thought to be harder to interpret compared to the risk difference or relative risk, the measure has desirable mathematical properties relating to its sampling distribution and its suitability for modelling [4]. In some situations, the risk difference can be more informative than the odds ratio or relative risk as it reflects the baseline risk as well as the change in risk with the intervention [5]. However, if the baseline event rate varies across trials, the risk difference is not usually appropriate. Each measure has advantages and disadvantages and the choice of which to use will depend on the example under consideration and the clinical questions being addressed. As issues surrounding meta-analysis of binary data are not considered here, the reader is referred to Deeks [6] for a comprehensive discussion of choosing a suitable summary statistic.

For continuous data (i.e. a numerical result for each patient e.g. systolic blood pressure) the difference in mean response in one treatment group compared to another (often referred to as the mean difference) is a commonly used measure of treatment effect for an individual study. Provided the outcome is measured on the same scale in each study in a meta-analysis, these results can be pooled to give a 'weighted mean difference'. If the outcome is measured on different scales in different studies (e.g. different scales to measure pain), the 'standardised mean difference' (an unitless summary measure) is sometimes used.

This thesis addresses issues surrounding the meta-analysis of time-to-event data (i.e. the time taken from some origin to a pre-defined end point of interest e.g. time to death after surgery) and is largely motivated by examples from epilepsy where time-to-event outcomes, such as time to seizure, are common. Time-to-event data are frequently summarised by the hazard function which is defined as the instantaneous event rate for an individual not experiencing the event to some time *t*. A treatment effect summarised as a hazard ratio (HR) is simply the ratio of hazards of the event of interest at any time for an individual on the experimental treatment relative to an individual on the standard treatment. The hazard ratio is the appropriate measure of treatment effect for failure time data as censoring, whereby the event of interest is not observed for a particular individual, and the time taken to achieve the event, are both properly allowed for with this measure. If only aggregate data (AD) are available, specific methods exist which enable the log hazard ratio (log(HR)) and its variance to be estimated in an individual trial. Further details of methods for extracting and estimating summary statistics to

undertake meta-analysis of aggregate time-to-event data are given in Chapter 2. Methods for meta-analysis of individual patient time-to-event data are discussed in Chapter 3.

1.2.2. Pooling results

The main objective when undertaking a meta-analysis is to obtain an estimate of the overall treatment effect pooled across J included studies. By pooling the relative treatment effect estimates from each individual study the 'within' study randomised comparison between treatment and control group is maintained. This is an important property of meta-analysis as it ensures that the advantage of a randomised controlled trial is maintained i.e. that there are no systematic differences in measurable or un-measurable characteristics between patients in the treatment groups that are compared.
The general approach to meta-analysis involves calculating a weighted average of individual trial treatment effects. Different methods exist to undertake these calculations and will depend on the type of data under consideration and what assumptions the meta-analyst is willing to make regarding the consistency of included trial results. Two common assumptions are made and are usually approached by considering two alternative models for meta-analysis; the *fixed effect* and *random effects* models.

Fixed effect approach

In the fixed effect approach, an assumption is made that the true treatment effect is homogenous (i.e. fixed) across studies and that each study is estimating the same common underlying treatment effect denoted θ. In other words, we assume

$$\theta_1 = \theta_2 = \ldots = \theta_j = \ldots = \theta_J = \theta$$

where θ_j is the treatment effect in trial j (j=1,2,…,J). Any variability in treatment effect estimates between studies is assumed to be due to sampling variability. In order to account for the differing precision of treatment effect estimates in each study, a weighted average is used to estimate the pooled treatment effect given by,

$$\hat{\theta} = \frac{\sum_{j=1}^{J} w_j \hat{\theta}_j}{\sum_{j=1}^{J} w_j} \qquad (1.1)$$

where w_j is the weight associated with the *jth* trial. Various methods are available to calculate w_j but these will not be discussed here. For a detailed description the reader is referred to Whitehead and Whitehead [7], Berlin *et al* [8], Yusuf *et al* [9], Hedges and Olkin [10]. Any choice of weight will provide an unbiased estimate of the pooled treatment effect, however w_j is usually taken to be the inverse of the variance of the treatment effect estimate in each trial as it provides the most precise estimate of the true treatment effect [11] i.e.

$$w_j = \frac{1}{v_j}$$

where $v_j = \text{var}(\hat{\theta}_j)$. By further assuming that

$$\hat{\theta}_j \sim N(\theta, v_j)$$

and that the weights w_j are known, a confidence interval for θ can be calculated using the result that

$$\hat{\theta} \sim N(\theta, 1/\sum_{j=1}^{J} w_j)$$

and an approximate 95% confidence interval for θ is given by

$$\hat{\theta} \pm 1.96\sqrt{\frac{1}{\sum_{j=1}^{J} w_j}}$$

Random effects approach

In a random effects meta-analysis the studies are regarded as a random sample from a population of possible treatment evaluations that can be used to estimate the mean

treatment effect and corresponding variance for the population [12]. It is commonly assumed that the treatment effect estimates ($\hat{\theta}_j$) follow a $N(\theta_j, v_j)$ distribution, and the true underlying treatment effects θ_j are themselves a sample of independent observations from a $N(\theta, \tau^2)$ distribution, where θ is the overall average treatment effect and τ^2 expresses the degree of variability between trial effects.

An estimate of the variability in treatment effect estimates across studies (τ^2) is incorporated into the model as an additional source of variation, which can result in a wider confidence interval for the estimated pooled effect. As the marginal distribution of $\hat{\theta}_j$ is given by,

$$\hat{\theta}_j \sim N(\theta, v_j + \tau^2)$$

and v_j and τ^2 are assumed to be known and equal to their estimated values, weights given by

$$w_j^* = \frac{1}{w_j^{-1} + \hat{\tau}^2}$$

are used in a random-effects model using the general formula (1.1) for calculating $\hat{\theta}$ with w_j replaced by w_j^*.

1.3. Heterogeneity

Statistical heterogeneity in meta-analysis can be defined as variation in the true underlying treatment effect between studies. A test for heterogeneity (or test of homogeneity) can be used to detect if the variation between study results is greater than that expected due to chance alone. Potential sources of statistical heterogeneity are,

(i) differences in clinical features (clinical heterogeneity, or clinical diversity) such as baseline characteristics or interventions used,

(ii) differences in methodological features such as randomisation methods or blinding,

(iii) differences in characteristics that have not been recorded or are simply not known.

Although a meta-analysis combines information from studies addressing the same (or very similar) clinical question, it is likely that some differences will exist between these studies. Interpreting a 'fixed effect' meta-analysis in the presence of heterogeneity can be misleading as the confidence interval is too narrow in terms of extrapolating the results to future trials or patients, since the extra variability between the results is ignored [13]. It is therefore important that heterogeneity is recognised and potential sources are explored so as to increase the clinical relevance of the conclusions drawn and the scientific understanding of the studies reviewed [13].

1.3.1. Detecting and quantifying heterogeneity

The extent of heterogeneity in meta-analysis can affect the interpretation of an overall pooled estimate. Problems of interpretation will depend on how substantial the heterogeneity is, since this determines the extent to which it might influence the conclusions of the meta-analysis [14].

Many formal hypothesis tests, which assess the evidence for heterogeneity are available, the most popular of which is Cochran's chi-square test [15] more commonly referred to as the 'Q test' which takes the form

$$Q = \sum_{j=1}^{J} w_j(\hat{\theta}_j - \hat{\theta})^2$$

Under the null hypothesis of homogeneity, the Q statistic has an approximate χ^2 distribution with J-1 degrees of freedom (df). For a detailed description of alternative test statistics and a discussion of the advantages and disadvantages of each, the reader is referred to Gavaghan *et al* [16], Takkouche *et al* [17], Hardy and Thompson [11]. These formal tests of homogeneity suffer from the disadvantage that they have low power, particularly when data are sparse [18], [19] and statistical heterogeneity may fail to be detected. To allow for this, a cut-off significance level of 0.10 is recommended rather than the conventional 0.05 value. Higgins and Thompson [14] have recently proposed a selection of three statistics to quantify the degree of heterogeneity and its impact on meta-analysis. In contrast to the commonly used Q statistic, the summary measures they propose do not depend on the number of trials in the meta-analysis. They conclude that

two statistics, which they refer to as H and I^2, are particularly useful summaries of the impact of heterogeneity and they recommend that one or both should be presented in published meta-analyses in preference to the usual test for heterogeneity [14]. The I^2 statistic describes the percentage of total variation across studies that is due to heterogeneity rather than chance and is calculated as

$$I^2 = \left(100\frac{Q-df}{Q}\right)\%$$

Negative values of I^2 are set to zero so that I^2 lies between 0% and 100% where a value of 0% indicates no observed heterogeneity, and larger values show increasing heterogeneity.

1.3.2. Investigating heterogeneity

If heterogeneity in the treatment effect between studies is evident, a strategy is needed for exploring potential sources, which should be specified *a priori* to avoid over exploring the data and to limit the potential for spurious findings. The extent of this exploratory analysis will depend mainly on the availability of data and the number of trials included in the meta-analysis. Furthermore, the possibility of obtaining several different explanations for heterogeneity means that such investigations are not always straightforward.

A popular approach is to undertake subgroup analyses in which the effect of treatment is examined in pre-specified and clinically important subgroups. Subgroups can either be defined at the trial level, in which all patients in a trial appear in only one subgroup, or at the patient level, in which some patients in a trial appear in one subgroup and other patients from the same trial appear in another subgroup. Although relatively straightforward, subgroup analyses have several limitations and should always be interpreted with caution. A particular problem for subgroup analyses using AD is that many studies may not present results for the subgroups of interest. This may be because the subgroups were not examined or alternatively, it may be that the results obtained were not significant and therefore were not reported. Hahn *et al* [20] discuss the potential for bias introduced by the selective reporting of subgroup analyses within

individual studies. Gelber and Goldhirsch [21] and Yusuf *et al* [22] discuss more general problems associated with subgroup analyses in more detail.

Following detection of heterogeneity and an investigation into possible causes, many reviewers adopt a random effects approach to incorporate the additional unexplained between trial variation into the model. A common criticism of the random effects approach is that more imprecise estimates from smaller studies are given more weight compared to the corresponding fixed effect approach. As the degree of heterogeneity between studies increases, this discrepancy also increases. Many researchers criticise the actual assumption made in a random effects model that the trials involved are considered a random sample from some hypothetical population of trials as this contradicts the underlying principle that a meta-analysis undertaken as part of a systematic review should incorporate all available trials that address the question of interest. On the other hand, the assumption of a common underlying treatment effect made by the fixed effect approach may be considered as overly restrictive since trials included inevitably encompass a substantial variety of specific treatment regimens, types of patients, and outcomes [13]. The choice between a fixed effect and random effects approach to meta-analysis should be made by the individual meta-analyst and is not the focus of this thesis. In the author's opinion, a selection of factors that may cause heterogeneity require careful thought prior to analysis and should always be considered.

1.4. Publication Bias

Publication bias is a common problem in meta-analysis which arises when unpublished studies or outcomes are not identified or not included in the analysis and the reason for this is related to their results. As research with statistically significant results is more likely to be published compared to research with non-significant results, combining only published studies can lead to an over-optimistic conclusion [23], [24]. The potential for publication bias can be minimised by adopting a comprehensive search strategy that includes sources of unpublished studies and foreign language journals. In addition, contact should be made with experts in the field and pharmaceutical companies or manufacturers (if appropriate), and relevant journal/conference abstracts should also be hand-searched. Several methods exist to detect the potential presence of publication bias and to compensate for this in the analysis. These issues will not be addressed in this

thesis but it should be noted that the availability of individual patient data may be useful to overcome the problem of unreported outcomes if the required data are provided.

1.5. Discussion and structure of the thesis

A meta-analysis within a rigorous systematic review can be very informative and is considered by many as providing the highest level of evidence in medical research. Although much of the methodology surrounding meta-analysis has been well researched, many issues remain unresolved. The main focus of the remaining Chapters of this thesis is to evaluate, develop and compare methods for meta-analysis with time-to-event outcomes. The main motivational example consists of a suite of systematic reviews and meta-analyses based on IPD which compare alternative drugs for epilepsy. Although an IPD approach is regarded as the "yardstick" against which other forms of systematic review should be measured [25], the additional demand imposed on resources mean that the majority are based on AD. In many situations an AD approach may be the only option and it is important to establish the reliability of such results when compared to IPD. In Chapter 2, methods for estimating summary treatment effect measures for time-to-event outcomes based on AD are discussed with particular emphasis on improving a current method and assessing the value and reliability of general AD approaches with this type of data. This leads into Chapter 3 where commonly used methods for meta-analysis with IPD for time-to-event outcomes are described and explored. Individual meta-analysis results for the collection of epilepsy drug trial systematic reviews are described in Chapter 4.

The additional resources required for an IPD approach may be justified if the IPD can be fully exploited by investigating potential prognostic factors and including a thorough investigation of possible sources of heterogeneity. In Chapter 5, methods for modelling heterogeneity with AD or IPD are described. Alternative random effects models, and programs for fitting these models, based on IPD are developed and extended. An assessment of how models based on AD or IPD compare in terms of exploring heterogeneity is also undertaken as this must be considered when deciding whether the extra investment required for an IPD approach to meta-analysis is worthwhile. In Chapter 6, the available IPD for several clinically related systematic reviews of epilepsy drug trials is fully exploited by considering methodology for indirect comparisons that is

extended to facilitate the simultaneous analysis of all available evidence from these reviews in what is termed here as the *totality of evidence* analysis. The final chapter provides a summary of preceding chapters, some concluding remarks and discussion of future research ideas.

CHAPTER 2

Aggregate data meta-analysis with time-to-event outcomes

An aggregate data (AD) meta-analysis uses summary statistics either extracted directly from trial reports or requested from the authors of unpublished articles. An alternative approach is to request individual patient data (IPD) from authors, resulting in a more rigorous meta-analysis, particularly when dealing with time-to-event outcomes. However, due to the increase in time and resources required, or the unavailability of data, most meta-analyses are based on AD. Furthermore AD can be very useful if a preliminary analysis is required before deciding whether collecting IPD is worthwhile or if conducting a meta-analysis as part of safety monitoring during a new trial. Further discussion about the practicalities and benefits of using IPD will be given in Chapter 3 and the reader is referred to Stewart and Clarke [26] for an in-depth description of the IPD meta-analysis process.

2.1. Time-to-event data

Time-to-event data, or survival data, are frequently encountered in medical research and arise when data relating to the time taken from some origin to the event of interest are collected e.g. time from surgery to death, or time from start of treatment to first seizure.

Two common features of time-to-event data are (i) data are typically skewed, and (ii) data are frequently censored, whereby the event of interest (e.g. death) is not observed for a particular individual. Consequently, standard statistical methods for dealing with continuous data cannot be used to analyse time-to-event data.

The assumption of non-informative censoring is made throughout this thesis. That is, the actual time to an event for an individual is independent of any mechanism which causes that individual's time to event to be potentially censored.

2.1.1. Survivor function and hazard function

The s*urvivor function* S(t) and *hazard function* h(t), defined below, are the functions most commonly used to describe time-to-event data.

The s*urvivor function* is defined as the probability that the event is not observed between the time origin and some time t. In other words, the probability that the time to an event is greater than or equal to t,

$$S(t) = P(T \geq t) = 1 - F(t)$$

where t is the actual time to an event of an individual, T is the random variable associated with the time-to-event, and F(t) is the distribution function of T given by,

$$F(t) = P(T < t) = \int_{u=0}^{t} f(u)\mathrm{d}u$$

where *f(u)* is the underlying probability density function of T.

The h*azard function* is defined as the probability that the event occurs instantaneously at time t, conditional on not having the event up to that time.

$$h(t) = \lim_{\delta t \to 0}\left\{\frac{P(t \leq T < t + \delta t \mid T \geq t}{\delta t}\right\}$$

2.2. Methods of analysis of time-to-event data

The analysis of time-to-event data requires specialist methods that take censoring and the data structure into account. Some of the most common approaches of analysis, which are considered throughout this thesis, are described below with further details and discussion provided in later sections and Chapters.

Non-parametric methods

The most common non-parametric approaches for estimating the survivor function for a single sample of survival times are the Kaplan-Meier [27] and actuarial, or life-table, methods with results typically summarised graphically. The actuarial method is particularly well suited for grouped data where the number of events and number of censored observations within a particular interval of time may be the only available information. For ungrouped data, in which the actual event times are available, the Kaplan-Meier estimate is most appropriate [29].

When interest centres on comparing the time-to-event experiences of two or more groups, a particularly informative summary may be achieved from plotting estimates of the survivor function for each group on the same graph. The log-rank test [28] is a non-parametric procedure for assessing the evidence against the null hypothesis of no difference in the time-to-event experiences between two or more groups of individuals.

From a clinical perspective, a summary of treatment effect and measure of precision is usually required. Examples of possible measures of treatment effect include the hazard ratio, difference in median survival times, difference in crude event rates and the difference in survival estimates at fixed time points. The hazard ratio, with underlying assumption that the hazard at time t for a patient in one group is proportional to the hazard at the same time for a patient in another group, takes both censoring and time to an event into account and is the most appropriate measure of treatment effect for this type of data. Further details for estimating the hazard ratio non-parametrically are given in section 2.3 and 3.2. Each of the remaining possible measures suffers from at least one disadvantage compared to the hazard ratio. It may not be possible to calculate the median survival time in one or both groups if the survivor function remains greater than 0.5, therefore the difference in median survival times cannot always be calculated. The

difference in survival estimates at a fixed time point has an interpretation which is appealing to clinicians since it summarises how much greater or smaller is the probability of surviving beyond certain clinical milestones for one group compared to another. However, information on the whole time scale is not taken into account and since later parts of the time scale contain less information, precision of this treatment effect measure will depend on the time point considered. The difference in crude event rates ignores information on censoring as well as ignoring the whole time scale and is therefore even less appealing.

Semi-parametric method

To examine relationships between time-to-event experiences of individuals and their clinical characteristics, a modelling framework may be adopted. The most commonly used approach is the semi-parametric proportional hazards model proposed by Cox [30]. The assumption that the hazard of an event at any given time for an individual in one group is proportional to the hazard at the same time for an individual in a different group is referred to as an assumption of proportional hazards and underlies this method of analysis. Since no distributional assumptions are imposed upon the time-to-event distribution of individuals, the model is a *semi-parametric* model and will be referred to as the Cox regression model in remaining Chapters. One particular advantage with this approach is that an estimate of the hazard ratio and appropriate confidence interval may be obtained directly from results of fitting this model to the data.

Parametric methods

If the assumption of a particular probability distribution for the data is valid, inferences based on such an assumption will be more precise and estimates of hazard ratios will tend to have smaller standard errors compared to a model without the distributional assumption [29]. Fully parametric models assume a specific baseline distribution for the survival times and the effect of covariates on the baseline function is also fully specified. Commonly used distributions include the exponential, weibull, log-logistic and log-normal distribution. Although fully parametric models may offer advantages in terms of efficiency, they involve stronger assumptions and require an assessment of the appropriateness of the chosen distribution. Due to these added complications, models

making further distributional assumptions will not be considered further in the current thesis but could certainly offer some advantages over the more flexible but potentially less efficient semi-parametric model.

The log hazard ratio rather than hazard ratio is used as a measure of treatment effect for meta-analysis of time-to-event data because the hazard ratio can take values between zero and infinity with a different interpretation for values between 0 and 1, or 1 and infinity. As this scale is not symmetric, a log transformation of the hazard ratio is taken to transform the measure to a symmetric scale (around zero) ranging between minus infinity and infinity. Methods for extracting and estimating the log hazard ratio and its variance from individual studies to enable a meta-analysis to be undertaken will be described in the following sections.

2.3. Estimating the log hazard ratio and its variance using aggregate data

At a very basic level, if an estimate of the log hazard ratio ($\log(HR_j)$) and its variance ($\text{var}(\log(HR_j))$ are presented in the manuscript of each trial j (j=1,..,J) in a systematic review, undertaking a meta-analysis can be straightforward provided that a clear description of the end-points are given. Furthermore, if these summary statistics are not reported directly, several methods are available to allow estimation using alternative types of AD e.g. the log-rank test p-value and number of events. However, the presentation, quality and consistency in reporting of time-to-event data analysis in the literature is diverse. Altman *et al* [31] undertook a systematic review of survival analyses published in cancer journals and found that the majority of papers gave an unclear description of at least one study end point, almost half of the papers did not summarise length of follow-up, and the results of log-rank and multivariate analyses were frequently only summarised using a p-value. In addition, they found that many papers presented survival curve plots but the quality of these plots was poor.

A description of methods for estimating $\log(HR_j)$ and its variance using AD is given by Parmar *et al* [32], who also propose a specific approach using Kaplan-Meier survival curves. In one particular example comparing the length of survival for women with advanced breast cancer randomised to different treatments, the authors note that the method using Kaplan-Meier survival curves to estimate $\log(HR_j)$ and its variance

appears to perform reasonably well except in a few cases. The methods proposed by Parmar *et al* [32] for estimating $\log(HR_j)$ and its variance within each trial are described in the following sections. A further modification to the approach based on published survival curves is reviewed in this thesis (published by Williamson, Tudur Smith *et al* in *Statistics in Medicine* [33]). The modified approach incorporates additional information from the 'numbers at risk'. Two separate studies are used in section 2.4 to illustrate and compare results obtained from the two survival curve based methods of estimation whilst in section 2.5, the practicality, reliability and value of the AD meta-analysis with time-to-event outcomes are investigated using two further empirical examples of meta-analysis (published by Tudur, Williamson *et al* in *JRSSA* [34]).

2.3.1. Direct method

Usually, no further estimation is required if $\log(HR_j)$ and its variance are quoted directly in the trial manuscript. Authors may report the coefficient of treatment effect and variance (more usually the standard error) estimated from a Cox proportional hazards model. These parameters correspond directly to estimates of $\log(HR_j)$ and its variance (or standard error). However, interpretation can be difficult if the results of multivariate Cox regression models are reported as each study will rarely adjust for the same covariates in the model.

2.3.2. Indirect method 1: Estimating var(log(HR$_j$)) from a confidence interval

If an estimate of $\log(HR_j)$ is reported without its variance or standard error but rather with a $(1-\alpha)100\%$ confidence interval denoted here by $(LLIM_j, ULIM_j)$, an estimate of $\text{var}(\log(HR_j))$ can be obtained using

$$\text{var}(\log(HR_j)) = \left[\frac{ULIM_j - LLIM_j}{2\Phi^{-1}(1-\alpha_j/2)}\right]^2 \qquad (2.1)$$

where Φ^{-1} is the inverse cumulative probability for the normal distribution.

2.3.3. Indirect method 2: Estimating log(HR_j) and its variance using the quoted p-value of the log-rank test

The p-value of the log-rank test is frequently quoted without the log-rank test statistic being given [31]. Provided the total number of events across both treatment groups (O_j) is given, this information along with the quoted (to at least 2 decimal places) two sided p-value (p_j) can be used to estimate log(HR_j) and its variance as described below.

As described by Parmar *et al* [32], the following standard results are used to estimate log(HR_j)

$$\log(HR_j) = \frac{O_{ej} - E_{ej}}{V_{rj}}$$

$$\frac{O_{ej} - E_{ej}}{\sqrt{V_{rj}}} = \Phi^{-1}(1 - p_j/2)$$

where the subscript e denotes the experimental treatment group, and V_{rj} denotes the approximation to the variance of the log-rank statistic for trial j (j=1,…,J). It follows that

$$\log(HR_j) = \frac{\Phi^{-1}(1 - p_j/2)}{\sqrt{V_{rj}}} \qquad (2.2)$$

with

$$\text{var}(\log(HR_j)) = \frac{1}{V_{rj}} \qquad (2.3)$$

Three different estimates for V_{rj} are given by,

$$V_{rj} = \frac{O_j}{4} \qquad (2.4)$$

$$V_{rj} = \frac{O_j R_{ej} R_{cj}}{(R_{ej} + R_{cj})^2} \qquad (2.5)$$

$$V_{rj} = \frac{O_{ej} O_{cj}}{O_j} \qquad (2.6)$$

where O_{ej} and O_{cj} denote the observed number of events for the experimental and control group whilst R_{ej} and R_{cj} denote the number randomised to the experimental and control group. The two approximations (2.4) and (2.5) are identical if there are equal sample sizes in both groups of the study ($R_{ej} = R_{cj}$), whilst approximations (2.4) and (2.6) are identical if the number of events in each group are equal ($O_{ej} = O_{cj}$). Collette *et al* [35] have investigated the three approximations for V_{rj} by simulation of meta-analyses of 10 trials. They conclude that all three approximations provide very close estimates to the overall individual patient data log-rank variance. In particular, approximation (2.6) is the most precise for trials with a low percentage of censoring, and (2.5) is preferred for trials with unequal sample sizes.

2.3.4. Indirect method 3a: Estimating log(HR_j) and its variance from survival curves

Survival curves are frequently presented to display the results of a time-to-event analysis graphically. The information summarised within the plotted survival curves may be extracted and used to estimate log(HR_j) and its variance. The method is briefly described here but further details are given by Parmar *et al* [32].

For each trial the time axis of each survival curve should be split into non-overlapping intervals such that the event rate within each interval is relatively small. An estimate of survival probability for each treatment group at each specified time-point should be read off the published curves. The number of patients at risk is estimated for each time interval by using an expression involving the estimated number of events and censored observations, whereby the number of events during an interval is estimated using an expression involving the extracted survival probabilities and estimated number at risk, and the number of censored observations during an interval may be estimated by assuming a model for censoring during the interval. For the assumption that patients are

censored at a constant rate across the entire follow-up period, the minimum and maximum follow-up times are required. This information may be provided directly in trial reports or estimated from other sources such as the article publication date.

The log hazard ratio and its variance are estimated within each interval, denoted by subscript *i*, using the following expressions

$$\log(HR_i) = \log\left(\frac{d_{ei}(t)/n_{ei}(t)}{d_{ci}(t)/n_{ci}(t)}\right)$$

$$\text{var}(\log(HR_i)) = \frac{1}{d_{ei}} - \frac{1}{n_{ei}} + \frac{1}{d_{ci}} - \frac{1}{n_{ci}}$$

where d_{ei}, d_{ci} denote the number of deaths estimated within an interval for experimental and control group respectively and n_{ei}, n_{ci} denote the estimated number at risk for each group. An overall estimate of log hazard ratio and its variance for a particular trial may finally be estimated by calculating an inverse variance weighted average of interval-specific estimates. Parmar *et al* [32] give further detail and description of relevant expressions for estimating the required quantities.

2.3.5. Indirect method 3b: Estimating log(HR_j) and its variance from survival curves and numbers at risk

To estimate the number of events and number at risk for method 3a, some assumption about the pattern of censoring across the period of follow-up is required, for example an assumption of constant censoring during the trial. The numbers at risk are often quoted below a survival curve plot or indeed within the text of a trial publication. Pocock *et al* [36] undertook a review of publications quoting survival analyses and found that where there was variable length of follow-up almost all trials presenting survival curves had also presented numbers at risk somewhere near the plot or in the text of the trial report. The availability of numbers at risk at various time points provides additional information on the censoring pattern within a trial. Including this information would mean that strong assumptions about censoring patterns across the entire follow-up

period would not be needed. This is what the following method proposes (published by Williamson, Tudur Smith *et al* in *Statistics in Medicine* [33]).

Suppose the numbers at risk $n_{k,1},....,n_{k,p}$ for each treatment group k (k=1,2), are given either on the survival curve or in the text of a report at each of p time-points $t_1,....,t_p$ respectively. Survival probabilities should be read off the curves at $t_1,....,t_p$ and are denoted by $s^*_{k,1},....,s^*_{k,p}$. By definition, let $t_0 = 0$, $s^*_{k,0} = 1$, $n_{k,0} =$ the number randomised in treatment group k.

The general method is to estimate the log hazard ratio and its variance within each time interval $[t_{i-1},t_i)$, i=1, …, p and to combine these estimates using an inverse variance weighted average across intervals. For this approach, estimates of both the number of events and the number at risk during the interval are required.

Following the actuarial approach [37] in which censoring is assumed to be constant within each time interval, but not necessarily across intervals

$$s^*_{k,i} = s^*_{k,i-1}\left[1-\frac{d^*_{k,i}}{n_{k,i-1}-(c^*_{k,i}/2)}\right] \qquad (2.7)$$

$$n_{k,i} = n_{k,i-1} - d^*_{k,i} - c^*_{k,i} \qquad (2.8)$$

where $d^*_{k,i}$ = number of events in $[t_{i-1},t_i)$ and $c^*_{k,i}$ = number censored in $[t_{i-1},t_i)$. Rearranging (2.7) and (2.8) gives

$$d^*_{k,i} = \frac{(n_{k,i-1}+n_{k,i})(s^*_{k,i-1}-s^*_{k,i})}{(s^*_{k,i-1}+s^*_{k,i})} \qquad (2.9)$$

$$c^*_{k,i} = \frac{2(n_{k,i-1}\,s^*_{k,i}-n_{k,i}s^*_{k,i-1})}{(s^*_{k,i-1}+s^*_{k,i})} \qquad (2.10)$$

$$n^*_{k,i} = \frac{(n_{k,i-1} + n_{k,i})s^*_{k,i-1}}{(s^*_{k,i-1} + s^*_{k,i})} \qquad (2.11)$$

where $n^*_{k,i}$ is the number at risk during the interval $[t_{i-1}, t_i)$. The method assumes that censoring is uniform over the intervals defined by the numbers at risk.

For an individual trial, the log hazard ratio and its variance within each interval $[t_{i-1}, t_i)$ are then estimated using one of the following sets of formulae.

$$\log(HR)_i = \log\left[\frac{d^*_{2,i}/n^*_{2,i}}{d^*_{1,i}/n^*_{1,i}}\right] \qquad \text{var}(\log(HR)_i) = \frac{1}{d^*_{2,i}} - \frac{1}{n^*_{2,i}} + \frac{1}{d^*_{1,i}} - \frac{1}{n^*_{1,i}} \qquad (2.12)$$

For **(2.13)-(2.14)** below,

$$\log(HR)_i = \frac{d^*_{2,i} - e^*_{2,i}}{v_i}, \quad \text{var}(\log(HR)_i) = \frac{1}{v_i}, \quad \textit{where} \quad e^*_{2,i} = (d^*_{2,i} + d^*_{1,i}) * \frac{n^*_{2,i}}{n^*_{2,i} + n^*_{1,i}}$$

$$v_i = \frac{d^*_{2,i} + d^*_{1,i}}{4} \qquad (2.13)$$

$$v_i = (e^*_{2,i} + e^*_{1,i}) * \frac{(n^*_{2,i} * n^*_{1,i})}{(n^*_{2,i} + n^*_{1,i})^2} \qquad (2.14)$$

Approximation (2.13) and (2.14) are identical if the number at risk during an interval are equal for both treatment groups. Approximation (2.14) can also be written as $v_i = e^*_{2,i} * e^*_{1,i}/(e^*_{2,i} + e^*_{1,i})$ where $e^*_{2,i}, e^*_{1,i}$ are commonly crudely estimated by $d^*_{2,i}, d^*_{1,i}$. However, as this additional estimation in the approximation is not necessary here, it will not be considered further.

The overall estimates for the *jth* trial are then calculated from

$$\log(\hat{HR}_j) = \frac{\sum_{i=1}^{p} w_i \log(\hat{HR})_i}{\sum_{i=1}^{p} w_i}, \quad \textit{and} \quad \text{var}(\log(\hat{HR}_j)) = \frac{1}{\sum_{i=1}^{p} w_i}$$

where $w_i^{-1} = \text{var}(\log(\hat{HR})_i)$.

2.4. Comparison of methods based on survival curves

Two studies for which individual patient data are available are used to compare estimates obtained from the two approaches based on survival curves (method 3a and method 3b) because frequently a meta-analyst will only have survival curves to work from.

2.4.1. Example 1

The first example is a randomized controlled trial [38] including 466 patients with epilepsy taken from a systematic review of five randomized controlled trials comparing two anti-epileptic drugs, carbamazepine (CBZ) and sodium valproate (VPS) [39]. One of the outcomes of interest is time to first seizure following randomisation. The Kaplan-Meier survival curves with number of patients at risk and survival probabilities at each time point, generated using individual patient data since curves were not presented in the published manuscript, are summarised in Figure 2.1.

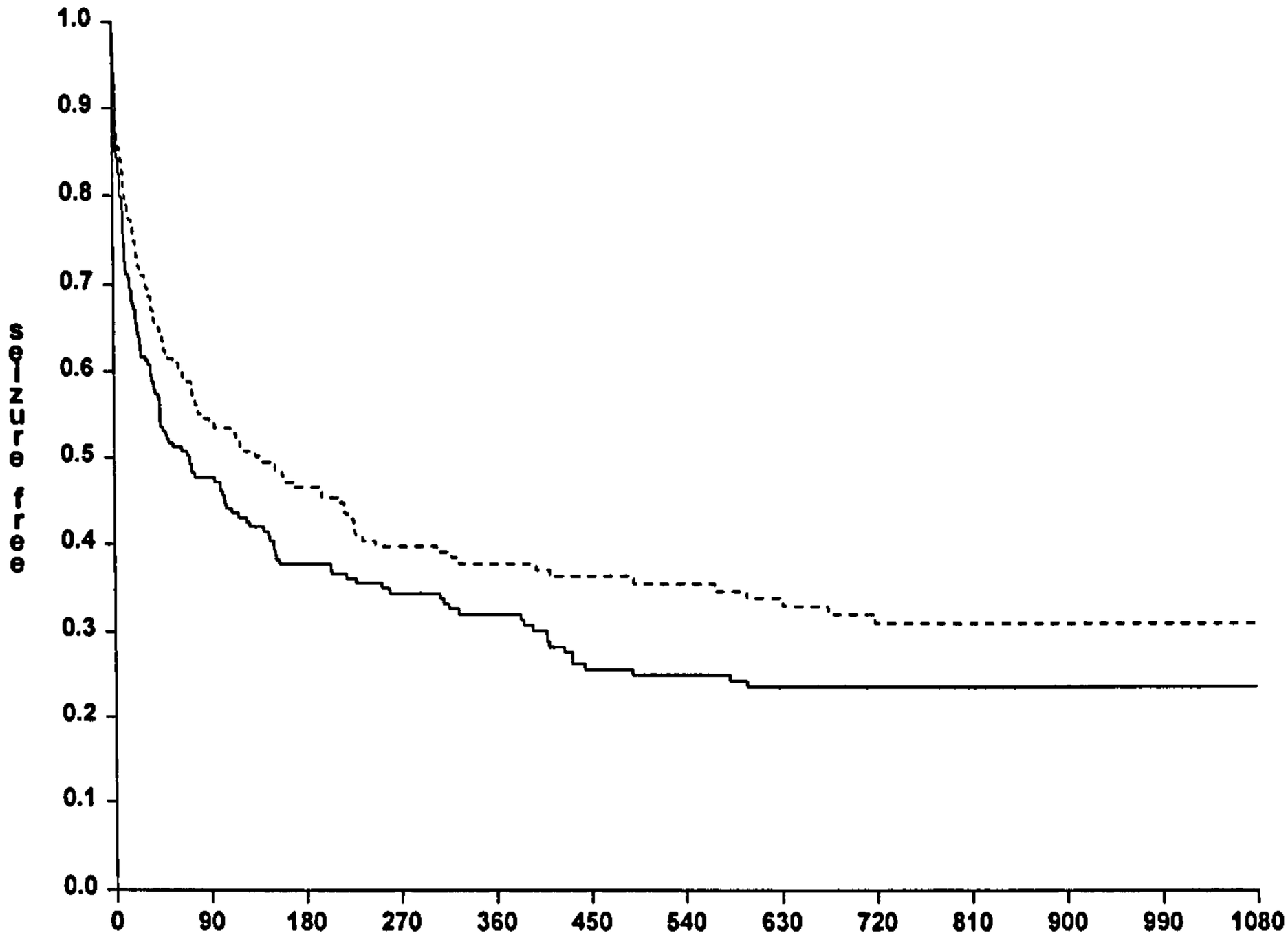
Table 2.1 shows how the results differ depending on the formulae used for calculating the log hazard ratio and its variance within an interval (2.12, 2.13, 2.14), whether numbers at risk are used (method 3a versus 3b), whether survival probabilities are read off the survival curve rather than calculated directly, the range over which probabilities are taken and the effect of varying interval widths. For indirect method (3a), Parmar *et al* [32] propose using expression (2.12) for approximating the log hazard ratio and variance within an interval. However, the alternative approximations given by (2.13) and (2.14) could be used and are explored in more detail for examples 1 and 2.

In this trial, patients were censored quite heavily early on. The assumption of constant

censoring across the entire follow-up period (method 3a) has led to the numbers at risk being overestimated for the earlier times (Figure 2.2) and hence the variance is underestimated. The estimate of log(*HR*) is slightly closer to the IPD estimate when the assumption of constant censoring is made, whereas the estimate of SE(log(*HR*)) based on numbers at risk is closer to the IPD results. For both approaches (method 3a and 3b), the estimate of SE(log(*HR*) is most precise using estimators (2.13) and (2.14) based on the log-rank observed and expected number of events.

The effect of using survival probabilities read off the survival curve rather than calculated exactly is found to be minimal in this example. As the range over which probabilities are taken increases, estimates of log(HR) decrease substantially away from the IPD value whilst the SE(log(HR)) decreases only slightly. As interval lengths are decreased, the estimate of log(HR) increases towards the IPD estimate, whereas the estimate of SE(log(HR)) increases above the IPD estimate.

Figure 2.1. Example 1: Kaplan–Meier estimates of time to first seizure of 466 patients with epilepsy treated with sodium valproate (VPS) or carbamazepine (CBZ) in a single randomized controlled trial [38].



Patients at risk

vp	———	239	93	70	59	53	39	36	34	29	22	20	18	18
cbz	- - - - -	227	101	78	62	56	46	43	39	31	30	30	26	24

Survival probabilities

vps	1	0.477	0.378	0.344	0.320	0.256	0.249	0.235	0.235	0.235	0.235	0.235	0.235
cbz	1	0.546	0.466	0.398	0.378	0.363	0.355	0.338	0.310	0.310	0.310	0.310	0.310

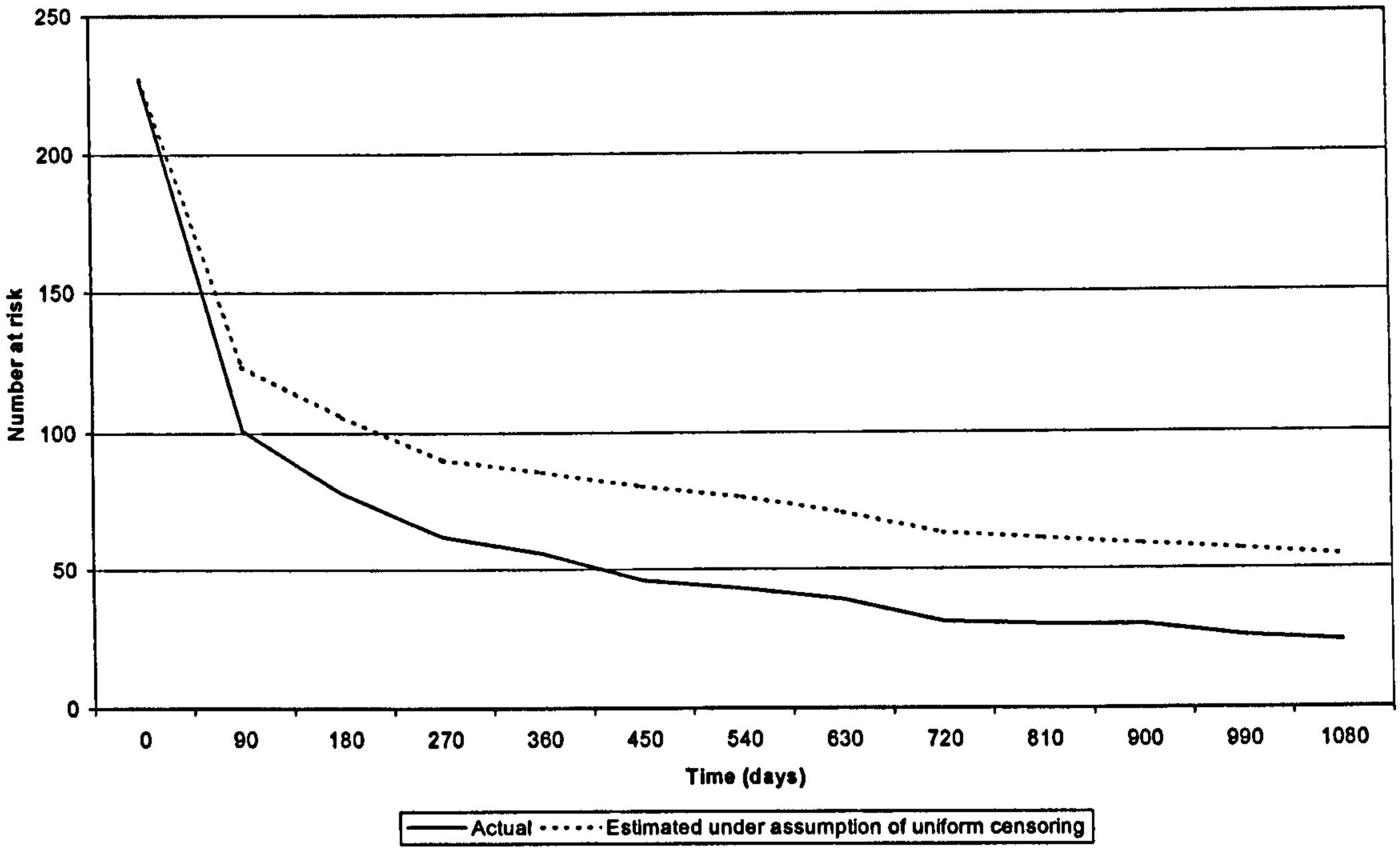
Table 2.1. Example 1: Comparison of estimates of overall log(HR) and SE(log(HR)) (pooled across intervals)

Method		*log(HR)*	SE(*log(HR)*)
Individual patient data		0.206	0.1162
Method 3a	**Estimators**		
Survival probabilities	**(2.12)**	0.1689	0.0854
from IPD results every	**(2.13)**	0.1674	0.1086
90 days up to 3 years[1]	**(2.14)**	0.1678	0.1091
Method 3b Survival probabilities from IPD results every 90 days up to 3 years and numbers at risk	**Estimators** **(2.12)** **(2.13)** **(2.14)**	0.1610 0.1638 0.1640	0.0906 0.1169 0.1170
Method 3b Survival probabilities read off figure 1 every 90 days up to 3 years and numbers at risk [2]		0.1581	0.1170
Method 3b Survival probabilities and numbers at risk from IPD results every 90 days up to:[2]	**(a) 2 years** **(b) 3 years** **(c) 5 years**	0.1640 0.1640 0.1346	0.1170 0.1170 0.1158
Method 3b Survival probabilities and numbers at risk from IPD results up to 3 years:[2]	**(a) every 30 days** **(b) every 60 days** **(c) every 90 days** **(d) every 180 days**	0.2071 0.1859 0.1640 0.1496	0.1173 0.1172 0.1170 0.1164

[1] Censoring rate assumed constant across follow-up period from 365-2190 days

[2] Expression (2.14) used to estimate *log hazard ratio* and its variance within each interval

Figure 2.2. Example 1: Comparing actual and estimated numbers at risk for CBZ group (similar pattern for VPS group).



2.4.2. Example 2

The second example illustrates the estimation of log hazard ratio and its standard error from a life-table summarizing the survival experience of males and females with cerebral palsy born between 1966 and 1984 in the Mersey region [40]. Since data have been continually accrued for this cohort, an up-dated life-table has been used in the calculations that follow. Estimates of log hazard ratio and standard error are displayed in Table 2.2 in addition to the results obtained from IPD for comparison.

In contrast to example 1, censoring is heavier towards the end of this study. From reading the manuscript, the period of follow-up would appear to be from 0–27 years. The assumption of constant censoring across this entire follow-up period (method 3a) has led to the numbers at risk being underestimated for the earlier times (Figure 2.3) resulting in the variance being overestimated.

The estimates of log(HR) are closer to the IPD estimate when numbers at risk are incorporated (method 3b) using estimators (2.13) and (2.14). In fact, estimator (2.12) is particularly poor compared to (2.13), (2.14) and the IPD estimate and would require

further theoretical exploration of these alternative estimators to establish a potential explanation for this discrepancy. If numbers at risk are not included, the estimates are substantially poorer. The estimates of SE(log(HR)) are more precise in all cases using (2.13) and (2.14), based on the observed and log-rank expected number of events. The availability of IPD in this example allowed us to identify that in fact no censoring had occurred before 9 years. The last row of Table 2.2 includes the estimates obtained from assuming constant censoring (method 3a) across the period 9-27 years. As expected, by reducing the period over which uniform censoring is assumed, the SE(*log*(*HR*)) decreases and the estimated numbers at risk pattern (Figure 2.3) is then similar to the pattern observed in Example 1 (Figure 2.2), resulting in the variance being underestimated compared to the approach incorporating numbers at risk.

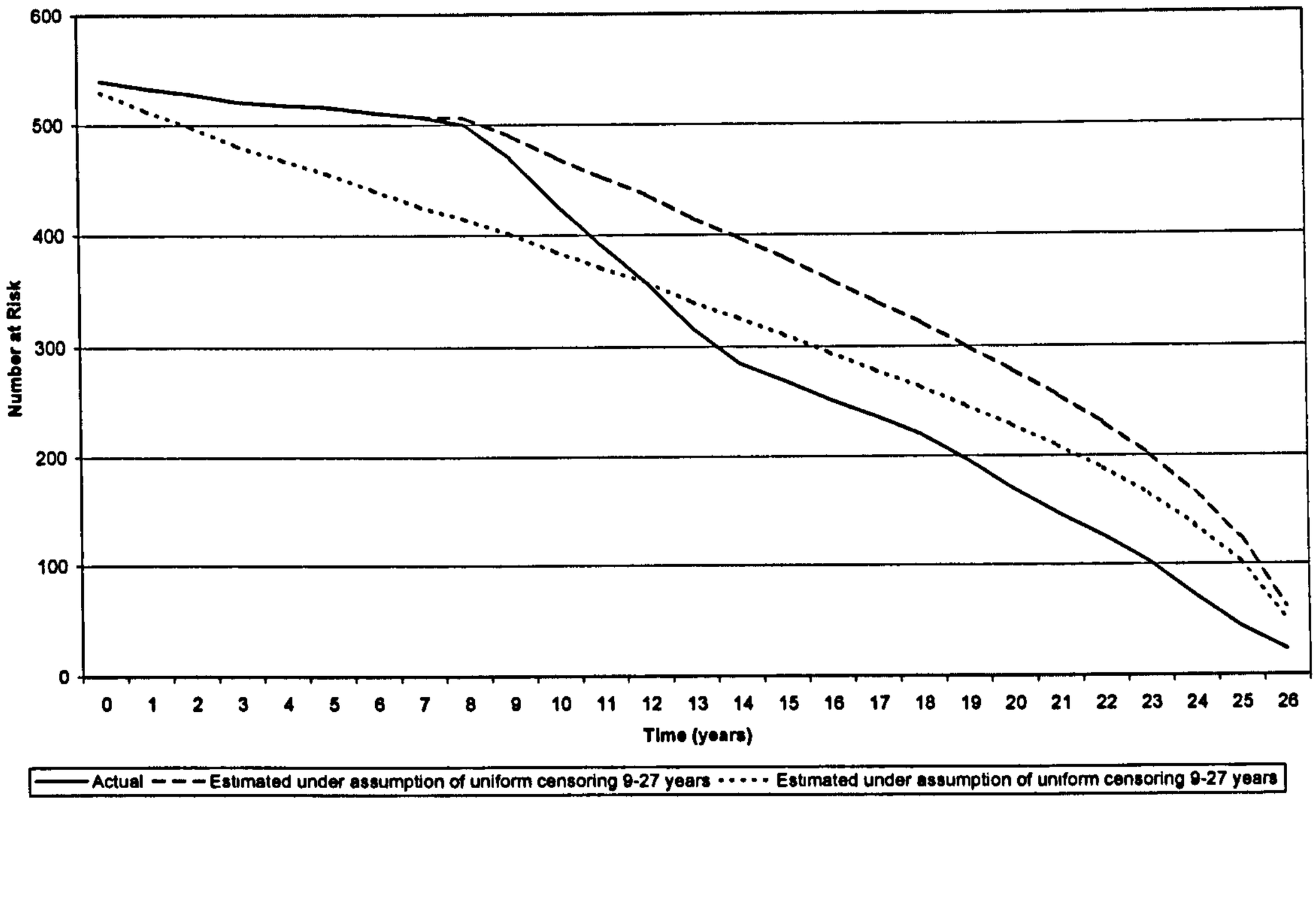
Table 2.2. Example 2: Comparison of estimates of overall log(HR) and SE(log(HR)) (pooled across intervals)

Method	*log(HR)*	SE(*log(HR)*)
Individual patient data	0.1321	0.1696
Method 3a Without numbers at risk [1] **Estimators**		
(2.12)	0.0167	0.1908
(2.13)	0.1664	0.1696
(2.14)	0.1693	0.1711
Method 3b Incorporating numbers at risk **Estimators**		
(2.12)	0.0129	0.1854
(2.13)	0.1414	0.1667
(2.14)	0.1441	0.1683
Method 3a Without numbers at risk[2] **Estimators**		
(2.12)	0.0217	0.1792
(2.13)	0.1788	0.1584
(2.14)	0.1819	0.1597

[1] Censoring rate assumed constant across period from 0-27 years.

[2] Censoring rate assumed constant across period from 9-27 years.

Figure 2.3. Example 2: Comparing actual and estimated numbers at risk for males (similar pattern for females).



2.5. Reliability of methods for aggregate data meta-analysis

Parmar *et al* [32] have examined 209 randomised comparisons in advanced breast cancer. In 62 of these comparisons, insufficient AD were available to enable the log hazard ratio and its variance to be estimated for the end-point of interest. For 48 comparisons, either direct or indirect methods (indirect method 1, indirect method 2) and the method proposed based on survival curves (indirect method 3a) could be employed to estimate $\log(HR_j)$ and its variance. This enabled a comparison to be made between direct or indirect based estimates with those obtained from survival curves (method 3a). In the comparisons examined the authors believe the survival curve estimate of $\log(HR_j)$ (method 3a) appears to perform reasonably well except in a few cases suggesting that overall there was no evidence of systematic bias in the survival curve estimate. On the other hand, estimates of $\text{var}(\log(HR_j))$ obtained from direct or indirect approaches appeared to be underestimated by the survival curve approach (method 3a). In particular it appears that this discrepancy increases as the actual variance increases. Parmar *et al* [32] suggest that this underestimation is due to the uniform censoring assumption that is made when adopting the survival curve approach. The method proposed in section 2.3.5 (method 3b) utilising numbers at risk attempts to improve estimation by

overcoming the need for this assumption. There is clearly a need to examine further examples of meta-analysis to compare results between different aggregate data based approaches (methods 1, 2, 3a and 3b). A comparison and investigation into the practicality, reliability and value of AD meta-analysis with time-to-event outcomes is undertaken in the next section.

2.5.1. Examples

Evaluation and comparison of AD based methods is undertaken using two examples of meta-analysis from different clinical fields. In the first example, only AD were extracted from randomised trials comparing the effectiveness of TIPS (transjugular intra-hepatic portosystemic shunt) and ES (endoscopic sclerotherapy) in the treatment of variceal bleeding. In the second example, IPD were available for some but not all trials comparing palliative chemotherapy versus supportive care in colorectal cancer.

Example 3: Aggregate data MA proposed

In this example IPD were not collected initially due to resource constraints. Khan *et al* [41] identified eleven randomised controlled trials suitable for inclusion in a systematic review to compare the effectiveness of TIPS (experimental group) and ES (control group) for the treatment of variceal bleeding, one of the most frequent and severe complications of chronic liver disease. Several outcomes were examined in the review, but time to death, the primary outcome, is the only end-point considered here.

Example 4: IPD available for some trials

The colorectal cancer collaborative group [42] included thirteen randomised controlled trials in a systematic review of the benefits and harms of palliative chemotherapy (experimental group) compared with supportive care (control group) for patients with locally advanced or metastatic colorectal cancer. Individual patient data were sought from all trialists, but for various reasons data were only obtained for seven trials. Again, several outcomes have been examined in this review but only time to death is considered here.

2.5.2. Data extraction

In order to examine the reliability of the estimated summary statistics, as much as possible of the following information was extracted from the published manuscript (or sought from authors) of each trial:

i) $\log(HR_j)$ or HR_j and corresponding variance, standard error or confidence interval if quoted directly
ii) coefficient of treatment effect and corresponding variance from an adjusted or unadjusted Cox proportional hazards model
iii) log-rank test statistic and corresponding p-value
iv) total number randomised and total number of deaths
v) actuarial or Kaplan-Meier survival curve probability estimates at 6 month intervals
vi) whether numbers at risk were given on survival curve or in text
vii) minimum and maximum follow-up times

Tables 2.3 and 2.4 summarise the information available in the trial reports of examples 3 and 4 respectively. Several reports in both examples presented results adjusted for a variety of different covariates. A number of trials indicated that a Cox proportional hazards model had been fitted but a treatment coefficient and standard error was only presented in one trial. Similarly, although a log-rank test was performed in most trials, the test statistic was not quoted in any. The general quality of published survival curves varied. Survival probabilities were extracted with attention to accuracy, although results are obviously approximate. Where information for minimum and maximum follow-up times could not be extracted or approximated, indirect method 3a could not be used.

Table 2.3. Example 3: Summary of information available in each trial.

Trial	1	2	3	4	5	6	7	8	9	10	11
Hazard Ratio	-	-	-	Adjusted and unadjusted with 95% CIs from Cox models	-	Coefficient and standard error in adjusted Cox model	Adjusted with 95% CI from Cox model	Adjusted with 95% CI from Cox model	-	-	-
Number randomised	63	81	58	49	80	83	126	46	65	85	75
Number of deaths	11	17	23	16	19	23	16	11	30	33	29
Logrank[1] test p-value	0.74	0.50	Reported as 'NS'	Reported as '>0.2'	0.03	0.617	-	Reported as '<0.02'	-	-	-
Survival curves	Actuarial	Kaplan-Meier	Kaplan-Meier	Adjusted Kaplan-Meier	Kaplan-Meier	Kaplan-Meier	Kaplan-Meier	Kaplan-Meier	-	-	-
Numbers at Risk	No	Yes	Yes	Yes	No	Yes	Yes	Yes	-	-	-
Follow-up	Min and max	Accrual dates	Accrual dates	Mean	Median	Median	Median	Min and max	-	Median	Mean

[1] No trial reported a log-rank statistic, p-value assumed to be two-sided when not stated.

Trial	1	2	3	4	5	6	7	8	9	10	11	12	13
Hazard Ratio	Unadjusted with 95% CI	-	-	Adjusted with p-value from Cox model	-	-	-	Adjusted with 95% CI from Cox model	Adjusted and Unadjusted with 95% CI's from Cox model	Unadjusted with 95% CI from Cox model	-	-	-
Number randomised	100	157	44	279	67	21	54	61	182	163	36	170	57
Number of deaths	85	-	38	194	65	20	53	53	157	-	33	-	57
Log-rank test p-value[1]	0.03	0.0016	-	0.0001	0.919	-	0.0039	-	0.13	'Reported as <0.02'	0.006	'NS'	-
Survival curves	Kaplan-Meier	Kaplan Meier	Actuarial % in text	Kaplan-Meier	Kaplan-Meier	-	Kaplan Meier	-	Kaplan-Meier	Kaplan-Meier	Kaplan Meier	Kaplan Meier*	Kaplan Meier
Numbers at Risk	Yes	No	Yes	No	Yes	-	Yes	-	Yes	Yes	No	Yes*	Yes
Follow-up	Min and max	Median	-	Median	Accrual dates	Min and Max	Min and Max	-	Min and Max	Mean	Accrual dates	Accrual dates	-

[1] No trial reported a log-rank statistic, p-value assumed to be two-sided when not stated.
* Kaplan-Meier survival curve and 'numbers at risk curve' provided by trialist.

Table 2.4. Example 4: Summary of information available in each trial.

2.5.3. Estimation of summary statistics

Tables 2.5 and 2.6 show the estimates of $\log(HR_j)$ and its standard error for each trial derived from the methods described earlier (section 2.3) for each example respectively.

Example 3

Insufficient information was given in the abstract of trial 9, 10 and 11 to estimate $\log(HR_j)$ and its variance. Comparing estimates from direct and indirect methods was not possible as only one adjusted direct estimate was given. The agreement between estimates from indirect methods 2 and 3 is varied with no particular pattern to the direction of differences. For trial 8, estimates from indirect method 2 and method 3b suggest a significant benefit for TIPS, whilst estimates from indirect method 3a incorrectly suggest there is no significant difference between the two groups. Some of the intervals chosen for indirect method 3a did not contain any events, which may explain the result obtained.

There was generally good agreement between the three estimates given by (2.4), (2.5) and (2.6). Since the amount of censoring was relatively large in each trial, and the sample size in both treatment groups were approximately equal, the estimates obtained using approximation (2.4) were used in the meta-analysis as recommended by Collette *et al* [35]. The estimates obtained using indirect methods 3a or 3b differed somewhat, and were certainly less consistent with each other than the estimates obtained from indirect method 2.

Example 4

IPD were available for trials 1, 4, 6, 7, 8, 9 and 10. There is generally good agreement between the IPD estimates and those obtained from indirect method 1 or 2. Less agreement is seen between IPD estimates and those obtained using indirect method 3a but method 3b performs much better and estimates from this approach are closer to IPD estimates.

Table 2.5. Example 3: estimates of $\log(HR_j)$ and corresponding standard error

Trial	Direct	Indirect 1		Indirect 2			Indirect 3	
	Adjusted	Unadjusted	Adjusted	2.4	2.5	2.6	3a	3b
1				**-0.2001** 0.6030	**-0.2001** 0.6031	**-0.2009** 0.6056	**-0.7807** 0.6918	
2				**0.3272** 0.4851	**0.3276** 0.4857	**0.3277** 0.4859	**0.1441** 0.4821	**0.2095** 0.5177
3							**0.2203** 0.3934	**0.2083** 0.4193
4		**0.0198** 0.5002	**-0.6349** 0.5409				**0.0645**[1] 0.6674	**-0.6143**[1] 0.4999
5				**0.9957** 0.4588	**0.9960** 0.4590	**1.0321** 0.4756	**0.9730** 0.4138	
6	**0.0920** 0.4879			**0.2086** 0.4170	**0.2086** 0.4170	**0.2088** 0.4174	**0.2133** 0.4369	**-0.1846** 0.4346
7			**0.2469** 0.6051				**-0.1579** 0.4128	**-0.2025** 0.5004
8			**-1.8326** 0.5194	**-1.4028** 0.6033	**-1.4054** 0.6036	**-1.7681** 0.6770	**-0.4490** 0.7480	**-1.3035** 0.6372
9								
10								
11								

[1] Estimate obtained from an adjusted Kaplan-Meier survival curve.

Table 2.6. Example 4: estimates of log(HR_j) and corresponding standard error

Trial	IPD	Indirect 1		Indirect 2			Indirect 3	
		Unadjusted	Adjusted	2.4	2.5	2.6	3a	3b
1	**-0.4338**	**-0.5108**		**-0.4708**	**-0.4710**	**-0.4740**	**-0.3267**	**-0.3339**
	0.2159	0.2207		0.2170	0.2170	0.2177	0.1811	0.2221
2								
3								
4	**-0.5964**		**-0.5365**	**-0.5587**	**-0.6391**	**-0.6020**		
	0.1646		0.1631	0.1439	0.1536	0.1490		
5				**0.0252**	**0.0253**	**0.0252**	**0.0884**	**-0.0058**
				0.2480	0.2484	0.2482	0.1669	0.2471
6	**-0.7985**							
	0.4903							
7	**-0.7841**			**-0.7929**	**-0.7940**	**-0.7932**	**-0.6110**	**-0.7554**
	0.2939			0.2748	0.2750	0.2748	0.2274	0.3425
8	**-0.4303**		**-0.4620**					
	0.2871		0.3429					
9	**-0.2217**	**-0.2485**	**-0.3285**	**-0.2417**	**-0.2417**	**-0.2417**	**-0.1590**	**-0.2178**
	0.1533	0.1655	0.1792	0.1597	0.1597	0.1597	0.1196	0.1519
10	**-0.3305**	**-0.4005**					**-0.3247**	**-0.3248**
	0.1631	0.1526					0.1356	0.1664
11				**-0.9567**	**-1.0762**	**-1.0335**	**-0.7723**	
				0.3481	0.3693	0.3619	0.2729	
12							**0.2469**	**0.1753**
							0.1030	0.1615
13							**-0.2743**	
							0.2805	

There was also good agreement between the three estimates given by (2.4), (2.5) and (2.6) with slight discrepancies in trial 4 and 11. The percentage of censoring was relatively low in each trial, therefore estimates obtained using approximation (2.6) were used in the meta-analysis as recommended by Collette *et al* [35]. However, twice as many patients were randomised to chemotherapy in trial 11, therefore the estimates obtained using approximation (2.5) were used. Study recruitment periods and dates of study termination were provided in four trials. Further methods for calculating F_{min} and F_{max} such as using date of publication were required for trials 4, 5, 10, 11 and 12 as sufficient information was not given in the manuscript. Although survival probabilities could be obtained, F_{min} and F_{max} could not be estimated for trial 2 or 3, hence indirect method 3 could not be used to estimate $\log(HR_j)$ and its variance in this case.

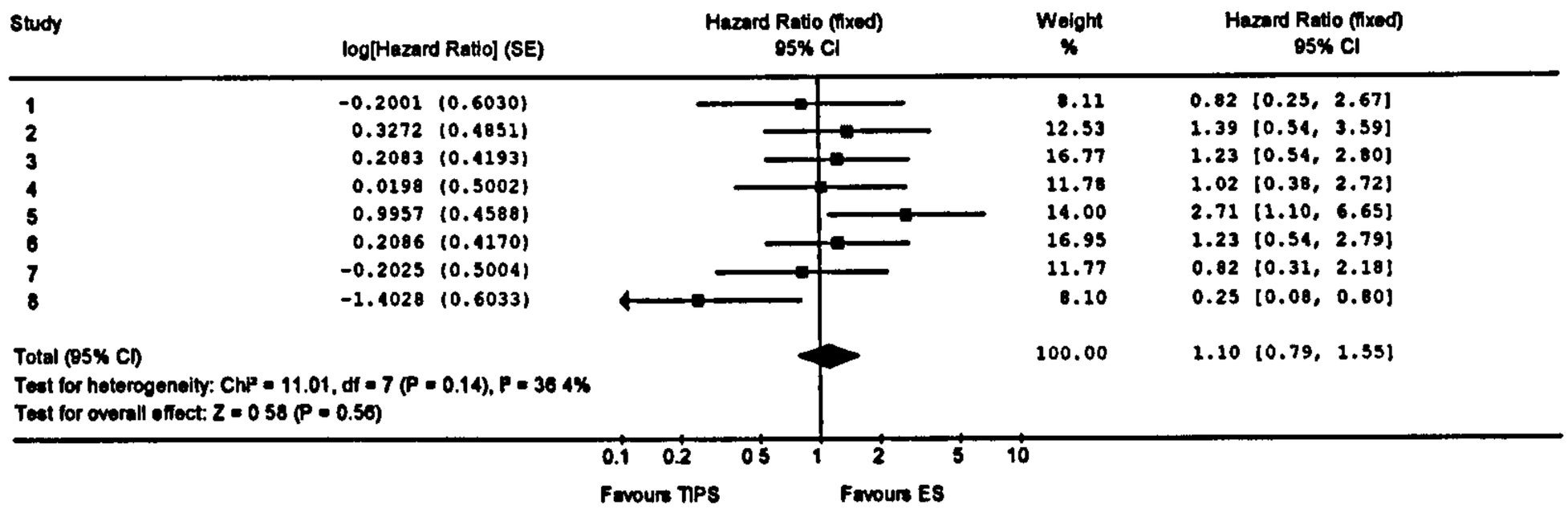
2.5.4. Meta-analysis

In these examples, estimates of $\log(HR_j)$ and its variance to be used in each meta-analysis were chosen according to the following hierarchy. Unadjusted direct estimates are given priority, but in their absence unadjusted estimates obtained from indirect method 1 were used. If indirect method 1 was not appropriate for a particular trial, the estimates obtained using indirect method 2 were used. Finally if these estimates were not available, the estimates obtained from applying indirect method 3 were used with priority given to method 3b which uses additional information regarding the numbers at risk.

Example 3

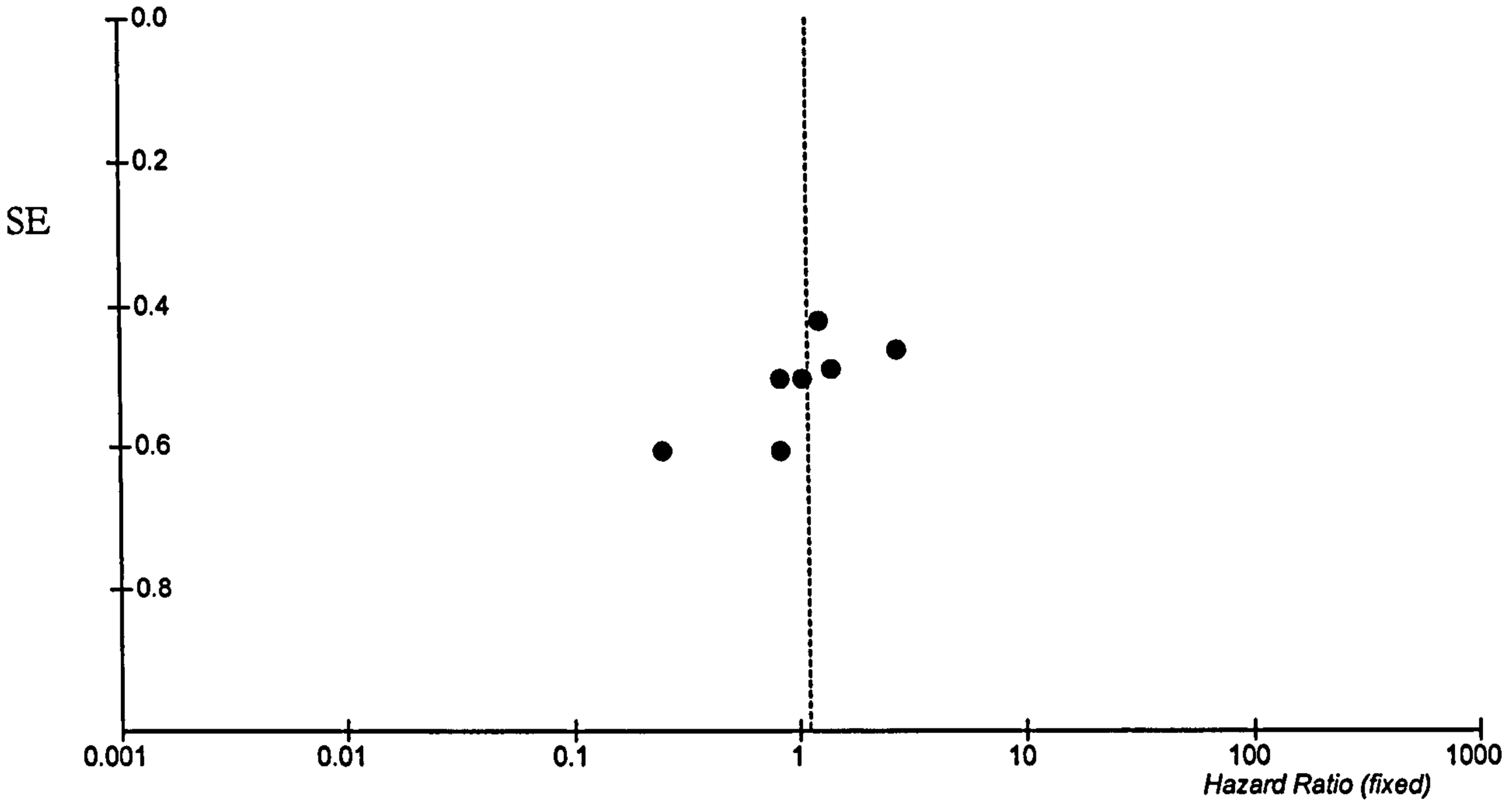
The meta-analysis using unadjusted estimates is displayed in Figure 2.4. Overall there is insufficient evidence to suggest that a treatment difference is present between TIPS and ES, although in the clinical opinion of the review author, important clinical differences with respect to mortality cannot be excluded.

Figure 2.4. Example 3: Meta-analysis of unadjusted results



The test for heterogeneity revealed no statistically significant evidence for heterogeneity between trials (chi-square=11.01, df=7, p=0.14). The funnel plot (Figure 2.5) shows that the smaller trials indicate a potential benefit for TIPS (experimental group), whilst the larger trials indicate a potential benefit for ES (control group). This pattern is consistent with the pattern observed when publication bias is present and therefore may be a potential problem for this example. However, interpreting funnel plots can be very difficult and subjective, particularly when the number of trials is small as in this example. Exploring publication bias is not dealt with further in this thesis.

Figure 2.5. Example 3: Funnel plot using unadjusted estimates



Example 4

Trials were grouped according to whether chemotherapy was administered regionally or systemically. As a combination of IPD and AD were available, three meta-analyses were undertaken as follows.

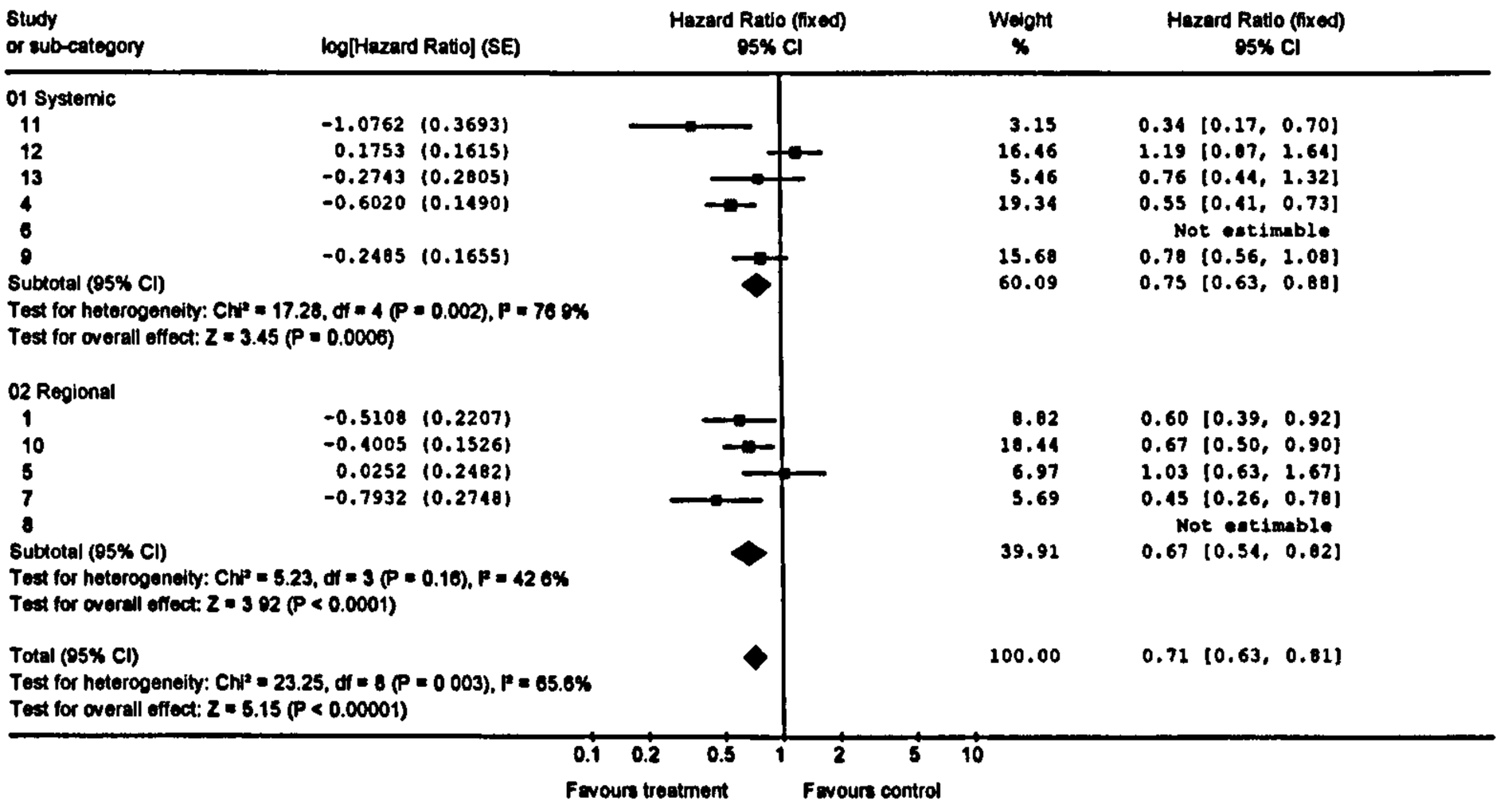
i) Figure 2.6 displays the meta-analysis using AD only, which is based on 9 trials, 1114 randomised patients and 1000 deaths.

ii) Figure 2.7 displays the meta-analysis using IPD only, which is based on 7 trials, 866 randomised patients and 753 deaths.

iii) Figure 2.8 displays the meta-analysis using IPD for the 7 trials where they were available and AD for 4 other trials where no IPD were available but AD estimates could be extracted from reports. This analysis is based on 1196 randomised patients and 1073 deaths.

All analyses suggest that the risk of death in the treatment group is significantly reduced in both the systemic and regional subgroups.

Figure 2.6. Example 4: Meta-analysis using AD estimates only



There is significant evidence of heterogeneity in the systemic group (p=0.002) but not in the regional group (p=0.16) when AD estimates are used (Figure 2.6). However, when only IPD estimates are used (Figure 2.7) there is no significant evidence of heterogeneity (p=0.18 systemic, p=0.61 regional). When both IPD and AD estimates are used (Figure 2.8), there is significant evidence of heterogeneity in the systemic group (p=0.004) but not in the regional group (p=0.32). The relatively large trial 12, with results in the opposite direction to other trials, appears to be the main cause of heterogeneity. There is no evidence for heterogeneity when only IPD estimates are used (Figure 2.7) as trial 12 is not included in this analysis.

Figure 2.7. Example 4: Meta-analysis using IPD estimates only

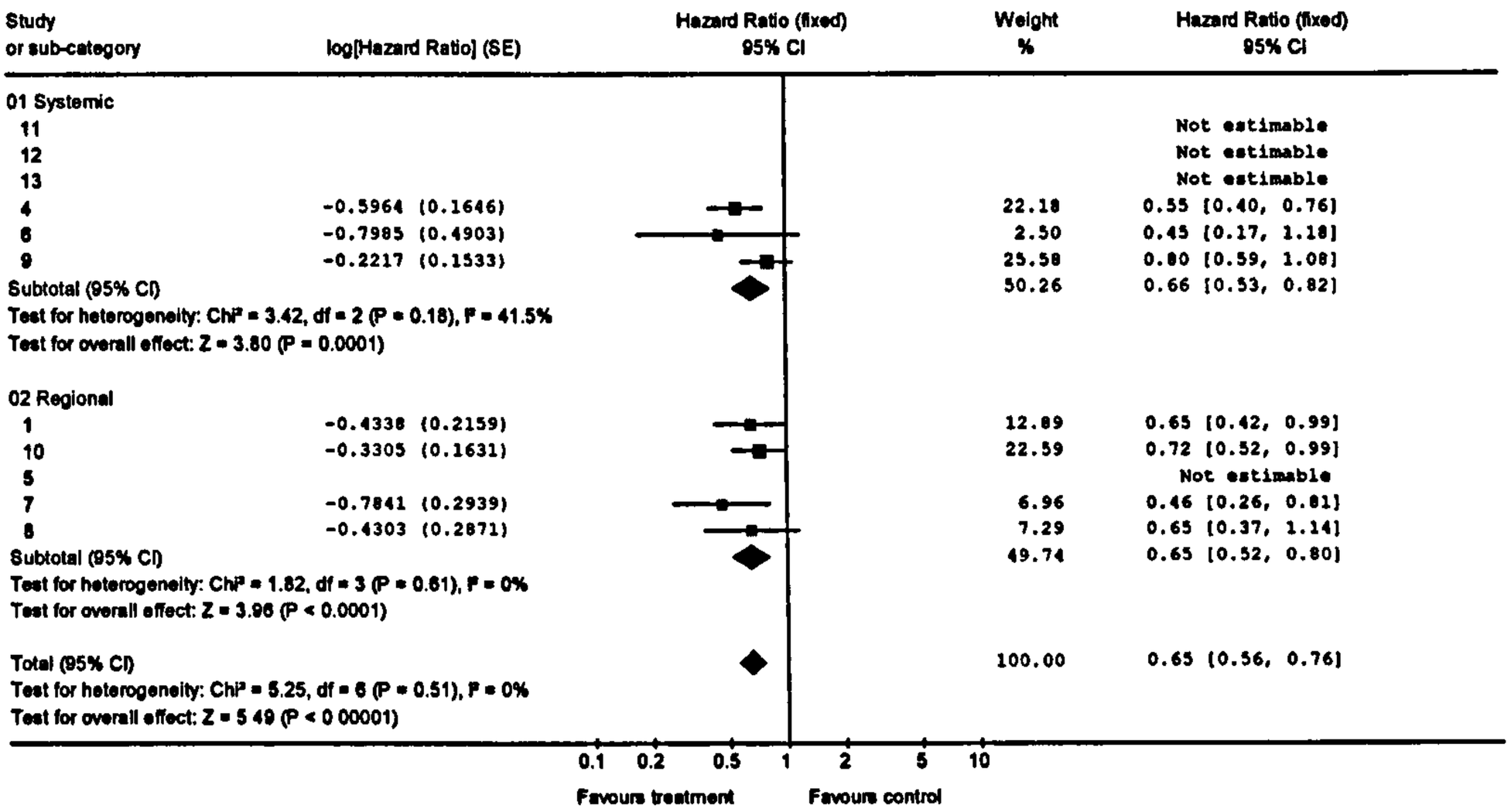


Figure 2.8. Example 4: Meta-analysis using IPD and AD estimates

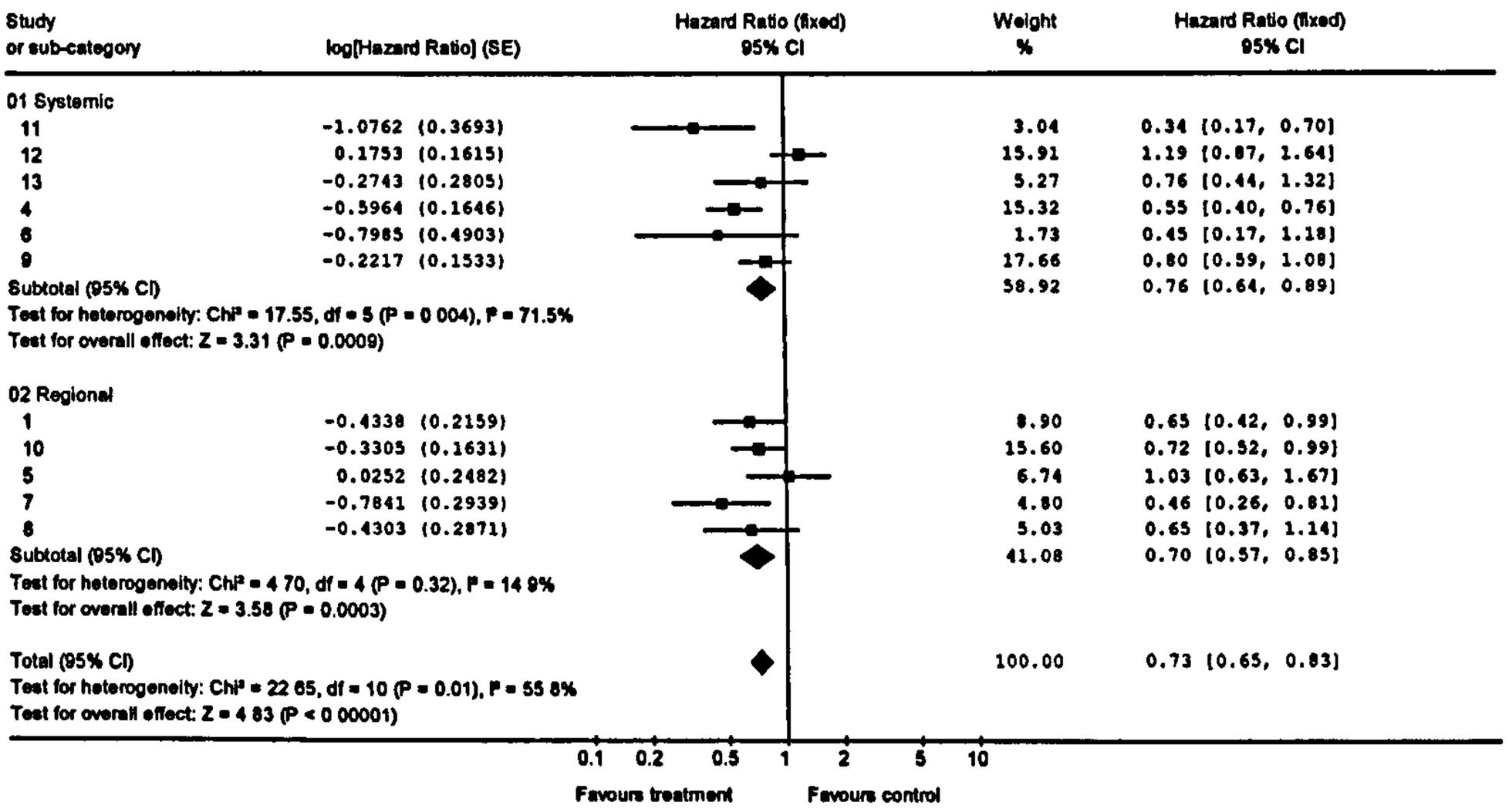
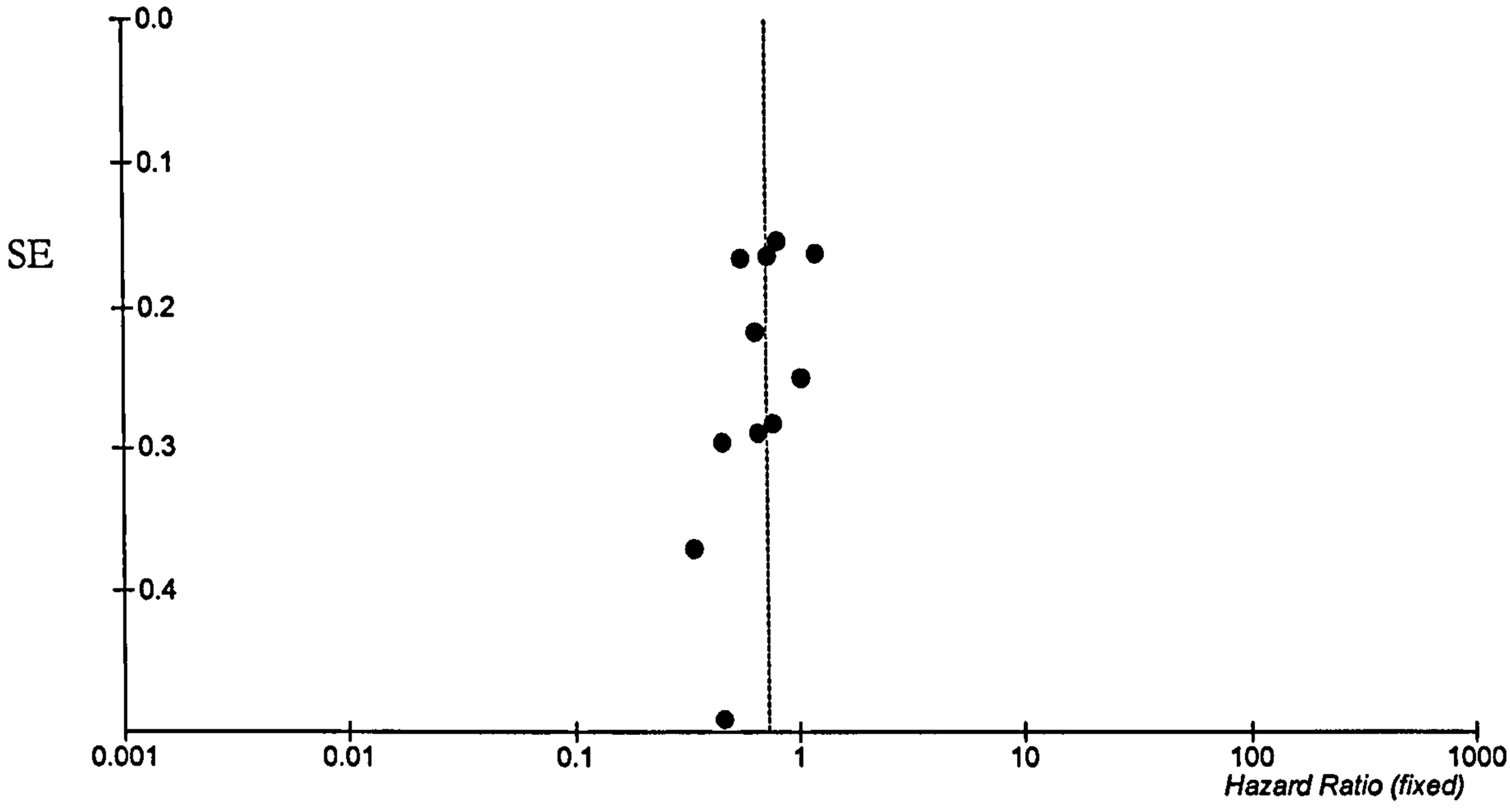


Figure 2.9. Example 4: Funnel plot using IPD and AD unadjusted estimates



The funnel plot in Figure 2.9 using a combination of IPD and AD estimates displays some evidence of asymmetry suggesting that some bias may be present. In particular, it may be that small trials favouring control treatment (supportive care) have not been published. Investigating potential sources of heterogeneity and publication bias is beyond the scope of this investigation. The reviewers are currently trying to contact the authors of trials 2 and 3 but the abstracts suggest that treatment is significantly superior in trial 2 but no difference is evident in trial 3.

In this example, there is good agreement between the meta-analysis based on AD alone (Figure 2.6) and that based on both AD and IPD (Figure 2.8). This similarity is mainly due to the following:

(i) The number of trials, patients and events are very similar in both analyses. The latter analysis includes two additional trials (trials 6 and 9 with a total of 82 randomised patients) for which IPD but not AD estimates were available.

(ii) The AD methods based on survival curves, which are likely to be less reliable, were only used in two trials (trial 12 and 13). As no IPD were available for these trials, the same estimates were used in both analyses.

The meta-analysis based on IPD alone (Figure 2.7) and that based on both AD and IPD (Figure 2.8) do not agree quite so well. This is mainly because the latter analysis includes 4 additional trials (320 events, 330 randomised patients) for which AD but not IPD were available.

Using a combination of AD and IPD has been useful here as (i) the number of trials, patients and events included in the analysis could be maximised (ii) heterogeneity in the systemic subgroup was highlighted which may not have been recognised if only IPD were used, and (iii) we have gained increased confidence in the results obtained in the regional subgroup.

2.6. Assessing the assumption of proportional hazards

If the relative effect of two treatments changes over time, and the trials included in a meta-analysis vary in terms of the length of follow-up, this will introduce heterogeneity which may be evident from graphical tests that detect bias [43]. Methods for assessing the proportional hazards (PH) assumption using IPD have been proposed [44]. The capacity to detect violations of the PH assumption if only aggregate data were available for each trial would be valuable to meta-analysts. Several methods for detecting such violations based on aggregate data are now developed and applied to five randomised trials included in the systematic review comparing CBZ and VPS [39]. If the PH assumption is deemed appropriate for a particular trial, one might expect it to be approximately appropriate for all trials if the treatments under comparison are expected to behave in a similar way across different trial settings considered in the meta-analysis. The principle of meta-analysis suggests that this would be the case for the majority of situations. The assumption may not be realistic if variation across trials introduces some differential effect on one or other treatments such that hazards are no longer proportional within a particular trial. This would suggest an underlying time varying covariate with a differential effect across trials that would induce non-proportional hazards in a selection of trials. This work has been published by Williamson, Tudur Smith *et al* in *Statistics in Medicine* [33].

2.6.1. Overall log(*HR*) estimate only available for each study

If the relative effect of two treatments does not remain constant over time, one would expect the log(HR_j) estimates from trials with differing periods of follow-up to vary. Informally, one could assess plots of the log(HR_j) against length of follow-up. The effect of average follow-up period on the estimate of the treatment effect could be investigated using meta-regression, a procedure which is described in more detail in Chapter 5. This test is likely to have low power in most situations as the number of trials in a meta-analysis may be small and there may be little variation in the overall follow-up time across trials.

For time to first seizure examined in the CBZ/VPS meta-analysis [39], the test for heterogeneity in treatment effect between trials was non-significant using IPD ($\chi^2_{(4)}=5.89$, $p=0.21$) and no obvious trend was evident between the treatment effect estimate and the summary measure of follow-up for each trial.

2.6.2. Log cumulative hazard plot (log-log plot)

A plot of the log cumulative hazard versus time (log-log plot) is a standard graphical tool, which can be used to indicate a violation of the PH assumption using IPD. The logarithm of the survival time is plotted against the estimated log cumulative hazard $(\log[-\log(\hat{s}(t))])$. If the plotted curves for the two treatment groups are approximately parallel, the PH assumption is reasonable. Estimated survival probabilities $s^*_{k,1},....,s^*_{k,p}$ at specific time points $t_1,....,t_p$ extracted from published survival curves can be used to produce an approximate log cumulative hazard plot as a crude method of assessing the plausibility of the PH assumption.

A log cumulative hazard plot for each trial in the CBZ/VPS systematic review using IPD is displayed in Figure 2.10. The corresponding log cumulative hazard plots using aggregate data (survival probabilities read off Kaplan-Meier curves at 200-day intervals) are displayed in Figure 2.11.

In this example, although the plots using IPD (Figure 2.10) are difficult to interpret as the curves are generally quite close together, they suggest that the assumption of PH

may be invalid in at least 3 trials (De Silva 1996 [45], Verity 1995 [46], Mattson 1992 [38]).

Broadly speaking, the plots produced using aggregate data (Figure 2.11) show similar patterns in that two trials favour CBZ (Richens 1994 [47], Mattson 1992 [38]), two trials tend to favour VPS (De Silva 1996 [45], Verity 1995 [46]) and neither drug is particularly favoured in one trial (Heller 1995 [48]). The interpretation is made difficult in this example as the survival curves are generally quite close together. Furthermore, the time-points at which survival probabilities are read off need to be chosen carefully with the first time-point chosen close to the origin (7 days in this example) to avoid missing potentially important information. On the whole, this approach may not be particularly useful and may lead to incorrect interpretations.

Figure 2.10. Log cumulative hazard plots using individual patient data (IPD) for five trials included in CBZ/VPS systematic review.

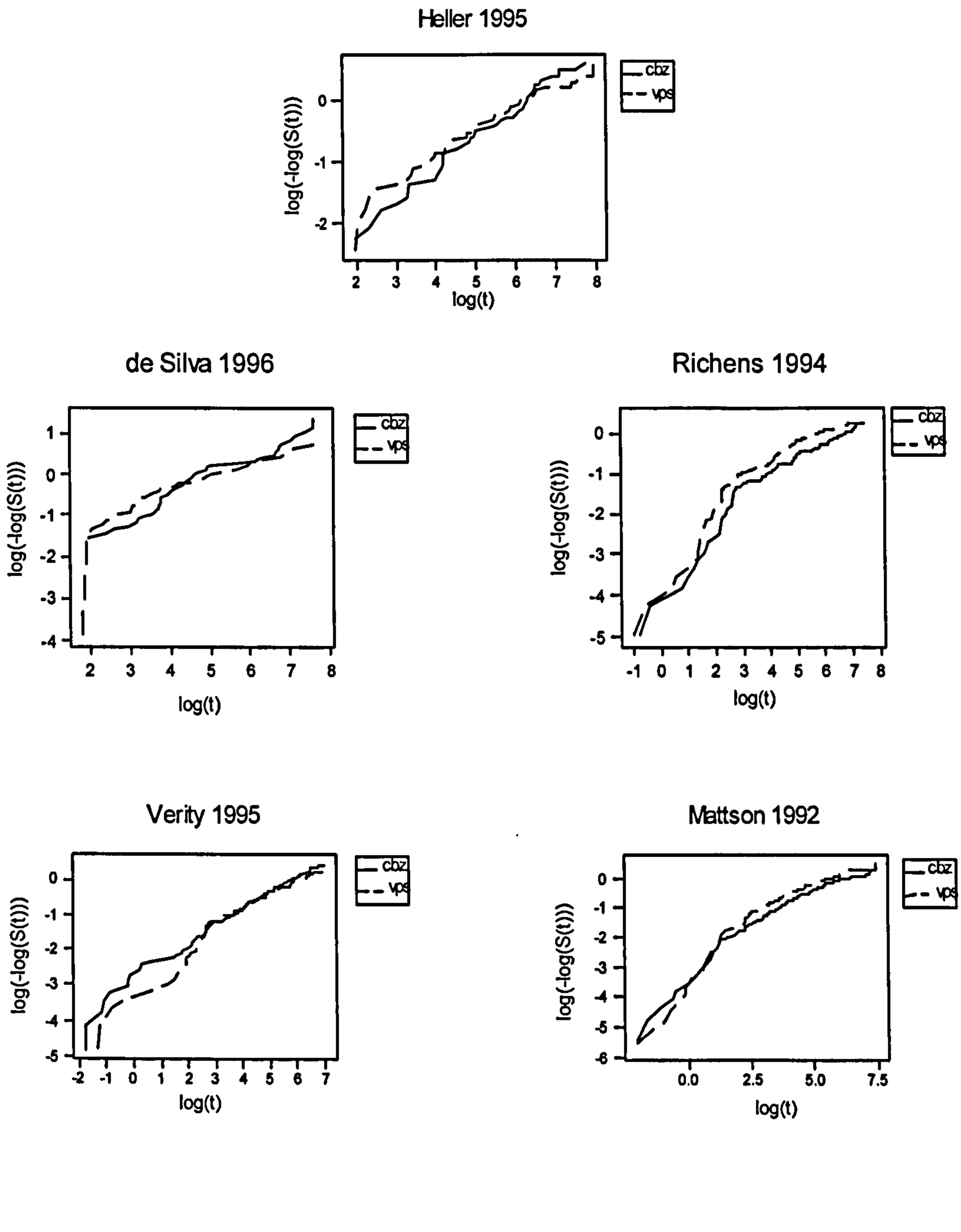
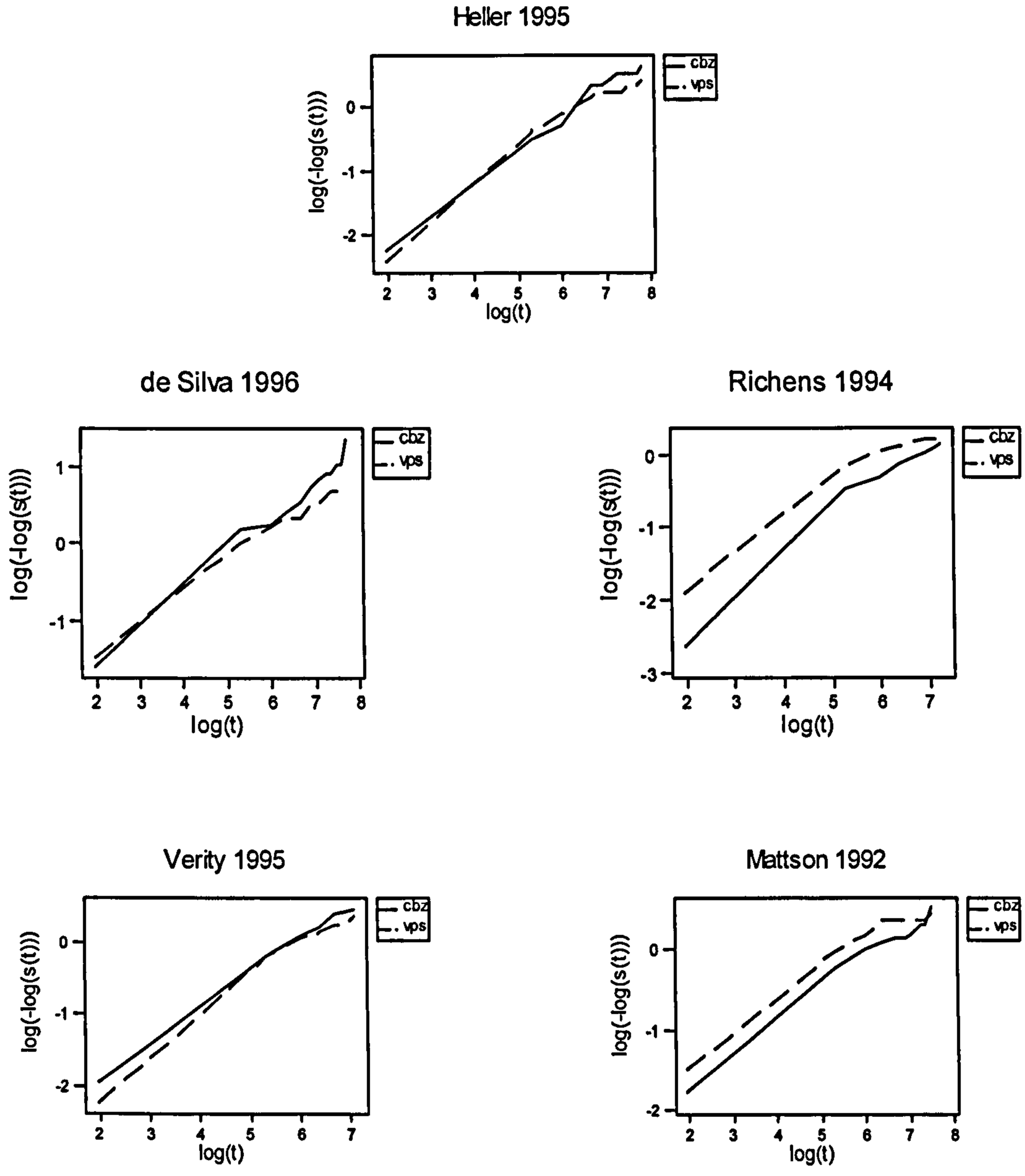


Figure 2.11. Log cumulative hazard plots using aggregate data (AD) for five trials included in CBZ/VPS systematic review.



2.6.3 Estimate of log(*HR*) available for different time intervals

If an estimate of the log hazard ratio and its variance are available for a number of separate intervals within each trial, an estimate of *log(HR)* pooled across trials can be calculated for each interval. If the assumption of PH is reasonable one would expect the pooled estimates of *log(HR)* to remain approximately constant when plotted against time.

To investigate the assumption of PH in such a way, or if the results extracted from survival curves for several trials are to be pooled together to form an overall survival curve [32], then intervals defined by the same time-points $t_1',....,t_r'$ need to be used for all trials. As numbers at risk are unlikely to be reported at the same time points for each trial, a further modification to the method described in section 2.3.5 (method 3b) is required and now described in order to utilise this information.

Using the notation introduced in section 2.3.5, suppose survival probabilities $s^*_{k,1},....,s^*_{k,p}$ are read off the curves at time points $t_1,....,t_p$ where numbers at risk $n_{k,1},....,n_{k,p}$ are given. Within each time interval $[t_{i-1},t_i)$ ($i=1 \ldots p$) the method described in section 2.3.5 is used to obtain estimates of $d^*_{k,i}$, $c^*_{k,i}$, and $n^*_{k,i}$. Furthermore, suppose survival probabilities $s^*_{k,1}',....,s^*_{k,r}'$ are read off the curves at required time points $t_1',....,t_r'$ that are necessarily identical for each trial. Within each time interval $[t_{l-1}',t_l')$ ($l=1 \ldots r$) estimates $d^*_{k,l}'$, $c^*_{k,l}'$, and $n^*_{k,l}'$ are required to calculate *log(HR)* and its variance.

For the sub-interval $[t_{i-1},t_l')$, where $t_{l-1}'<t_{i-1}<t_l'<t_i$, it can be shown that

$$d^*_{k,([i-1,l'))} = 0.5*\left[\frac{s^*_{k,l}'((c^*_{k,i}+2d^*_{k,i})*(s^*_{k,l}'-s^*_{k,i-1})-(s^*_{k,i}c^*_{k,i}))+(s^*_{k,i}s^*_{k,i-1}c^*_{k,i})}{s^*_{k,l}'(s^*_{k,i}-s^*_{k,i-1})}\right] \quad (2.15)$$

$$c^{*}_{k,([i-1,l'))} = \left[\frac{s^{*}_{k,l}{}'((2s^{*}_{k,i}n_{k,i-1}) - s^{*}_{k,i-1}(c^{*}_{k,i} + 2n_{k,i})) - (s^{*}_{k,i}s^{*}_{k,i-1}c^{*}_{k,i})}{s^{*}_{k,l}{}'(s^{*}_{k,i} - s^{*}_{k,i-1})} \right] \quad (2.16)$$

For the sub-interval $[t_l', t_i)$, it can be shown that

$$d^{*}_{k,([l',i))} = 0.5 * \left[\frac{s^{*}_{k,l}{}'((c^{*}_{k,i} + 2d^{*}_{k,i})(s^{*}_{k,i} - s^{*}_{k,l}{}') + (s^{*}_{k,i-1}c^{*}_{k,i})) - (s^{*}_{k,i}s^{*}_{k,i-1}c^{*}_{k,i})}{s^{*}_{k,l}{}'(s^{*}_{k,i} - s^{*}_{k,i-1})} \right] \quad (2.17)$$

$$c^{*}_{k,([l',i))} = -\left[\frac{s^{*}_{k,l}{}'(s^{*}_{k,i}(2n_{k,i-1} - c^{*}_{k,i}) - 2s^{*}_{k,i-1}n_{k,i}) - (s^{*}_{k,i}s^{*}_{k,i-1}c^{*}_{k,i})}{s^{*}_{k,l}{}'(s^{*}_{k,i} - s^{*}_{k,i-1})} \right] \quad (2.18)$$

The estimated number at risk at time point t_l' is then given by

$$n_{k,l}{}' = n_{k,i-1} - d^{*}_{k,([i-1,l'))} - c^{*}_{k,([i-1,l'))} \quad (2.19)$$

Finally for the interval $[t_{l-1}', t_l')$, the *log(HR)* and its variance can be approximated using the following estimates

$$d^{*}_{k,l}{}' = d^{*}_{k,([l-1',i-1))} + d^{*}_{k,([i-1,l'))} \quad (2.20)$$

$$c^{*}_{k,l}{}' = c^{*}_{k,([l-1',i-1))} + c^{*}_{k,([i-1,l'))} \quad (2.21)$$

$$n^{*}_{k,l}{}' = n^{*}_{k,l-1}{}' - \frac{c^{*}_{k,l}{}'}{2} \quad (2.22)$$

The estimate $\log(\hat{HR})_l{}'_{pooled}$ pooled across trials for each interval $[t_{l-1}', t_l')$, (estimated using an inverse variance weighted average) or its exponential can be plotted against time.

A formal significance test could be applied as follows. The null hypothesis of proportional hazards for the effect of treatment can be described by

$$H_0 : \log(H\hat{R})_1{}'_{pooled} = \log(H\hat{R})_2{}'_{pooled} = \cdots = \log(H\hat{R})_r{}'_{pooled} = \hat{\theta}_p$$

where $\hat{\theta}_p$ denotes the overall pooled log hazard ratio across J trials. A suitable test statistic, similar to the usual test of homogeneity of summary statistics across trials, is

$$Q = \sum_{l=1}^{r} w_l{}'_{pooled} [\log(H\hat{R})_l{}'_{pooled} - \hat{\theta}_p]^2$$

where

$$w_l{}'_{pooled} = \text{var}(\log(H\hat{R})_l{}'_{pooled})^{-1}$$

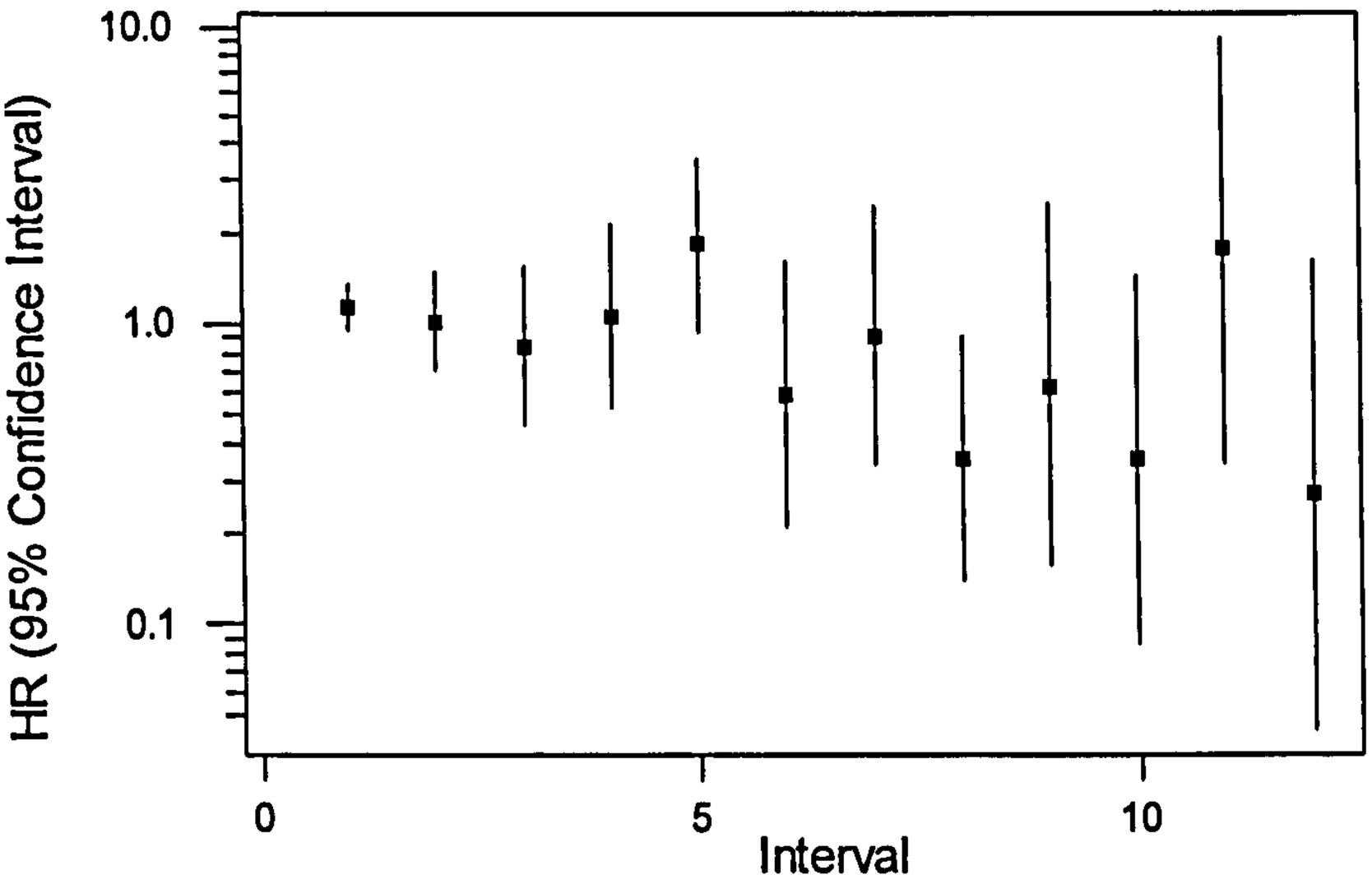
Under the null hypothesis, $Q \sim \chi^2_{r-1}$ asymptotically. The power of this test is likely to be low as the alternative hypothesis is general and an alternative test for trend in the log hazard ratios with time may be more appropriate although not considered further here. The above approach pools the within interval estimates across all trials, an approach which assumes that each trial is estimating similar hazard rates and that the pattern with time is similar across trials. This assumption may be explored further by examination and visual comparison of the individual within trial plots of log hazard ratios and confidence interval across all intervals.

For the trials in the CBZ/VPS example, estimates of *log(HR)* and its variance obtained using the method described above agree well compared to corresponding estimates obtained from directly incorporating numbers at risk for the intervals of interest.

Figure 2.12 shows a plot of the *hazard ratio* within each interval $[t_{l-1}', t_l')$ pooled across trials against time for the CBZ/VPS example using intervals of 90 days within each trial. In this plot, the estimated hazard ratio appears to remain fairly constant as time increases although the variation increases towards the end of the time scale where there is much less information and this is reflected in the increase in confidence interval

width. A Q value of 16.21 (11 degrees of freedom) provides no statistical evidence against this observation (p=0.13). Similar analyses were undertaken using interval widths of 30 days over ranges of 0-1080 days and 0-720 days resulting in Q values of 34.24 (35 degrees of freedom, p=0.51) and 25.17 (23 degrees of freedom, p=0.34) respectively. These results suggest that assuming proportional hazards may not be unreasonable in this example.

Figure 2.12. Hazard Ratio pooled across trials within each interval (90-day) plotted against time for 5 trials in CBZ/VPS systematic review



2.7. Discussion

Methods for undertaking meta-analysis with time-to-event outcomes using aggregate data extracted from trial reports or made available from trialists are accessible but their value is limited by the lack of accurate reporting of suitable data. As an example, the p-value quoted for the log-rank test should be reported to at least 2 decimal places and care is required when interpreting the direction of effect for the log hazard ratio which may not always be obvious especially when the two treatment groups have similar effects on the outcome (Tudur *et al.* [34]). The approaches based on extracting survival probabilities from published survival curves may be particularly prone to bias introduced by inter-reader variability extracting data, highlighting the importance of having at least two independent reviewers to undertake data extraction. The quality of published survival curves are likely to vary considerably. Some improvement in data extraction accuracy may be gained by enlarging the published plots or using specialist software developed for this purpose. The current author has not examined the accuracy of estimates extracted electronically which may be interesting to investigate in future. The survival curve based approach that ignores information from numbers at risk (method 3a) requires some assumption to be made about the pattern of censoring across the entire follow-up period. Less severe assumptions are made in the approach incorporating numbers at risk (method 3b) which improves accuracy of estimates for the empirical examples examined herein. Further empirical examples are required to validate these observations, particularly in comparison to results obtained from individual patient data. However, such comparisons are likely to be complicated by a number of differences between the data used in IPD and AD analyses [49]. In addition, for many empirical examples IPD may be sought for the very reason that AD are unavailable from trial reports. A pragmatic comparison of results from these approaches may therefore be limited and a simulation study may be useful to help quantify bias and complement empirical results. In practice, the biases are likely to vary according to example and reviewers should be made aware of potential pitfalls with AD meta-analysis of time-to-event outcomes. The author is of the opinion that for AD based meta-analyses, data should be extracted to enable as many of the methods as possible for estimating log hazard ratio and its variance to be utilised and compared. This would allow consistency to be explored and could identify potential areas for concern. If estimates from all approaches are available the author would recommend that the

hierarchy of indirect method 1, indirect method 2, indirect method 3b and indirect method 3a be used for choosing which estimate to use for meta-analysis.

One advantage of collecting IPD for meta-analyses involving time-to-event outcomes is the potential to examine the underlying assumption of proportional hazards which underpins the interpretation of the hazard ratio. Methods for assessing this assumption using aggregate data have been proposed in this chapter. Evidence for non-proportional hazards was not detected by assessing the extent of variability in treatment effect according to length of follow-up. This formal statistical test of homogeneity has low power when the number of trials is small, as in the example used for illustration. A log cumulative hazard plot can be constructed using aggregate data although this may be difficult to interpret, in a similar way to the difficulty of interpretation in a single survival study. A further graphical display of the estimated hazard ratio pooled across trials within contiguous intervals, and a statistical test similar to the chi-square test of homogeneity, failed to demonstrate non-proportional hazards. To improve estimation, the time-points in each trial should be chosen such that the number of events are non-zero for each trial within a particular interval. Results should also be interpreted with caution as the choice of interval width could potentially alter conclusions. Furthermore, the power of the test is likely to be low as the alternative hypothesis is general. A more specific alternative hypothesis such as a linear trend with time may be more suitable. More in-depth exploration of the proposed approaches for assessing proportional hazards with AD is required to assess reliability. If the assumption does not hold or there are clear clinical reasons to expect the relative effect to change with time the availability of IPD will be particularly beneficial as models that do not require making the PH assumption may be explored and may be more appropriate. The availability of IPD provides greater flexibility to further explore alternative model structures.

It is clear that further evidence is needed to establish in which situations and areas of health care an IPD approach is most beneficial and, conversely, to characterise when an AD approach will be reliable. Specific issues relating to IPD analyses are detailed in following sections. Williamson, Marson, Tudur *et al* [49] have undertaken a review of the empirical evidence related to the comparison of IPD and AD meta-analyses, concluding that more evidence is needed. However larger scale IPD projects may not always be feasible or the data may not be available and the AD based methods outlined in this

chapter may be particularly valuable in these situations. In terms of calculating an overall pooled hazard ratio and confidence interval, the present author is of the opinion that an aggregate failure time data meta-analysis will be adequate under the following conditions: the outcome is well and consistently defined across trials, the AD presented in all included trials consists of a direct estimate of log hazard ratio and its variance, SE or confidence interval; the within-trial analyses include data for all randomised individuals following an intention to treat approach; a clear description is provided of the model adopted, covariates included, and direction of treatment comparison. Deviations from these conditions are likely to introduce some bias, and should be highlighted when interpreting the results and drawing conclusions from AD based meta-analyses.

CHAPTER 3

Individual patient data meta-analysis

Systematic reviews which include a meta-analysis of individual patient data (IPD) have been described as the 'yardstick' against which all systematic reviews should be measured [25]. An approach of this kind can be resource intensive as the process of requesting, collecting, organising and cleaning data to ensure a standard format, can be lengthy and expensive compared with the traditional approach based on extracting data from trial reports or supplied by authors of unpublished articles. Although an IPD approach may require greater resources compared with an aggregate data approach, there are several advantages particularly in systematic reviews that focus on time-to-event outcomes. As described in Chapter 2, limitations of reporting suitable aggregate data for time-to-event outcomes often preclude an aggregate data meta-analysis of this type of outcome. Having IPD from each trial permits the standardisation of outcome definitions and the possibility of analysing previously unreported outcomes. If multiple outcomes are of interest but only a subset are reported in a particular trial, it could be that arriving at the original decision of which outcomes to report was based on presenting the most significant. Without IPD, the trial in question would be excluded if aggregate data for the outcome were not reported and this is likely to introduce bias to the meta-analysis. Further advantages include the possibility of undertaking a more complete and updated analysis as follow-up data collected after the original trial publication can be included in the meta-analysis and re-analyses based on all randomised

patients to achieve an intention to treat analysis can be undertaken with IPD. In addition, the availability of IPD allows more thorough data validation and quality assessments to be undertaken. Many further benefits such as identification of further unpublished trials can also be gained as a result of collaborating with original trial authors in the relevant clinical field. The practicalities, advantages and disadvantages of undertaking meta-analyses based on IPD are described in detail by Stewart and Clarke [26], Clarke *et al* [50], and the Early Breast Cancer Trialists' Collaborative Group [51].

Although IPD meta-analyses are considered the gold-standard, it is important to establish circumstances in which such an approach is worthwhile and warrants the extra investment of time and money that is inevitably required. However, the empirical evidence comparing the main meta-analysis results and clinical interpretations obtained from systematic reviews using AD compared with IPD, is largely inconclusive with respect to establishing how much gain is to be achieved with IPD. This is mainly because various sources of bias are introduced at different levels making a straightforward comparison problematic. For example, a 'gold-standard' AD meta-analysis would require the correct AD to be reported accurately and completely for every eligible trial. However, in practice, an AD approach may be more susceptible to publication bias, within study selective reporting bias, and biases associated with inaccurate or incomplete reporting of the required data, which hinders a comparison of 'gold-standard' AD and IPD meta-analyses. In most circumstances, such a comparison would require using IPD to generate AD. Obtaining additional follow-up information is likely to be more successful for certain end-points that are commonly examined in particular clinical areas. For example, time to death analyses are common in cancer trials and information on whether a patient has died following the end of a clinical trial can be collected through ONS, making up to date analyses possible. On the other hand, seizure data following the end of trial in epilepsy patients is rarely collected reliably which is likely to prevent 'further follow-up' analyses in these trials. Comparisons between IPD and AD may therefore differ according to specific clinical areas. Further differences in outcome definition, patient exclusions imposed by original authors, and restricted reported data make a pragmatic comparison between AD and IPD difficult to interpret. Such biases and differences are likely to be inconsistent across meta-analyses making overall conclusions hard to reach.

Alternative methods for IPD meta-analysis of time-to-event outcomes have been established and utilised in practice. Two-stage methods, that involve using the IPD to estimate the treatment effect and variance within trial as the first stage then pooling across trials as the second stage, are the most commonly used approach to analysis (Mark Simmonds, personal communication). Examples of two-stage methods for time-to-event outcomes include calculating an inverse variance weighted average of within trial log hazard ratios or a stratified Log-rank analysis. In particular, the stratified log-rank analysis is the method of analysis adopted by the SCHARP software, a freely available package for undertaking IPD meta-analysis developed at the Medical Research Council's Cancer Trials Unit in collaboration with the Istituto di Ricerche Farmacologiche. In the authors opinion, there is a need to compare such approaches to ascertain whether the most appropriate techniques are adopted and highlight potential areas where methods may be preferred. Common methods for meta-analysis are compared in later sections using simulated data.

3.1. Comparison of IPD and AD based meta-analysis

A review of systematic reviews which have included a comparison of the main treatment effect results from IPD and AD meta-analyses has been undertaken (published by Williamson, Marson, Tudur *et al* in JECP [49]). Inclusion criteria for this review were systematic reviews of randomised controlled trials that include both an IPD meta-analysis and a comparative AD meta-analysis. The primary end-point of interest in each review, as defined by the original reviewers, was examined. Electronic databases MEDLINE, BIDS, The Cochrane Database of Systematic Reviews and The Cochrane Review Methodology Database were searched. Proceedings from the Cochrane colloquium and Oxford Symposium on Systematic Reviews were reviewed for all years. Experts in the field were contacted and asked to identify further unpublished studies. Seven published systematic reviews meeting the inclusion criteria were identified where the main treatment effect could be estimated from a meta-analysis of the literature (MAL) and meta-analysis of individual patient data (MAP). One further unpublished comparison could be included using available data from the systematic review of monotherapy trials comparing CBZ and VPS for epilepsy. Results for the comparison of main treatment effect from MAL and MAP approaches are displayed in Table 3.1.

Table 3.1. Empirical evidence of the comparison between meta-analysis methods. Table entries relate to the pooled treatment effect and 95% confidence interval. HR=Hazard Ratio, sdiff=difference in survival probabilities at 30 months, OR=odds ratio, RR=relative risk

Condition, Outcome, Intervention	Gold standard IPD (MAP)	Literature-based (MAL)
Epilepsy, time to withdrawal of treatment, Carbamazepine versus Sodium Valproate [49]	HR=0.97 (0.79, 1.18) N=1195	HR=1.02 (0.67, 1.56) N=705
Ovarian cancer, time to death, non-platinum single versus platinum combination chemotherapy [52]	HR=0.93 (0.83, 1.05) sdiff=0.025 N=1329	OR=0.71[ϕ] (0.52, 0.96) sdiff=0.075 N=788
Lung cancer, time to death, chemotherapy versus chemotherapy plus radiotherapy [53]	HR=0.86 (0.78, 0.94) N=2140	OR=0.65[φ] (0.57, 0.77) N=1911
Breast cancer, time to death, ovarian ablation versus control [54]	OR=0.76 (0.65, 0.88) N=1746	OR=0.86 (0.68, 1.07) N=1644
Non-small cell lung cancer, time to death (Parmar *et al.* personal communication)	HR=1.15 (1.04, 1.27) N=2145	HR=1.09 (0.98, 1.22) N=1927
Colorectal cancer, time to death, palliative chemotherapy versus supportive care [34]	HR=0.65 (0.56, 0.76) N=866	HR=0.74 (0.66, 0.84) N=1196
Myocardial infarction, death, ACE inhibitor versus control [55]	OR=0.93 (0.89, 0.98) N=98496	OR=0.93 (0.89, 0.98) N=96669
Recurrent miscarriage, livebirth, immunotherapy versus control [56]	RR=1.12 (0.97, 1.31) N=379	RR=1.29 (1.03, 1.60) N=202

[ϕ] Odds ratio rather than hazard ratio estimated from published survival curves. Both were translated into difference in survival probabilities
[φ] Sub-optimal MAL in that non-randomised patients were included

The level of statistical significance, and estimate of the treatment effect obtained from the two approaches varied but the differences were not consistent across reviews. In two reviews a significant result was observed using MAP but a non-significant result obtained using MAL. A conflicting pattern was seen in one further review with MAL suggesting a significant difference not identified by MAP. Agreement from both

approaches in terms of statistical significance were obtained for the remaining reviews examined. The treatment effect was greater using MAL in three reviews, greater for MAP in three further reviews and similar or equal estimates from both approaches in two reviews. This review of empirical comparisons indicates that different results and conclusions may be obtained from the two approaches to meta-analysis but no systematic pattern of differences can be established. Differences between approaches may be explained by differences in method of analysis, publication bias, exclusion of patients, and follow-up. Williamson, Marson, Tudur *et al* [49] propose that the following specific analyses should be undertaken when comparing AD and IPD to overcome these difficulties. Firstly, a meta-analysis of literature based AD (MAL) would represent the least resource intensive approach. Secondly, IPD from the same trials using identical patient and follow-up information would provide a 'method of analysis' comparison. To further quantify the effect of publication bias, any unpublished trial results obtainable from adopting an IPD approach should be incorporated. The effect of patient exclusion in published analyses can be assessed by using IPD to reinstate any originally excluded patients. The impact of further follow-up can be examined by including this additional level of data before undertaking a 'gold-standard' IPD analysis (MAP, meta-analysis based on individual patient data) incorporating all available trials and patients mentioned previously. These analyses provide a range of comparisons between the least (MAL, meta-analysis of the literature) and most (MAP) resource-intensive approaches.

Further empirical evidence with more specific comparisons are needed in order to establish whether the extra investment needed for IPD over and above AD is worthwhile. A systematic review of empirical comparisons for the main treatment effect [57] is currently underway within the Cochrane Collaboration. Until further evidence is available, the decision of whether the extra investment required for IPD analyses is worthwhile remains a decision in which several factors should be considered.

Several statistical methods are available for the analysis of time-to-event data for a single randomised controlled trial. The remaining chapters of this thesis will primarily consider the log-rank analysis and Cox proportional hazards model as these popular methods are widely accepted as standard for the analysis of time-to-event data. These general approaches are also likely to be the most readily accessible to researchers undertaking meta-analysis of this type of data.

3.2. Log-rank analysis

In an individual trial involving a time-to-event outcome, the log-rank analysis proposed by Mantel and Haenszel [28] is a common non-parametric approach for comparing the time-to-event experiences of two or more groups of individuals. For two treatment groups A and B, the general procedure entails ordering the distinct event times $t_{(k)}$ (k=1,2,...,r) across all groups and recording at each time the number of events for each group, denoted $d_{A(k)}$ and $d_{B(k)}$, and the number at risk for each group, denoted $n_{A(k)}$ and $n_{B(k)}$. At each event time $t_{(k)}$, a 2x2 table may be constructed as follows:

Group	Number of events at $t_{(k)}$	Number of non events at $t_{(k)}$	Number at risk just before $t_{(k)}$
A	$d_{A(k)}$	$n_{A(k)} - d_{A(k)}$	$n_{A(k)}$
B	$d_{B(k)}$	$n_{B(k)} - d_{B(k)}$	$n_{B(k)}$
Total	$d_{(k)}$	$n_{(k)} - d_{(k)}$	$n_{(k)}$

The null hypothesis of no difference in the time-to-event experiences of individuals in the two groups implies independence of event status and group in the table above. Furthermore, as $d_{A(k)}$ follows a hypergeometric distribution, the expected value and variance of $d_{A(k)}$ is given by

$$e_{A(k)} = \frac{n_{A(k)} d_{(k)}}{n_{(k)}}, \qquad v_{A(k)} = \frac{n_{A(k)} n_{B(k)} d_{(k)} (n_{(k)} - d_{(k)})}{n_{(k)}^{2} (n_{(k)} - 1)}$$

Evidence against the null hypothesis can be assessed by combining across all *r* event times, the difference between observed and expected numbers of events under the null hypothesis. The log-rank statistic given by

$$U_{LR} = \sum_{k=1}^{r} \left[d_{A(k)} - e_{A(k)} \right] \qquad (3.1)$$

has an approximate normal distribution with expected value of zero and variance

$$V_{LR} = \sum_{k=1}^{r} v_{A(k)} \qquad (3.2)$$

It follows that $\frac{U_{LR}}{\sqrt{V_{LR}}} \sim N(0,1)$ and therefore $\frac{U_{LR}^2}{V_{LR}} \sim \chi_1^2$ which provides a measure of the extent to which the observed time to event experiences in the two groups deviate from those under the null hypothesis of no difference, with large values indicating greater evidence against the null hypothesis. The approach assumes that the times to event for individuals are independent and identically distributed although no particular distribution is assumed. Further assumptions are that individuals are a random sample from the population of interest, any censoring that occurs is random, and the distribution of censoring times is independent of the time-to-event for individuals in the sample.

Now consider a meta-analysis. In order to preserve the randomisation within each trial in a meta-analysis involving IPD, a log-rank analysis stratified by trial is used to obtain an overall estimate of the hazard ratio and confidence interval. If the subscript *j* denotes trial, where *j=1...J,* the general procedure is to obtain the log-rank statistic and its variance for each trial, denoted U_{LRj} and V_{LRj} respectively, and sum these values over all trials included in the meta-analysis. To the author's knowledge, this approach was first described for undertaking meta-analysis of time-to-event individual patient data by the Early Breast Cancer Trialists' Collaborative Group [51] and is noted to be of maximal statistical sensitivity for the detection of modest treatment effects. An estimate of the typical treatment effect is given by summing the log-rank statistic over trials and dividing by the sum of the log-rank variance across trials (3.3) with variance for the treatment effect estimated by the reciprocal of the sum of the log-rank variance across

trials (3.4). In the 1990 manuscript, this treatment effect is called a "typical log odds ratio" but is referred to as the pooled log hazard ratio (β_{LR}) in this thesis.

$$\hat{\beta}_{LR} = \frac{\sum_{j=1}^{J} U_{LRj}}{\sum_{j=1}^{J} V_{LRj}} \qquad (3.3)$$

$$\mathrm{var}(\hat{\beta}_{LR}) = \frac{1}{\sum_{j=1}^{J} V_{LRj}} \qquad (3.4)$$

The estimate of β_{LR} is described as the "one-step" estimator since it is equal to the first step from a *log(hazard ratio)* of zero towards the maximum likelihood estimator in the Newton-Raphson iterative procedure. It is also referred to as the Peto estimator as it was first described by Yusuf, Peto *et al* [9]. Berry *et al* [58] note that the Peto estimator was originally proposed within the context of meta-analysis where the parameter of interest is an odds ratio rather than a hazard ratio.

3.3. The Cox proportional hazards model

The semi-parametric proportional hazards model proposed by Cox [30] is easily the most widely used regression model for the analysis of failure time data as there is no requirement to make any parametric assumption regarding the baseline hazard rate. The model is written in terms of hazard functions such that for the *ith* individual, i=1,2,...,n, in a single trial with covariate value x_i, the hazard function at time t is written as

$$\lambda_i(t) = \lambda_0(t)\exp(\beta x_i) \qquad (3.5)$$

If x_i is an indicator variable representing treatment group membership such that $x_i = 1$ for patients on the experimental treatment and $x_i = 0$ for patients on control treatment, $\lambda_0(t)$ is the hazard function for an individual on control treatment and is often referred to as the baseline hazard function whilst the hazard for an individual on experimental

treatment is given by $\lambda_0(t)\exp(\beta)$. The regression coefficient β is estimated by maximising the logarithm of the likelihood function given by

$$\log L(\beta) = \sum_{i=1}^{n} \delta_i \left\{ \beta x_i - \log \sum_{l \in R(t_i)} \exp(\beta x_l) \right\} \qquad (3.6)$$

where δ_i is a censoring indicator variable which takes the value 0 if the time to event for the *ith* individual is censored and unity if the event of interest was observed, and $R(t_i)$ represents the risk set, the set of individuals at risk of the event just prior to time t_i. As the likelihood function does not directly use the actual failure times of individuals the function is not a true likelihood function and is referred to as the partial likelihood function. Maximisation of the partial log likelihood function (3.6) is undertaken using the Newton-Raphson numerical procedure such that an estimate of β at step (h+1) of the iterative procedure is given by

$$\hat{\beta}_{h+1} = \hat{\beta}_h + I^{-1}(\hat{\beta}_h) U(\hat{\beta}_h)$$

where $U(\beta) = \frac{\delta \log L(\beta)}{\delta \beta}$ and $I(\beta) = -\frac{\delta^2 \log L(\beta)}{\delta \beta^2}$. An initial value of zero is usually taken and the iterative procedure is considered to have converged when the difference in partial log-likelihood function is sufficiently small. The standard error of $\hat{\beta}$ is approximated by $\sqrt{I^{-1}(\hat{\beta})}$.

Since meta-analysis of individual patient data involves combining individual patient responses across each trial, whereby patients within a trial are assumed to be more alike than patients from different trials, the structure of the data are naturally hierarchical with patients treated as *level-1 units* and trials as *level-2 units*. This 2-level hierarchical structure is adopted for the models considered throughout this paper. Recent advances in the meta-analysis of IPD have explored models for the analysis of binary data [59], continuous data [60], and ordinal data [61], using a hierarchical modelling approach. Bayesian and non-bayesian hierarchical models have been developed for time to event outcomes, and their application is particularly noticeable in the literature of multi-centre

clinical trial analysis [62], [63], [64], [65]. Although the hierarchical data structure is the same for meta-analysis of IPD and a multi-centre clinical trial, the number of *level-1 units* (patients within each trial/centre respectively) and *level-2 units* (trials/centres) is often quite different. Typically, meta-analyses are based on small numbers of trials with many patients, whereas multi-centre clinical trials include many centres with relatively few patients in each.

Alternative hierarchical formulations of the proportional hazards regression model can be used for meta-analysis and for detecting and exploring possible sources of heterogeneity. Assuming a fixed treatment effect across trials and fixed trial effect, two formulations of the Cox model are possible for undertaking meta-analysis as described by Williamson [90] and Whitehead [88]. These fixed effect models for the meta-analysis of trials comparing two treatments are reviewed and described in the following sections. Further details regarding models assuming random effects are deferred to Chapter 5.

3.3.1. Cox regression model with trial indicator variables and fixed treatment effect

For the *ith* individual in the *jth* trial ($i=1...n_j$, $j=1...J$), the hazard function at time t is written

$$\lambda_{ij}(t) = \lambda_0(t)\exp(\beta_{0j} + \beta_1 x_{1ij}) \qquad (3.7)$$

where the fixed parameter β_{0j} indicates trial membership (with β_{01} constrained to equal zero) for all individuals in the *jth* trial and x_{1ij} is a treatment indicator variable. In this model, the hazards within all trials are assumed to be proportional to the same common baseline hazard function $\lambda_0(t)$. The fixed parameter β_1 indicates the log hazard ratio of the event in experimental group relative to the control group, which is assumed to be identical across trials. Parameter estimates are obtained by maximising the partial log likelihood using the Newton-Raphson procedure as described earlier. However, since construction of the risk set involves ordered event times for all individuals from all trials, this model does not strictly compare patients from experimental and control groups within the same trial, an underlying desirable property for meta-analysis. The restrictive assumption that the hazards are proportional to a common baseline hazard

function makes model (3.7) unappealing for undertaking meta-analysis as different settings and patient populations are likely to give rise to different baseline hazard functions. A less restrictive assumption of proportional hazards within each trial, rather than overall, can be achieved using a stratified Cox regression model.

3.3.2. Cox regression model stratified by trial with fixed treatment effect

In this model, the hazard function for the *ith* individual in the *jth* trial is written as

$$\lambda_{ij}(t) = \lambda_{0j}(t)\exp(\beta_1 x_{1ij}) \qquad (3.8)$$

where λ_{0j} is the baseline hazard function in the *jth* trial. In model (3.8), the restriction of overall proportional hazards across trials is removed and hazards are only assumed to be proportional within each trial. The log likelihood for this model is equal to the summation of each trial log likelihood (equation (3.6)) constructed using only ordered event times for individuals from within each trial. The disadvantage of model (3.8) is that a direct estimate of the trial effect is not produced but this is not considered to be critical for meta-analysis as the effect of trial in isolation is rarely of interest. As with model (3.7), the fixed parameter β_1 of model (3.8) indicates the log hazard ratio of the event in experimental group relative to the control group, which is assumed to be identical in each trial.

As models (3.7) and (3.8) assume a common treatment effect across trials, no allowance is made for residual heterogeneity in these models. Alternative assumptions and corresponding models are described in Chapter 5.

3.4. Inverse variance weighted average

An alternative approach for meta-analysis is to estimate the log hazard ratio and its variance within each trial and compute the usual inverse variance (IV) weighted average of estimates across trials. Either the log-rank analysis or Cox regression model may be used to estimate the within-trial log hazard ratio and its variance at the initial step of this approach. As the log-rank analysis and Cox model can produce different estimates of

log hazard ratio, the IV weighted averages may also differ depending on choice of method adopted to estimate the within-trial log hazard ratios and their variance.

The IV weighted average estimate of log hazard ratio and its variance are given by

$$\hat{\beta}_{IV} = \frac{\sum_{j=1}^{J} \hat{\beta}_j / v_j}{\sum_{j=1}^{J} 1/v_j} \qquad (3.9)$$

$$\text{var}(\hat{\beta}_{IV}) = \frac{1}{\sum_{j=1}^{J} 1/v_j} \qquad (3.10)$$

where $\hat{\beta}_j$ and v_j are the log hazard ratio and its variance estimated from within the *jth* trial.

3.5. Comparison of methods

Since a number of methods are available for undertaking meta-analysis of time-to-event outcomes with IPD, an understanding of the behaviour of these methods is required to enable a choice to be made about which method is most appropriate. The following sections provide a summary of how these approaches compare to each other theoretically and in section 3.6 a small simulation study is undertaken to provide further insight. To the author's knowledge, an investigation and comparison of these methods for meta-analysis has not been undertaken previously. A further choice of whether to assume fixed or random effects is not considered until the fifth Chapter.

3.5.1. Log-rank analysis versus Cox regression model

Standard survival analysis text books (such as that by Collett [29]) note the close connection between the Cox regression model and log-rank analysis. Using previous notation introduced in section 3.2 for a single trial with two treatment groups, Collett [29] describes the close connection as follows.

If x_i is an indicator variable which is unity for individuals allocated treatment group A and zero for individuals allocated treatment group B, the Cox regression model for analysis of a single trial is given by expression (3.5). When there are no tied event times, the regression coefficient β is estimated by maximising the logarithm of the likelihood function in equation (3.6). For individuals in treatment group B, the variable x_i is given value zero and the log-likelihood function (3.6) can therefore be re-written as

$$\log L(\beta) = d_A \beta - \sum_{k=1}^{r} \log\left\{n_{A(k)} \exp(\beta) + n_{B(k)}\right\}$$

where d_A is the total number of events in treatment group A and the summation is taken over k=1,2,...,r event times.

One test of the null hypothesis that β=0 is the score test which is based on the test statistic

$$\frac{U^2(0)}{I(0)}$$

which follows a chi-squared distribution with one degree of freedom under the null hypothesis. In the above expression,

$$U(\beta) = \frac{\delta \log L(\beta)}{\delta \beta} = \sum_{k=1}^{r} \left(d_{A(k)} - \frac{n_{A(k)} \exp(\beta)}{n_{A(k)} \exp(\beta) + n_{B(k)}} \right)$$

and

$$I(\beta) = -\frac{\delta^2 \log L(\beta)}{\delta \beta^2} = \sum_{k=1}^{r} \frac{n_{A(k)} n_{B(k)} \exp(\beta)}{(n_{A(k)} \exp(\beta) + n_{B(k)})^2}$$

which are referred to as the efficient score and information function respectively. When β=0, these quantities are given by

$$U(0) = \sum_{k=1}^{r} \left(d_{A(k)} - \frac{n_{A(k)}}{n_{A(k)} + n_{B(k)}} \right)$$

and

$$I(0) = \sum_{k=1}^{r} \frac{n_{A(k)} n_{B(k)}}{(n_{A(k)} + n_{B(k)})^2}$$

which are equivalent to expressions for U_{LR} and V_{LR} of the log-rank test (equation (3.1) and (3.2) respectively) when there are no ties (i.e. $d_{(k)} = 1$).

As described in section 3.2, the one-step estimator of the *log(hazard ratio)* $\hat{\beta}_{LR}$ is given by U_{LR}/V_{LR} with variance equal to $1/V_{LR}$. From the above standard results, $U(0) = U_{LR}$ and $I(0) = V_{LR}$, and it follows that the estimate of β from the first iterative step of the Newton Raphson procedure for maximising the Cox partial log likelihood function is equivalent to the log-rank one-step estimator as described in the literature ([9] and [51]). Following from these results, it would seem that if the true value of β is some distance from zero, the maximum likelihood estimator ($\hat{\beta}_{ML}$) for the Cox regression model and the log-rank one-step estimator ($\hat{\beta}_{LR}$) will differ. This can be seen by assuming convergence is reached for the maximum likelihood estimator at step (h=H+1) of the iterative procedure such that

$$\hat{\beta}_{ML} = \hat{\beta}_h + I^{-1}(\hat{\beta}_h) U(\hat{\beta}_h)$$

$$= I^{-1}(0) U(0) + \sum_{h=1}^{H} I^{-1}(\hat{\beta}_h) U(\hat{\beta}_h)$$

$$= \hat{\beta}_{LR} + \sum_{h=1}^{H} I^{-1}(\hat{\beta}_h) U(\hat{\beta}_h)$$

The estimates $\hat{\beta}_{ML}$ and $\hat{\beta}_{LR}$ will be most similar as $\sum_{h=1}^{H} I^{-1}(\hat{\beta}_h) U(\hat{\beta}_h)$ tends towards zero.

3.5.2. Stratified log-rank analysis versus an IV weighted average

The Early Breast Cancer Trialists' Collaborative Group [51] note that summing the logrank statistic and variance across trials (leading to a stratified Log-rank analysis)

effectively leads to the results of each trial being given a 'weight' in the overall assessment that depends appropriately on the amount of statistical information provided by it. This can be shown explicitly as follows. If within-trial log-rank analyses are used for estimating β_j and v_j then for the *jth* trial

$$\beta_j = \frac{U_{LRj}}{V_{LRj}} \text{ and } v_j = \frac{1}{V_{LRj}}$$

which can be inserted into equation (3.9) and (3.10) for the inverse variance weighted average which can be written as

$$\hat{\beta}_{IV} = \hat{\beta}_{IV(LR)} = \frac{\sum_{j=1}^{J}\left(\frac{U_{LRj}}{V_{LRj}}\right)\Big/\left(\frac{1}{V_{LRj}}\right)}{\sum_{j=1}^{J} V_{LRj}} = \frac{\sum_{j=1}^{J} U_{LRj}}{\sum_{j=1}^{J} V_{LRj}} \qquad \textbf{(3.11)}$$

$$\text{var}(\hat{\beta}_{IV(LR)}) = \frac{1}{\sum_{j=1}^{J} 1/v_j} = \frac{1}{\sum_{j=1}^{J} V_{LRj}} \qquad \textbf{(3.12)}$$

where U_{LRj} and V_{LRj} are the log-rank statistic and variance within the *jth* trial. Equation (3.11) and (3.12) are the expressions for the one-step estimator of log hazard ratio $\hat{\beta}_{LR}$ and its variance ((3.3) and (3.4) respectively) from the stratified log-rank analysis and these two methods for meta-analysis are shown to be equivalent. If individual log-rank analyses are not used for every trial to estimate the log hazard ratio β_j and its variance v_j (equation (3.9)), the methods may not be equivalent.

3.5.3. Stratified Cox regression model versus an IV weighted average

A connection between the stratified Cox regression model and IV weighted average estimates of log hazard ratio and its variance is now derived. If J separate Cox regression models each with a single treatment indicator variable are used to estimate the within-trial log hazard ratio β_j and its variance v_j, the IV weighted average across trials

(denoted $\hat{\beta}_{IV(ML)}$) based on these estimates may not necessarily be similar to the corresponding maximum likelihood $\hat{\beta}_{ML}$ estimate from a stratified Cox model (one Cox regression model including all data stratified by trial). This can be shown as follows.

For the *jth* study, the estimate of β at step h+1 of the Newton Raphson iterative procedure is given by $\hat{\beta}_{h+1(j)} = \hat{\beta}_{h(j)} + I_j^{-1}(\hat{\beta}_{h(j)})U_j(\hat{\beta}_{h(j)})$ with variance $v_{h+1(j)} = I_j^{-1}(\hat{\beta}_{h+1(j)})$. At the first iterative step,

$$\hat{\beta}_{1(j)} = 0 + I_j^{-1}(0)U_j(0)$$

and the IV weighted average estimate across all trials would be estimated by

$$\hat{\beta}_{IV} = \hat{\beta}_{IV(ML)_1} = \frac{\sum_{j=1}^{J} \hat{\beta}_{1(j)} / v_{1(j)}}{\sum_{j=1}^{J} 1/v_{1(j)}} = \frac{\sum_{j=1}^{J} \hat{\beta}_{1(j)}.I_j(\hat{\beta}_{1(j)})}{\sum_{j=1}^{J} I_j(\hat{\beta}_{1(j)})} = \frac{\sum_{j=1}^{J} I_j^{-1}(0)U_j(0).I_j(\hat{\beta}_{1(j)})}{\sum_{j=1}^{J} I_j(\hat{\beta}_{1(j)})} \qquad \textbf{(3.13)}$$

with variance

$$\text{var}(\hat{\beta}_{IV(ML)_1}) = \frac{1}{\sum_{j=1}^{J} I_j(\hat{\beta}_{1(j)})} \qquad \textbf{(3.14)}$$

Under the assumption that each trial estimates a moderate treatment effect, $\hat{\beta}_{1(j)} \approx 0$ for all j and $I_j(\hat{\beta}_{1(j)}) \approx I_j(0)$. It follows that

$$\hat{\beta}_{IV(ML)_1} \approx \frac{\sum_{j=1}^{J} I_j^{-1}(0)U_j(0).I_j(0)}{\sum_{j=1}^{J} I_j(0)} = \frac{\sum_{j=1}^{J} U_j(0)}{\sum_{j=1}^{J} I_j(0)} \qquad \textbf{(3.15)}$$

$$\text{var}(\hat{\beta}_{IV(ML)_1}) \approx \frac{1}{\sum_{j=1}^{J} I_j(0)} \qquad (3.16)$$

Now consider the stratified Cox model. The log likelihood for this model is equal to the sum of log likelihood terms from each individual trial such that $\log L(\beta) = \sum_{j=1}^{J} \log L_j(\beta)$. It follows that

$$U(\beta) = \sum_{j=1}^{J} U_j(\beta) \text{ and } I(\beta) = \sum_{j=1}^{J} I_j(\beta)$$

At the first iterative step of the Newton Raphson procedure under the stratified Cox model,

$$\hat{\beta}_1 = 0 + I^{-1}(0)U(0) = \frac{\sum_{j=1}^{J} U_j(0)}{\sum_{j=1}^{J} I_j(0)} \qquad (3.17)$$

$$v_1 = I^{-1}(\hat{\beta}_1) \approx I^{-1}(0) = \frac{1}{\sum_{j=1}^{J} I_j(0)} \qquad (3.18)$$

which are equal to equations (3.15) and (3.16). Therefore, under the assumption that each individual trial estimate of log hazard ratio is close to zero, the IV weighted average of trial estimates from the first iterative step ((3.15) and (3.16)) will be approximately equal to corresponding stratified Cox model estimates ((3.17) and (3.18)).

At the second iteration, $\hat{\beta}_{2(J)} = \hat{\beta}_{1(J)} + I_j^{-1}(\hat{\beta}_{1(J)})U_j(\hat{\beta}_{1(J)})$ and $v_{2(J)} = I_j^{-1}(\hat{\beta}_{2(J)})$. The IV weighted average estimate across all trials is given by

$$\hat{\beta}_{IV(ML)_2} = \frac{\sum_{j=1}^{J} \hat{\beta}_{2(J)} / v_{2(J)}}{\sum_{j=1}^{J} 1/v_{2(J)}} = \frac{\sum_{j=1}^{J} \hat{\beta}_{2(J)} . I_j(\hat{\beta}_{2(J)})}{\sum_{j=1}^{J} I_j(\hat{\beta}_{2(J)})}$$

with variance

$$\text{var}(\hat{\beta}_{IV(ML)_2}) = \frac{1}{\sum_{j=1}^{J} I_j(\hat{\beta}_{2(J)})}$$

Replacing $\hat{\beta}_{2(J)}$ by $\hat{\beta}_{2(J)} = I_j^{-1}(0)U_j(0) + I_j^{-1}(\hat{\beta}_{1(J)})U_j(\hat{\beta}_{1(J)})$

$$\hat{\beta}_{IV(ML)_2} = \frac{\sum_{j=1}^{J}\left[I_j^{-1}(0)U_j(0)I_j(\hat{\beta}_{2(J)}) + I_j^{-1}(\hat{\beta}_{1(J)})U_j(\hat{\beta}_{1(J)})I_j(\hat{\beta}_{2(J)})\right]}{\sum_{j=1}^{J} I_j(\hat{\beta}_{2(J)})}$$

If convergence is achieved at the second iterative step then $I_j(\hat{\beta}_{2(J)}) \approx I_j(\hat{\beta}_{1(J)})$ and

$$\hat{\beta}_{IV(ML)_2} \approx \frac{\sum_{j=1}^{J}\left[I_j^{-1}(0)U_j(0)I_j(\hat{\beta}_{1(J)}) + I_j^{-1}(\hat{\beta}_{1(J)})U_j(\hat{\beta}_{1(J)})I_j(\hat{\beta}_{1(J)})\right]}{\sum_{j=1}^{J} I_j(\hat{\beta}_{1(J)})}$$

which simplifies to

$$\hat{\beta}_{IV(ML)_2} \approx \frac{\sum_{j=1}^{J}\left[I_j^{-1}(0)U_j(0)I_j(\hat{\beta}_{1(J)})\right]}{\sum_{j=1}^{J} I_j(\hat{\beta}_{1(J)})} + \frac{\sum_{j=1}^{J}\left[U_j(\hat{\beta}_{1(J)})\right]}{\sum_{j=1}^{J} I_j(\hat{\beta}_{1(J)})} \qquad (3.19)$$

with variance

$$\text{var}(\hat{\beta}_{IV(ML)_2}) \approx \frac{1}{\sum_{j=1}^{J} I_j(\hat{\beta}_{1(J)})} \qquad (3.20)$$

At the second iterative step of the Newton-Raphson procedure under the stratified Cox model,

$$\hat{\beta}_2 = \hat{\beta}_1 + I^{-1}(\hat{\beta}_1)U(\hat{\beta}_1)$$

which from (3.17) may be written as

$$\hat{\beta}_2 = \frac{\sum_{j=1}^{J} U_j(0)}{\sum_{j=1}^{J} I_j(0)} + I^{-1}(\hat{\beta}_1)U(\hat{\beta}_1) \qquad (3.21)$$

with variance

$$v_2 \approx \frac{1}{\sum_{j=1}^{J} I_j(\hat{\beta}_1)} \qquad (3.22)$$

The IV weighted average log hazard ratio (3.19) and its variance (3.20) are approximately equal to corresponding estimates from the stratified Cox model (equation (3.21) and (3.22) respectively) under the assumption that individual trial estimates are similar to each other and close to zero (i.e. assumption of homogeneity across trials and small treatment effect). If such assumptions are reasonable, $I_j(\beta_{1(j)}) \approx I_j(0)$ and

$$\hat{\beta}_{IV(ML)_2} \approx \frac{\sum_{j=1}^{J} U_j(0)}{\sum_{j=1}^{J} I_j(0)} + \frac{\sum_{j=1}^{J} U_j(\hat{\beta}_{1(j)})}{\sum_{j=1}^{J} I_j(\hat{\beta}_{1(j)})} \approx 2\frac{\sum_{j=1}^{J} U_j(0)}{\sum_{j=1}^{J} I_j(0)}$$

$$\text{var}(\hat{\beta}_{IV(ML)_2}) \approx \frac{1}{\sum_{j=1}^{J} I_j(0)}$$

which are approximately equal to equations (3.21) and (3.22) under the assumption that $\hat{\beta}_1 \approx 0$. A similar argument would apply if the Newton-Raphson procedure required further iterations to reach convergence. However, further iterations might indicate that the estimate of log hazard ratio is further from zero.

Deviations from the assumptions outlined above would result in different estimates for the IV weighted average log hazard ratio and variance compared with the stratified Cox model.

3.6. Simulation study

As indicated above by these theoretical results, a stratified Cox model, stratified log-rank analysis and IV weighted average may yield different estimates of pooled log hazard ratio and its variance under certain conditions. There is a clear need to investigate and compare the results and this is done with a small exploratory simulation study. As the stratified log-rank analysis estimates are equivalent to the IV weighted average of individual within-trial log-rank estimates (section 3.5.2), only the stratified log-rank analysis will be examined in the simulation study.

Although several factors could impact on parameter estimation for different methods of meta-analysis, this initial simulation study considers only the impact of underlying treatment effect, denoted by β_1, and underlying between trial variability in treatment effect, denoted by τ^2. In Chapter 5, a simulation study is undertaken to compare the behaviour of both fixed and random effect Cox regression models for meta-analysis and for exploring heterogeneity. As will be described in Chapter 5, the random effect Cox models require a large amount of computing time for parameter estimation and this restriction largely influenced the choice of simulation parameters to investigate in that particular simulation study. For consistency, and to allow comparisons to be made between results of the current simulation study with those in Chapter 5, a deliberate decision was made to ensure that the same simulation parameters and random number generator seeds were used in the current simulation study.

To allow different levels of between trial variation, the data are simulated under a model with random trial and random treatment effects that will be described in Chapter 5 (model 5.5 RE/RE). The model for the *ith* individual ($i=1,\ldots,n_j$) in the *jth* trial ($j=1,\ldots,J$), presented here to facilitate understanding of the simulation, is given by

$$\lambda_{ij}(t) = \lambda_0(t)\exp(b_{0j} + \beta_{1j}x_{1ij})$$

$$\beta_{1j} = \beta_1 + b_{1j}$$
$$b_{0j} \sim N(0, \sigma^2)$$
$$b_{1j} \sim N(0, \tau^2)$$
$$\text{cov}(b_{0j}, b_{1j}) = 0$$

where x_{1ij} represents treatment group membership coded as ± ½, β_1 is the average log hazard ratio for a population of possible treatment effects, and b_{1j} is the deviation of the log hazard ratio in the *jth* trial from this population average, random quantities that are assumed to follow a Normal distribution with mean zero and variance τ^2 which is a measure of the between trial variability in treatment effect i.e. a measure of the degree of statistical heterogeneity. The random quantities b_{0j} represent the deviation in the *jth* trial of the risk from the average of the two treatments and these random effects are assumed to follow a Normal distribution with mean zero and variance σ^2 representing the variation in baseline risk across trials.

Within each trial, the random quantities b_{0j} and b_{1j} are each generated from a Normal $(0, \sigma^2)$ and Normal $(0, \tau^2)$ distribution respectively. For each trial, the control group and experimental group log hazard rates are calculated as

$$\log(\lambda_{Cj}) = \log\lambda_0 + b_{0j} - 1/2(\beta_1 + b_{1j})$$
$$\log(\lambda_{Ej}) = \log\lambda_0 + b_{0j} + 1/2(\beta_1 + b_{1j})$$

where $\lambda_0 = 1$ in these investigations and β_1 is a fixed simulation parameter representing the underlying average log hazard ratio.

In order to reflect varying magnitudes of underlying average treatment effect, and varying degrees of statistical heterogeneity, values of 0, 0.1, 0.5 and 0.9, 1.5 and 1.9 (note that 1.5 and 1.9 are not examined in the simulation study of Chapter 5) are chosen for the parameter β_1 and 0, 0.1, 0.5, 0.9 for the parameter τ^2. Values of 0, 0.1, 0.5 for β_1 represent hazard ratios of 1, 1.1, 1.6, chosen to reflect a range of plausible values that may commonly be encountered in meta-analysis whilst values of 0.9, 1.5 and 1.9 represent hazard ratios of 2.5, 4.5, and 6.7 and were chosen to explore patterns for

examples with more extreme treatment effects. Values of 0, 0.1, 0.5 and 0.9 for the parameter τ^2 represent clinically plausible values of minimal, moderate and a more extreme degree of heterogeneity required to adequately explore the effect of increasing heterogeneity.

The hazard rates in each group are used to generate a potential exponential failure time for each individual in each trial. A potential censoring time for each individual is generated from a uniform distribution on [t2-t1, t2], where t2 indicates the time of analysis for a trial and [0,t1] denotes the accrual period for the trial with patient entry times assumed to be independent uniform random variables on this interval. For simplicity, t2=2 and t1=1 are explored throughout but future investigations could examine alternative values in order to control and examine the impact of amount of censoring. Finally, a censoring indicator and corresponding 'time to event' are obtained for each individual where,

(i) time to event = potential failure time,
censoring indicator =1 if potential failure time ≤ potential censoring time

(ii) time to event = potential censoring time,
censoring indicator =0 if potential failure time > potential censoring time.

In each meta-analysis situation, the between trial variability in baseline risk is assumed to be zero (i.e. $\sigma^2=0$) in these initial investigations which is actually equivalent to simulating data under model (5.3) of Chapter 5 which includes a fixed trial effect and random treatment effect. Furthermore, if the between trial variability is assumed to be zero (i.e. $\tau^2=0$), the model is equivalent to model (3.7) described earlier which includes fixed trial and treatment effects. The above simulation framework (under model 5.5) was chosen, rather than assuming a fixed trial effect, as it allows different values of σ^2 to be easily explored in the future if required.

Data are generated for 40 individuals in each of two treatment groups in each of 5 trials. Although a meta-analysis of 5 trials may not be representative of all IPD meta-analyses, especially those undertaken to compare therapies for treating cancer, this number does accurately reflect the number of trials included in IPD meta-analyses of drug trials in epilepsy as described in the next chapter. In fact, a recent systematic review (Mark

Simmonds, personal communication) revealed that IPD based meta-analyses typically include less than 10 trials. However, larger within-trial sample sizes should ideally be examined but due to the computing time required to fit multiple Cox regression models including random effects (described in Chapter 5), larger sample sizes are not explored. As an example of the restrictions imposed, the model with random trial and random treatment effects (RE/RE) using data from the CBZ/VPS meta-analysis in chapter 5 took 42 hours, 6 minutes to fit using a Pentium II processor 400Mhz, 128 MB system RAM. For the same reason, the number of repetitions of each meta-analysis is constrained to 100 throughout. A simulation study based on a small number of repetitions does require careful interpretation as the simulated data may not be representative of results from a larger (e.g. based on 1000 repetitions) simulation study. In particular, with only 100 repetitions the random error is much greater and the results and conclusions should therefore only be considered as exploratory.

For each set of simulation parameters, the mean of the estimated log hazard ratios was calculated along with the coverage over all 100 simulations. Coverage is defined as the percentage of 95% confidence intervals that contain the true underlying value of log hazard ratio.

Simulation Results

With no heterogeneity between trials ($\tau^2=0$), the stratified Cox model, stratified log-rank analysis and IV weighted average approach, estimate the true log hazard ratio accurately and with approximately 95% coverage when the true treatment effect is less than or equal to 0.9 (Table 3.2, Figure 3.1 and Figure 3.5). For larger treatment effects the stratified log-rank analysis tends to overestimate the log hazard ratio with coverage of less than 95% whereas the stratified Cox model and IV weighted average estimate the log hazard ratio accurately and with coverage of approximately 95%.

A similar pattern for treatment effect is seen when the underlying heterogeneity parameter τ^2 is 0.1 (Table 3.2, Figure 3.2) with slightly less agreement between the stratified Cox model and the IV weighted average at higher values of log hazard ratio, although the discrepancy is minimal. On the other hand, coverage values are much worse for all three approaches (Figure 3.6) compared to corresponding values when

τ^2=0 reflecting the need for an approach that correctly takes into account the additional level of variability. The low coverage values for all values of treatment effect indicate that when there is some heterogeneity that is not appropriately accounted for in the analysis, the 95% confidence interval for the log hazard ratio will contain the true value less than 80% of the time with even less coverage for the stratified log-rank analysis.

As the degree of heterogeneity increases further (τ^2=0.5), the log hazard ratio is underestimated on average for all three methods (Table 3.2, Figure 3.3). When the true log hazard ratio is less than or equal to 0.5, similar estimates are obtained from the three methods with slightly less bias for the stratified Cox model. For larger values of treatment effect the least biased estimate is from a stratified log-rank analysis and the IV weighted average the most biased. Coverage is very poor (<60%) for all models and all values of log hazard ratio (Figure 3.7).

As the degree of heterogeneity increases further (τ^2=0.9) a more striking but similar pattern of bias and underestimation of the log hazard ratio is seen (Table 3.2, Figure 3.4) compared to patterns when τ^2=0.5. However, there is a greater degree of bias with this larger value of τ^2 and the bias increases more severely as the true value of log hazard ratio increases. Coverage is again extremely poor (<53%) for all models and all values of log hazard ratio (Figure 3.8).

Table 3.2. Mean log hazard ratio (standard deviation) and coverage for parameter combinations in 100 simulated meta-analyses of 5 trials.

Simulation parameters		Stratified log-rank		Stratified Cox		IV (Cox) pooled	
τ^2	β_1	Mean(sd) of $\hat{\beta}_1$	% Coverage	Mean(sd) of $\hat{\beta}_1$	% Coverage	Mean(sd) of $\hat{\beta}_1$	% Coverage
0	0	-0.0019 (0.1149)	95	-0.0012 (0.1149)	95	-0.0013 (0.1143)	95
	0.1	0.1007 (0.1154)	94	0.1015 (0.1155)	94	0.1008 (0.1150)	94
	0.5	0.5094 (0.1234)	94	0.5053 (0.1206)	95	0.5023 (0.1202)	96
	0.9	0.9307 (0.1374)	92	0.9074 (0.1295)	95	0.9017 (0.1290)	94
	1.5	1.5855 (0.1500)	87	1.5137 (0.1421)	93	1.5027 (0.1400)	93
	1.9	2.0030 (0.1515)	85	1.9177 (0.1571)	94	1.9010 (0.1535)	93
0.1	0	-0.0182 (0.1896)	76	-0.0173 (0.1892)	76	-0.0176 (0.1865)	76
	0.1	0.0823 (0.1864)	77	0.0829 (0.1860)	77	0.0811 (0.1835)	78
	0.5	0.4912 (0.1874)	74	0.4873 (0.1828)	76	0.4798 (0.1810)	77
	0.9	0.9084 (0.2044)	74	0.8877 (0.1930)	79	0.8742 (0.1913)	79
	1.5	1.5482 (0.2162)	71	1.4867 (0.2048)	76	1.4628 (0.2028)	82
	1.9	1.9540 (0.2110)	75	1.8806 (0.2177)	81	1.8483 (0.2153)	80
0.5	0	-0.0302 (0.3311)	55	-0.0288 (0.3284)	55	-0.0287 (0.3130)	56
	0.1	0.0654 (0.3305)	56	0.0661 (0.3280)	56	0.0613 (0.3124)	56
	0.5	0.4544 (0.3344)	55	0.4512 (0.3292)	57	0.4265 (0.3134)	55
	0.9	0.8510 (0.3467)	51	0.8382 (0.3361)	54	0.7940 (0.3226)	54
	1.5	1.4444 (0.3541)	50	1.4077 (0.3429)	51	1.3353 (0.3338)	50
	1.9	1.8322 (0.3507)	54	1.7898 (0.3563)	53	1.6985 (0.3522)	53
0.9	0	-0.0372 (0.4138)	47	-0.0355 (0.4099)	47	-0.0364 (0.3795)	52
	0.1	0.0544 (0.4140)	49	0.0554 (0.4108)	49	0.0466 (0.3801)	53
	0.5	0.4317 (0.4186)	48	0.4296 (0.4150)	50	0.3902 (0.3839)	50
	0.9	0.8036 (0.4274)	45	0.7956 (0.4207)	46	0.7254 (0.3919)	46
	1.5	1.3671 (0.4379)	45	1.3458 (0.4329)	47	1.2295 (0.4096)	39
	1.9	1.7374 (0.4318)	46	1.7145 (0.4375)	50	1.5742 (0.4275)	46

Figure 3.1. Mean overall log hazard ratio estimated in 100 repetitions of 5 trial meta-analyses with underlying heterogeneity parameter $\tau^2 = 0$.

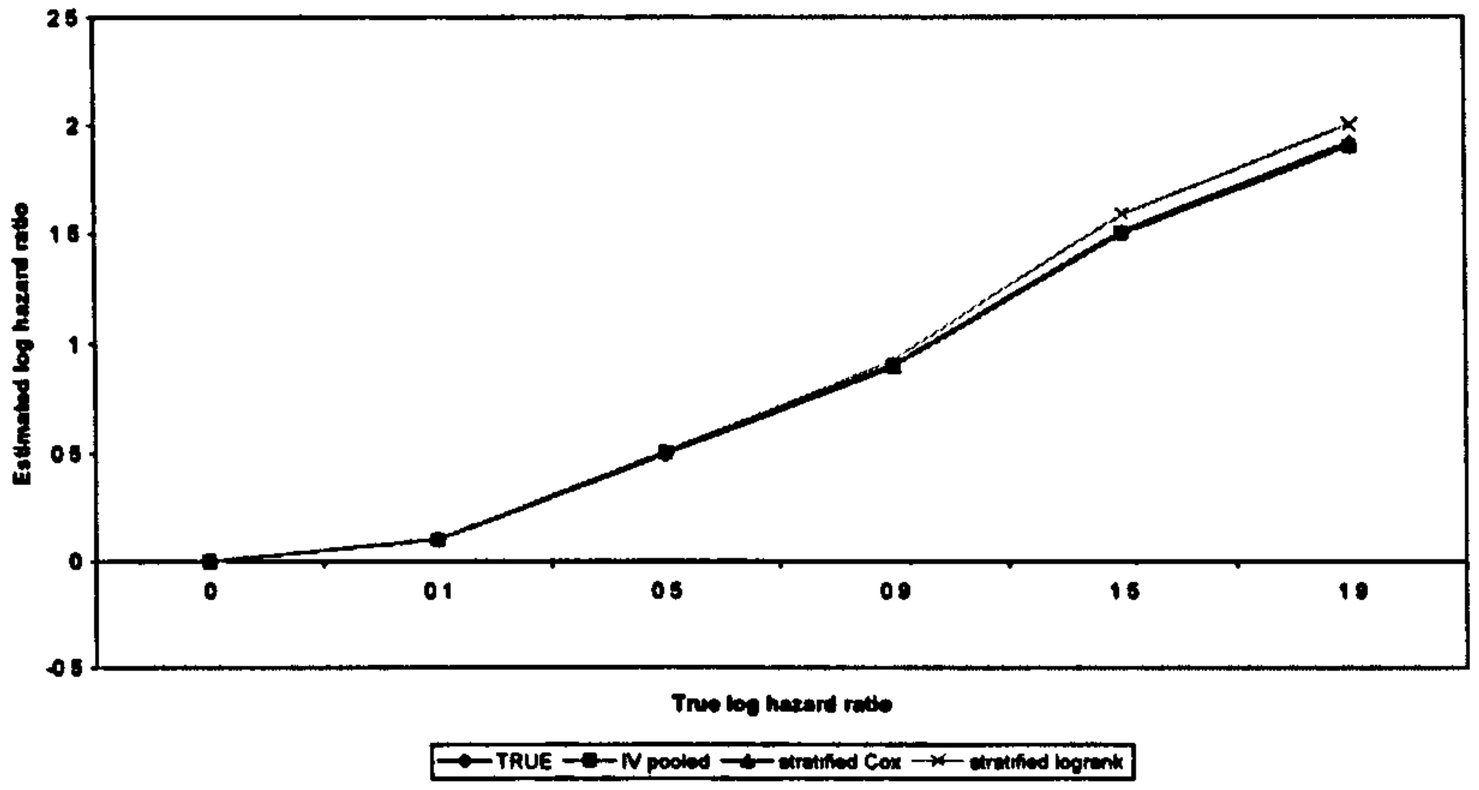


Figure 3.2. Mean overall log hazard ratio estimated in 100 repetitions of 5 trial meta-analyses with underlying heterogeneity parameter $\tau^2 = 0.1$.

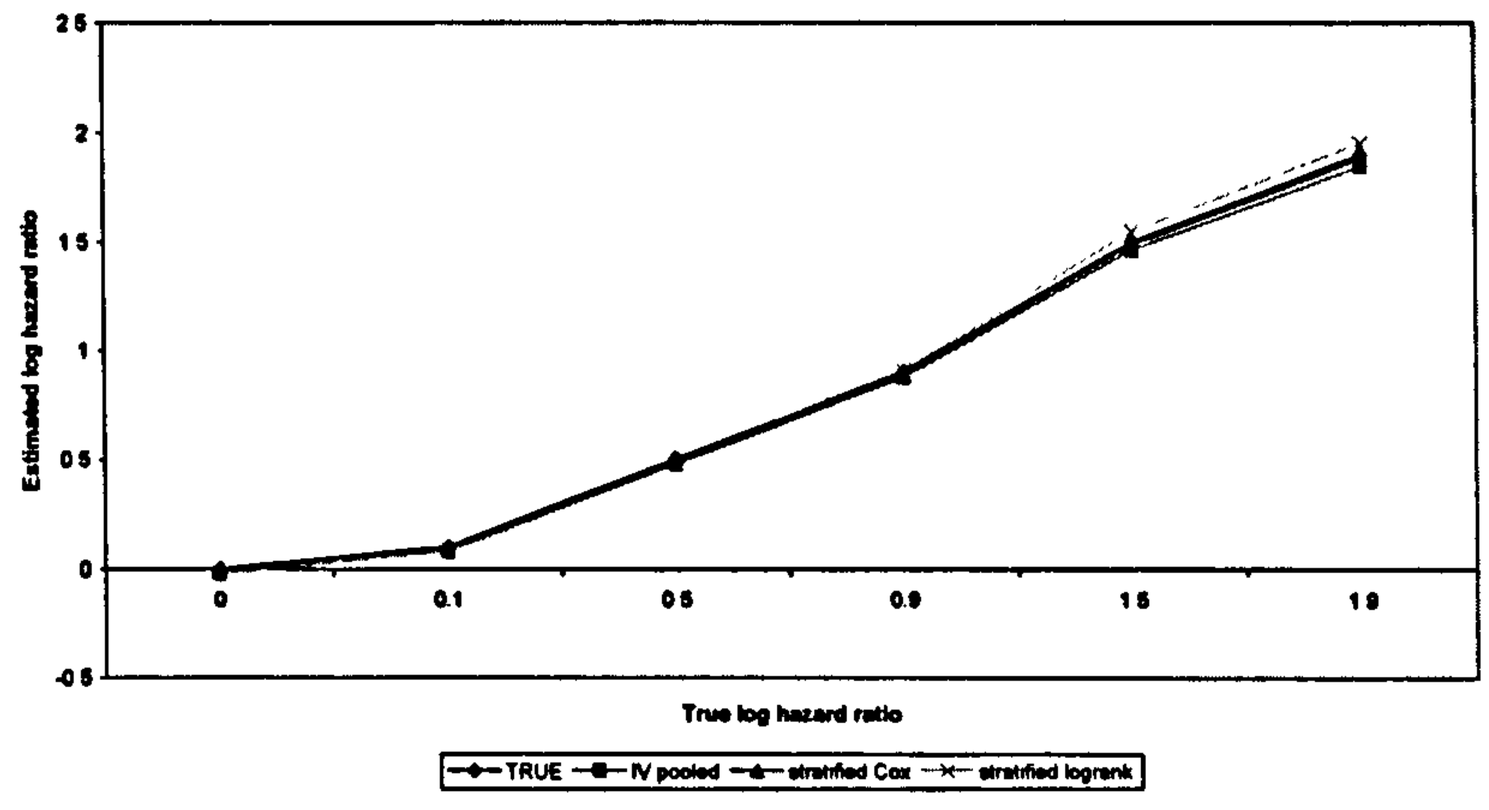


Figure 3.3. Mean overall log hazard ratio estimated in 100 repetitions of 5 trial meta-analyses with underlying heterogeneity parameter $\tau^2 = 0.5$.

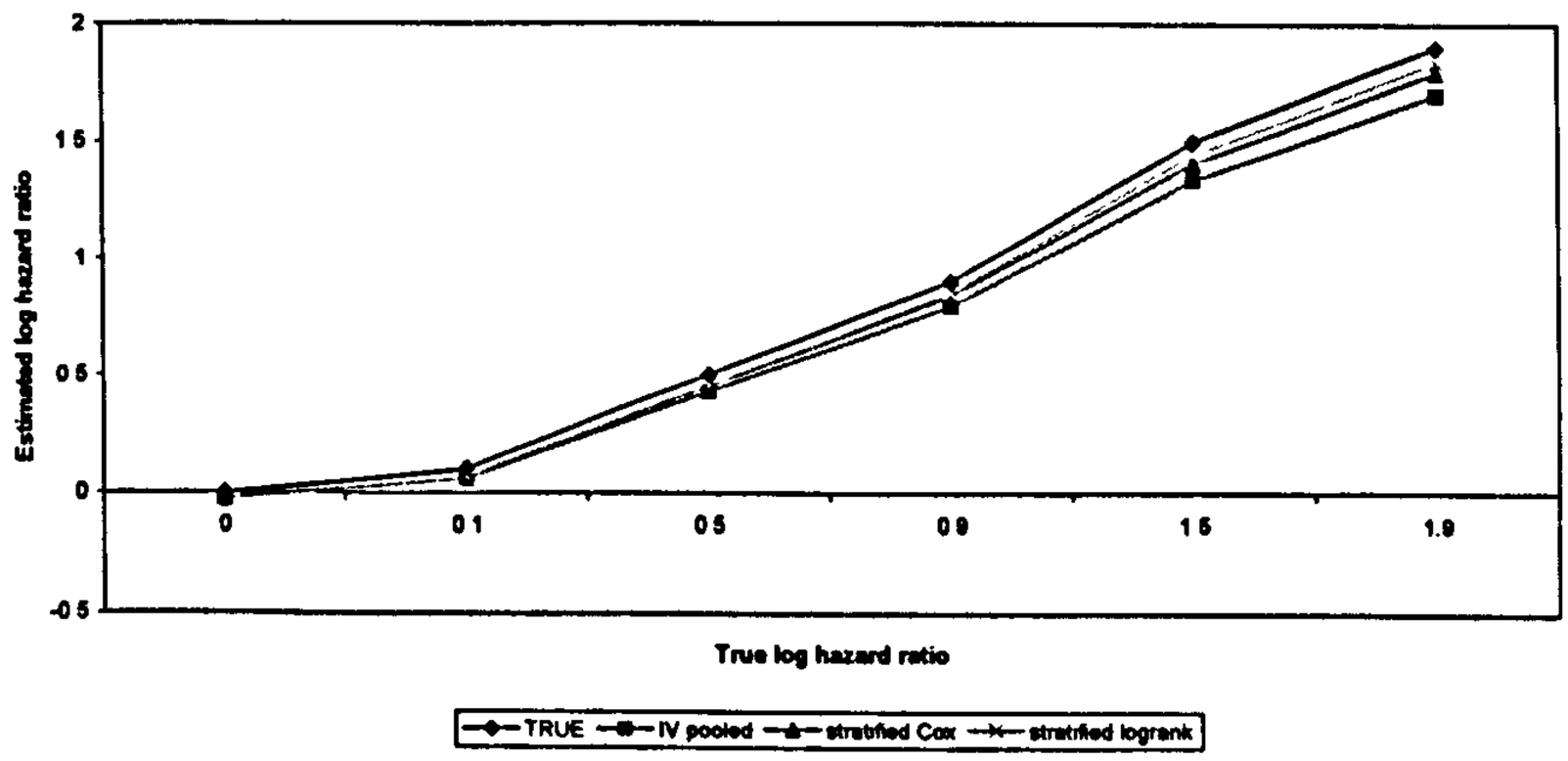


Figure 3.4. Mean overall log hazard ratio estimated in 100 repetitions of 5 trial meta-analyses with underlying heterogeneity parameter τ^2=0.9.

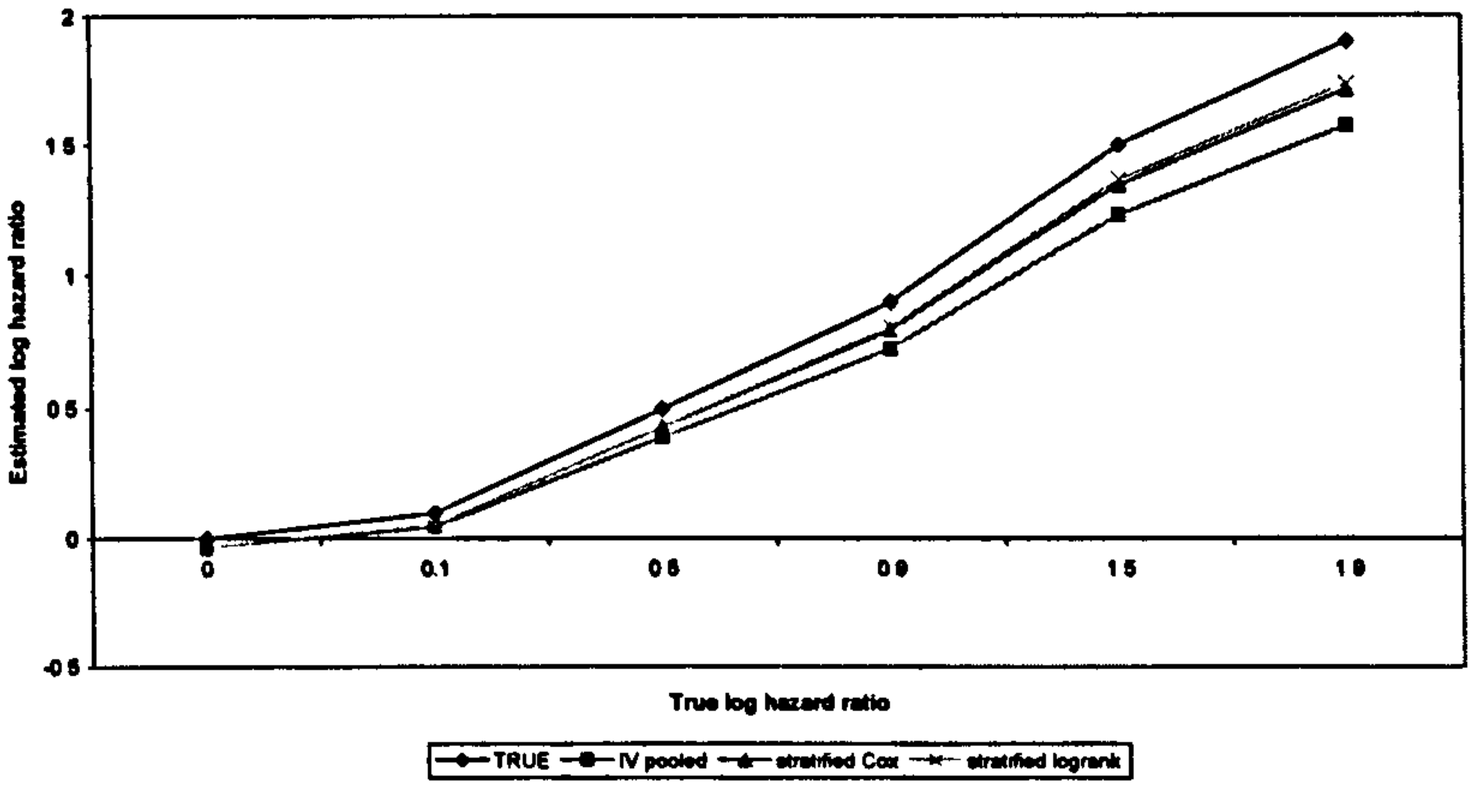


Figure 3.5. Coverage (percentage of 95% confidence intervals for the log hazard ratio that contain the true value) over 100 repetitions of 5 trial meta-analyses with underlying heterogeneity parameter τ^2=0.

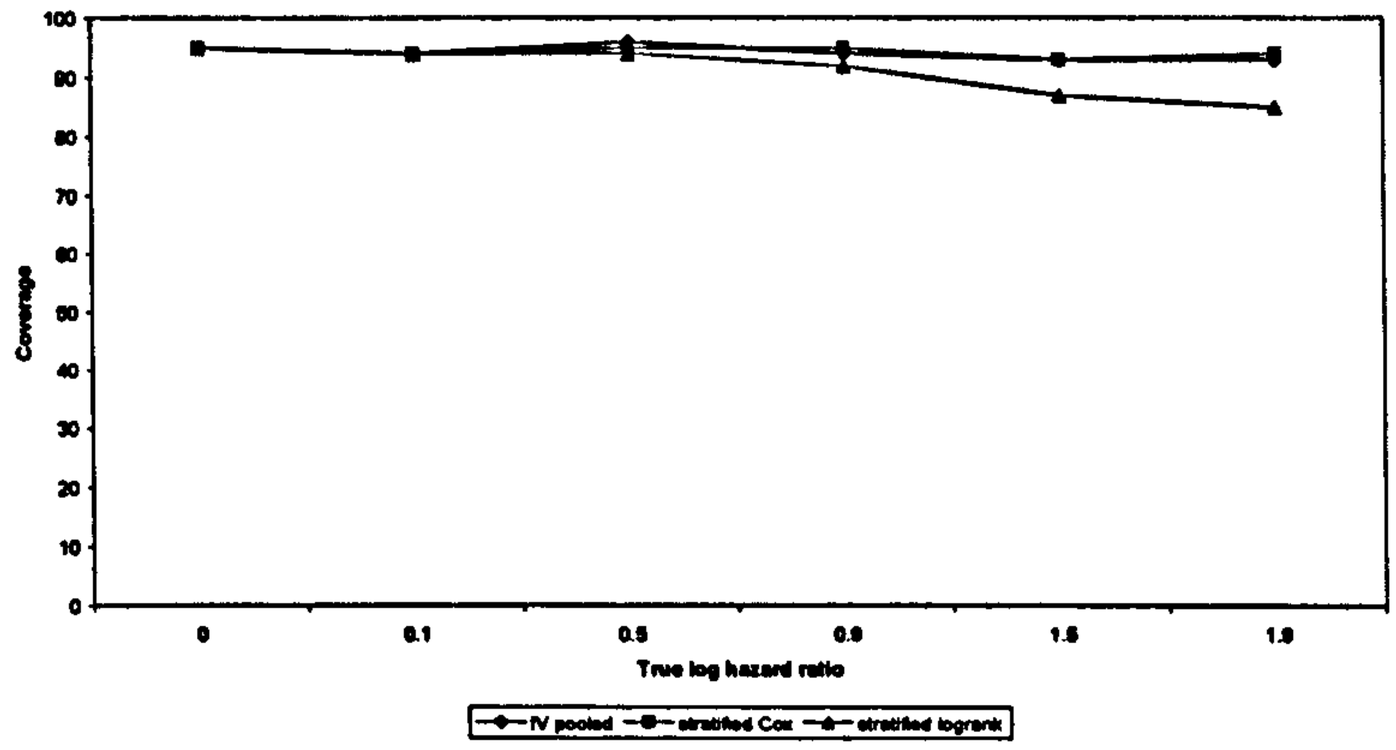


Figure 3.6. Coverage (percentage of 95% confidence intervals for the log hazard ratio that contain the true value) over 100 repetitions of 5 trial meta-analyses with underlying heterogeneity parameter τ^2=0.1.

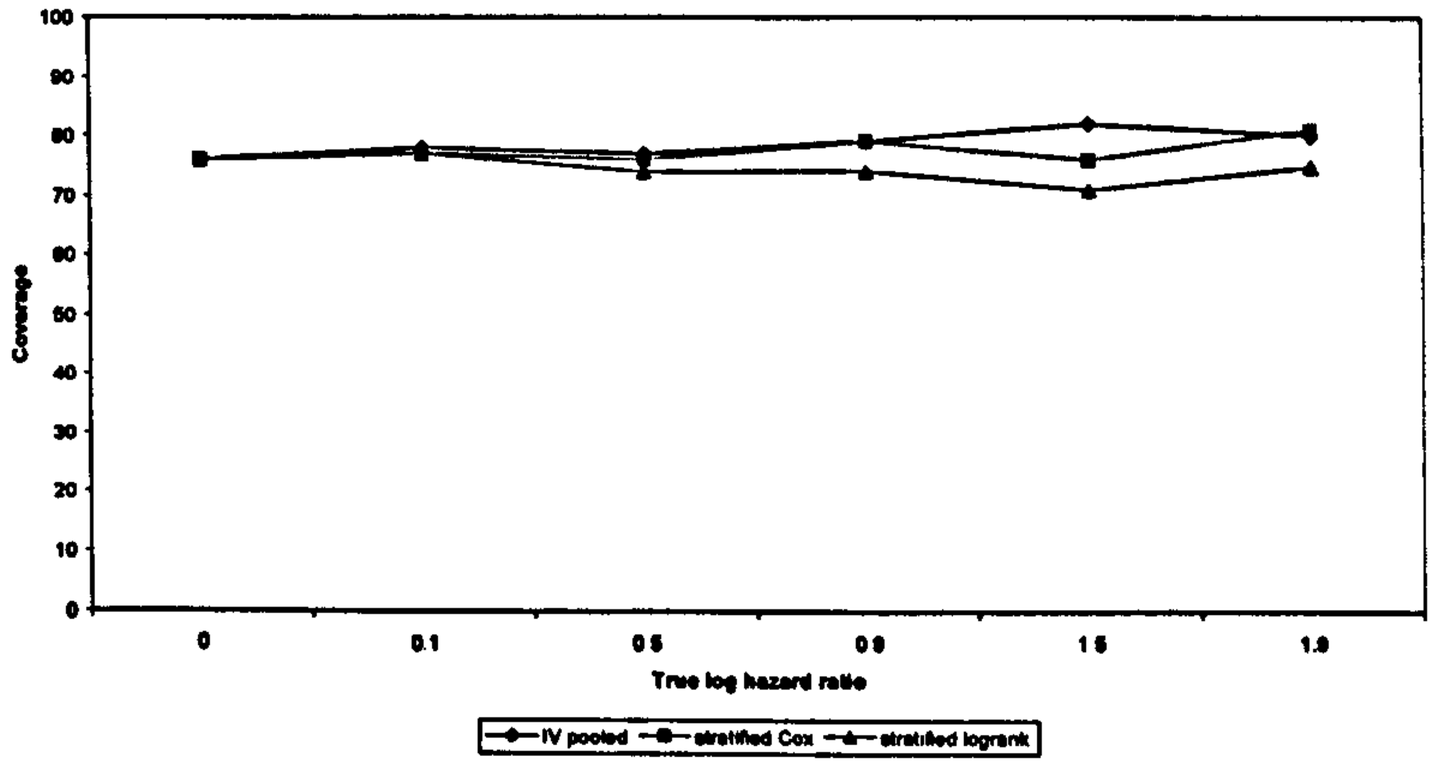


Figure 3.7. Coverage (percentage of 95% confidence intervals for the log hazard ratio that contain the true value) over 100 repetitions of 5 trial meta-analyses with underlying heterogeneity parameter τ^2=0.5.

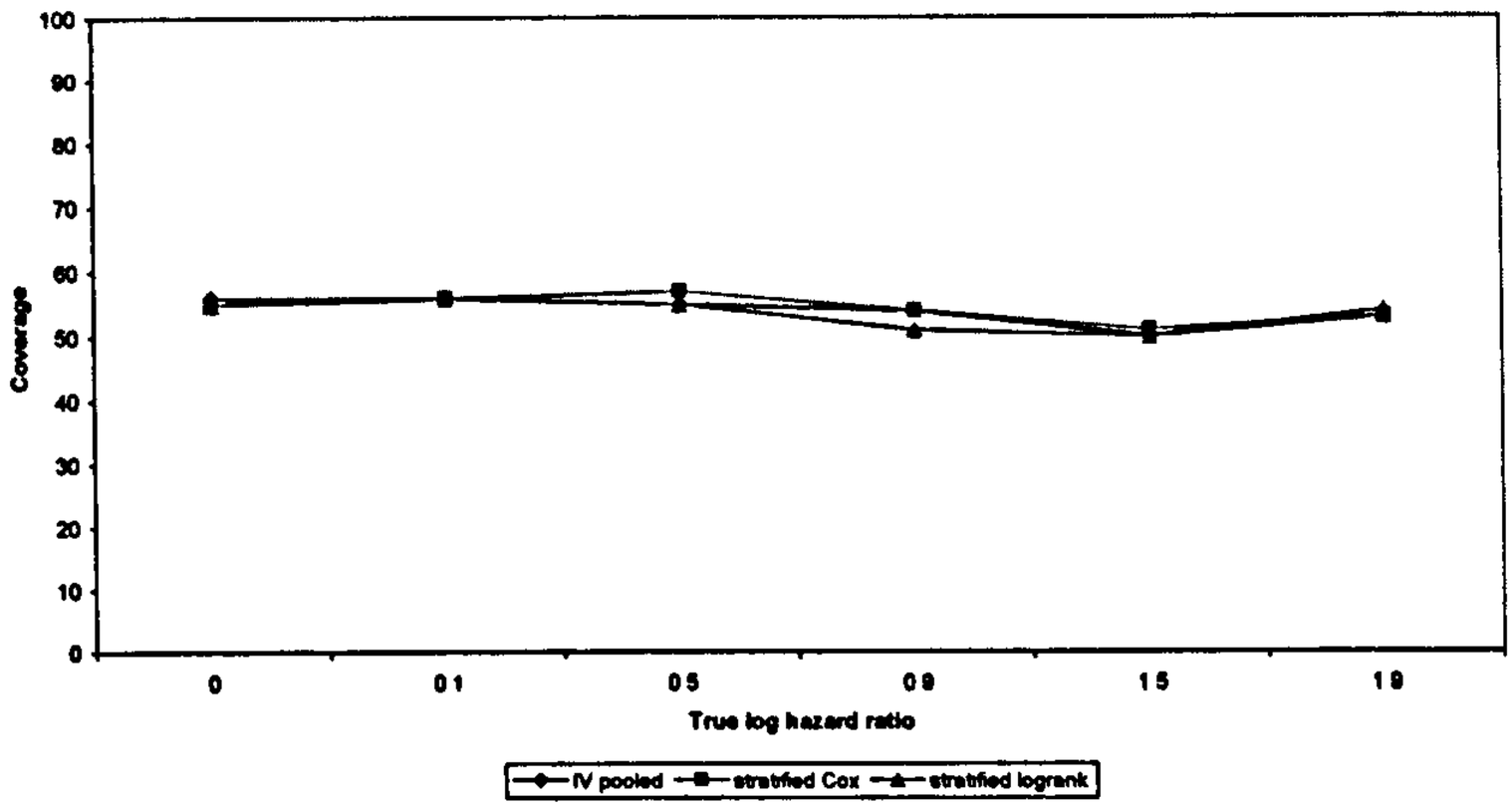
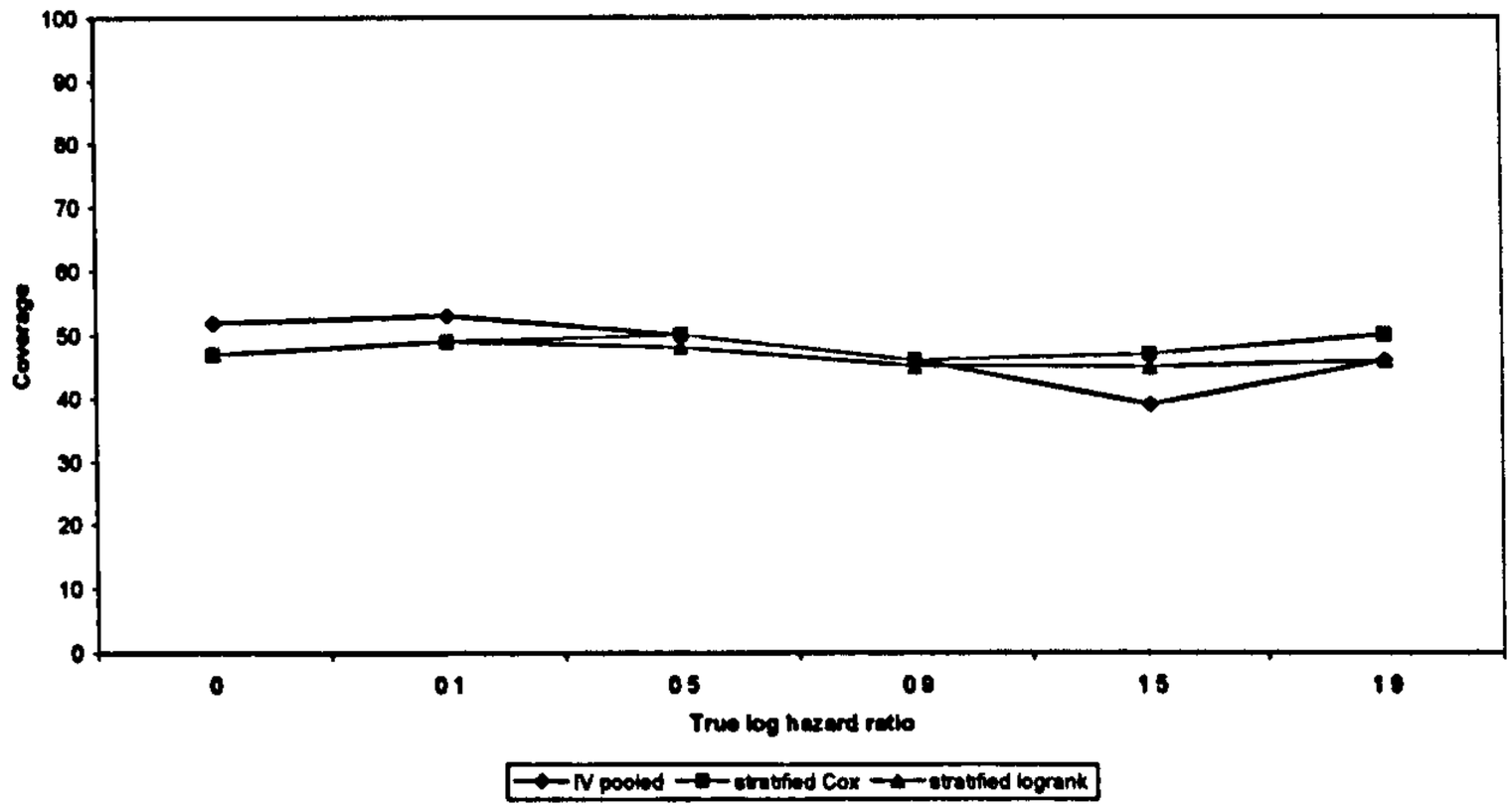


Figure 3.8. Coverage (percentage of 95% confidence intervals for the log hazard ratio that contain the true value) over 100 repetitions of 5 trial meta-analyses with underlying heterogeneity parameter τ^2=0.9.



In summary, when there is no heterogeneity, the stratified Cox model and IV weighted average will give similar estimates and are to be preferred to the stratified log-rank analysis when the underlying treatment effect is larger than 0.9. For values less than 0.9, all three methods perform well. Similar patterns in bias are seen as the degree of heterogeneity increases slightly but coverage values decrease quite dramatically. The bias increases and coverage decreases with further increases in the heterogeneity parameter particularly for larger treatment effects. The low coverage values highlight the need for models that appropriately account for the between trial variation. However, these

simulations are based on only 40 patients per treatment group in each trial. Further simulations with a larger number of patients and repetitions are required to establish more specific guidelines on choice of method.

3.7. Discussion

Individual patient data based meta-analyses with time-to-event outcomes are increasingly common but the literature regarding different methods for analysis and how they might compare is scarce. This Chapter has addressed a selection of approaches that assume a fixed treatment effect, namely the inverse variance weighted average of trial estimates, the stratified log-rank analysis and the stratified Cox regression model. Further discussion regarding random effects alternatives are provided in the fifth Chapter of this thesis.

The close connection between estimates from the log-rank test and Cox regression models have been noted elsewhere [29], [51]. Further theoretical results presented in this thesis indicate the equivalence of a stratified log-rank analysis and IV weighted average of log-rank estimates, and the connection between stratified Cox model estimates with those obtained from calculating a weighted average of within-trial Cox model estimates. These latter two Cox model based approaches are expected to provide similar estimates for modest treatment effects that are similar across trials.

The simulation study, although limited in terms of the number of trials, patients and repetitions examined, suggests that these theoretical results may apply in small studies. The results further suggest that when there is no heterogeneity, the stratified log-rank analysis may be best avoided for estimating larger treatment effects and all methods appear inadequate for increasing levels of heterogeneity. Future extensions to the current simulation study should also include presentation of I^2 statistics to give an indication of proportion of variation due to heterogeneity rather than chance.

A simulation study was undertaken by Greenland and Salvan [66] to examine the Peto one-step estimator of odds ratio in meta-analysis based on binary outcome data. They note that the one-step estimate will be asymptotically unbiased under the null hypothesis but will become increasingly biased as the trial level estimates of odds ratio get further from the value of 1. They further note that from numerical exploration the bias in the

pooled estimate of odds ratio can be positive or negative. The direction of bias for the stratified log-rank pooled estimate explored in the simulation study of this chapter also support their observations. Any bias in the one-step estimator is noted to be negligible in meta-analysis of randomised trials with small treatment effects and a reasonably large numbers of events [66].

All three methods of analysis are straightforward to implement using standard statistical software. The specialist software package, SCHARP, developed specifically for IPD meta-analysis by the MRC Clinical Trials Unit, Meta-analysis Group utilises the stratified log-rank approach described in detail by the Early Breast Cancer Trialists' Collaborative Group [51]. Two-stage methods such as an IV weighted average or stratified log-rank analysis, are the most commonly adopted approach to analysis (Mark Simmonds, personal communication). Possible reasons for the popularity of a stratified log-rank method might include availability of the SCHARP software package and the attractiveness of a non-parametric method. The stratified log-rank analysis is also noted to be of maximal statistical sensitivity for the detection of modest treatment effects [51]. A systematic review of IPD meta-analyses currently underway (Mark Simmonds, personal communication) could provide valuable information regarding how often the stratified log-rank analysis is used under these conditions.

The results presented in this Chapter may be useful to reviewers undertaking IPD meta-analysis as they highlight that different estimates and conclusions may be drawn from different approaches, a fact that should be considered at the protocol development stage for the review. However, further simulation and empirical work is required to make specific recommendations regarding choice of method. One further note to consider would be the availability of data. If IPD are not available for all included trials, there may be scope to extract AD for those trials without IPD. The analysis of all trials where a mixture of IPD and AD are available could not be undertaken using a fully stratified Cox regression model. On the other hand, if IPD are available for all included trials, the stratified Cox regression model may be advantageous since it extends easily to incorporate covariate data to enable exploring heterogeneity and factors for explaining heterogeneity in meta-analysis.

CHAPTER 4

Monotherapy drugs for Epilepsy: Meta-analyses based on individual patient data

In the current Chapter, the methodology available for undertaking meta-analysis of individual patient time-to-event data are illustrated with IPD from monotherapy drug trials in epilepsy. Some of the practicalities and arising issues are highlighted.

4.1. Introduction

Epilepsy is one of the commonest neurological conditions in the United Kingdom with an estimated 400,000 people in England and Wales affected by the condition [67]. Epilepsy is a condition in which people have seizures that occur when there is a disturbance in the normal electrical activity of the brain. Seizures are stereotyped attacks, the features of which depend upon which part of the brain is involved. They may involve abnormal sensations, movements, and are most commonly recognised when the individual has a convulsion. There are more than 30 types of seizures that fall into two general categories, partial and generalized. In partial seizures, at the start of the seizure, the abnormal electrical discharge is limited to one area of the brain, whereas in generalized seizures the whole of the cerebral cortex is involved from the outset.

Following a diagnosis of epilepsy, it is important to consider treatment with an anti-epileptic drug as they are associated with remission of seizures in 60-70% of those

treated. Anti-epileptic drugs (AEDs) are usually initially administered as single drug therapies, referred to as monotherapy. If seizures remain uncontrolled despite an appropriate maintenance dose, the patient may either withdraw from initial drug and receive a different drug taken as monotherapy, or stay on initial drug but receive additional anti-epileptic drug(s) in combination, referred to as either polytherapy, add-on therapy, adjunctive therapy or combination therapy. Anti-epileptic drugs can also cause side-effects that may be intolerable to the patient resulting in withdrawal from that particular drug. These patients may subsequently be tried on an alternative monotherapy treatment.

The National Institute for Clinical Excellence published the following guidelines [67] for the clinical management of epilepsy

- Adults with epilepsy should be treated with just one anti-epileptic drug where possible. If the first drug does not prevent seizures, another can be tried.
- Adjunctive or combination therapy should only be considered when attempts at monotherapy have not resulted in seizure freedom.
- A careful assessment of the risks and benefits of treatment with individual AEDs should be undertaken, particularly in relation to women of childbearing potential.
- A person who has a seizure for the first time should be seen by an epilepsy specialist as soon as possible, to find out exactly what type of epilepsy he or she has, so that the best treatment can be started.
- Treatment should be reviewed at regular intervals.

The epilepsy clinician is faced with a decision of prescribing one of several anti-epileptic drugs, some of which are thought to work better for one type of epilepsy than another, or for some people better than others [68]. For example, Valproate is the treatment of choice for generalized seizures and Carbamazepine the treatment of choice for partial seizures [69], [70] although there is little in the way of evidence from randomised controlled trials to support it [39], [70]. Up until 1973, four standard anti-epileptic drugs were thought to be effective in treating epilepsy (Phenobarbitone, Phenytoin, Carbamazepine, Valproate). Between 1989 and 2000, a series of 'new' anti-epileptic drugs have been identified and seven are currently licensed in the UK (Gabapentin,

Lamotrigine, Levetiracetam, Oxcarbazepine, Tigabine, Topiramate and Vigabatrin). Due to advances in drug regulation procedures, the new AEDs have been subjected to careful clinical evaluation prior to licensing, where they have been assessed in the first instance as add-on therapy within randomised placebo controlled trials [71], [68]. As drug regulation procedures were less scrupulous at the time of licensing the older 'standard' AEDs, randomised evidence from comparisons between standard AEDs and placebo controls is not available. Nevertheless, these 'standard' AEDs are accepted as clinically effective and remain a viable choice in practice.

4.2. Epilepsy data

The choice of appropriate anti-epileptic drug for a particular patient should be based on clinical judgement and evidence from randomised controlled trials. To summarise the current best available evidence from randomised controlled trials, the effects of four 'standard' (Carbamazepine (CBZ), Valproate (VPS), Phenytoin (PHT), Phenobarbitone (PHB)) and two 'new' anti-epileptic drugs (Oxcarbazepine (OXC), Lamotrigine (LTG)), when used as monotherapy in patients with partial seizures or generalized seizures, have so far been examined in eight separate Cochrane systematic reviews [39], [72], [73], [74], [75], [76], [77], [78].

For each review comparing two drugs, denoted AED 1 and AED 2 for convenience below, the following general review methods were adopted.

Objectives

To compare the efficacy and tolerability of AED 1 and AED 2 when used as monotherapy in patients with partial onset seizures or generalised onset tonic-clonic seizures.

Types of studies

1. Randomised controlled monotherapy studies comparing AED 1 and AED 2. Studies may be double, single or unblinded.
2. Studies using either quasi (e.g. by date of birth) or adequate methods of randomisation.

Types of participants

1. Children or adults with partial onset seizures (simple partial, complex partial, or secondarily generalising tonic-clonic seizures) or generalised onset tonic-clonic seizures.
2. A new diagnosis of epilepsy, or a relapse following anti-epileptic drug withdrawal, or who have failed on other therapies.

Types of interventions

AED 1 or AED 2 monotherapy.

Types of outcomes

(i) time to withdrawal of allocated treatment due to inadequate seizure control or intolerable adverse effects; participants achieve this outcome if allocated treatment is withdrawn for poor seizure control, adverse effects, non-compliance (assumed to reflect a patient's intolerance to drug or perceived ineffectiveness in terms of seizure control) or if additional add-on treatment is initiated (i.e. allocated treatment has failed). The outcome is censored if treatment was withdrawn because the individual achieved a period of remission or if the individual was still on allocated treatment at the end of follow-up. It is a combined outcome reflecting both efficacy and tolerability and is an outcome to which the individual makes a contribution. It is the primary outcome measure recommended by the Commission on Antiepileptic Drugs of the International League Against Epilepsy [79]

(ii) time to 12 month remission from seizures; individuals achieve this outcome if a continuous period of 12 months is experienced without any seizures and is a particularly important outcome for adults as the application of a UK driving license requires a seizure free period of at least 12 months

(iii) time to first seizure; individuals achieve this outcome as soon as the first seizure occurs after randomisation. A number of pharmaceutical industry based trials that examine this particular outcome ignore any seizures that occur within the first 6 weeks

following randomisation. This approach is adopted to allow drug dose to stabilise after the initial titration period. Since each systematic review and analyses presented in this thesis adopt an intention to treat approach, in order to evaluate the treatment policy, all seizure data collected from date of randomisation are used in calculations regardless of titration period.

Search strategy for identification of studies

1. MEDLINE (See search strategy for Epilepsy Group specialist register of RCTs).
2. The latest edition of the Cochrane Library (See search strategy for Epilepsy Group specialist register of RCTs).
3. Contacting manufacturers of AED 1.
4. Contacting manufacturers of AED 2.
5. Contacting original investigators of relevant trials found and experts in the clinical area

Individual patient data

Due to the lack of uniformity across trials in reporting these outcome measures, the desire to investigate time-to-event outcomes and examine treatment-covariate interactions, individual patient data (IPD) were requested from the authors of identified and eligible randomised trials.

Original trial authors were contacted asking whether they would collaborate with an IPD meta-analysis, and whether data from their trial could be made available. The response was favourable, and we proceeded to ask for the following participant data for each randomised patient within each trial: unique patient identifier, date of randomisation, drug allocated and dose, dates of follow-up, dates of dose changes, dates of all seizures (any type) post randomisation or seizure frequency data, date of treatment withdrawal, reason for treatment withdrawal, degree and methods of blinding, method of generation of randomisation list and method of concealment of randomisation. No independent prognostic factor studies in newly diagnosed patients had been undertaken. However, previous related studies [80], [81], [82], [83], [84], [85] in patients with epilepsy indicated that eight patient covariates may have a prognostic effect. These were, age,

sex, presence of neurological signs, seizure type at randomisation, number of seizures prior to randomisation, time from first ever seizure to randomisation, EEG results and CT/MRI results. These covariate data were requested for each randomised patient in each trial.

For each trial where IPD were supplied, a series of data validation assessments were performed and included (i) range and consistency checks with missing data, errors and inconsistencies identified and followed up with a nominated individual, (ii) trial details were cross checked against any published report of the trial and all possible results from the trial reports were reproduced using the provided IPD, (iii) review of the chronological randomisation sequence with missing allocation numbers followed up with the nominated individual, (iv) balance of prognostic factors were checked, taking account of factors stratified for in the randomisation procedure.

The outcome time to 12 month remission was calculated from the date of randomisation to the estimated date the individual had first been free of seizures for 12 months. The outcome time to first seizure was calculated from the date of randomisation to the date that a first seizure following randomisation was estimated to have occurred. If seizure data were missing for a particular visit, these outcomes were censored at the previous visit. These outcomes were also censored if the event of interest did not occur, an individual died, or follow-up ceased prior to the occurrence of the event of interest.

For a number of trials, seizure data were provided in terms of the number of seizures recorded between clinic visits. Linear interpolation was applied to approximate the dates on which seizures occurred. For example, if an individual recorded four seizures between two clinic visits that occurred on 01/03/90 and 01/05/90 (interval of 61 days), then date of first seizure would be approximated by dividing the 61 day period into five equal length segments and calculating the estimated date of first seizure as 13/03/90.

Analysis

The analysis was by intention to treat and included all randomised patients as far as possible, analysed in the treatment group to which they were allocated, irrespective of

which treatment they actually received. Log-rank analyses, stratified by trial were employed to obtain study-specific and overall estimates of hazard ratios (with 95% confidence intervals). Information provided by the stratified log-rank analyses were used to investigate the main effect of drug and to assess evidence for heterogeneity in drug effect between trials [51]. Clinical heterogeneity was assessed by reviewing the differences across trials in characteristics of randomised patients.

A summary of IPD available from all eligible RCTs and the comparisons examined in each review is given in Table 4.1 with a brief description of trial characteristics in Table 4.2. Summary data for the covariates of interest in each trial are provided in Table 4.3.

Table 4.1. Systematic reviews and availability of outcome data for monotherapy comparisons VPS: Valproate, CBZ: Carbamazepine, PHT: phenytoin, PHB: phenobarbitone, LTG: Lamotrigine, OXC: Oxcarbazepine.

Comparison examined	Trials included with IPD	Total number randomised	Percentage of IPD obtained from eligible trials[1]	Time to withdrawal	Time to 12 month remission	Time to first seizure
CBZ:VPS	De Silva 1996	103	95%	100	103	103
Marson *et al*	Heller 1995	122		118	122	122
2000 [39]	Mattson 1992	480		470	466	466
	Richens 1994	300		277	288	288
	Verity 1995	260		235	246	246
VPS:PHT	Craig 1994	166	60%	0	147	147
Tudur Smith *et al*	De Silva 1996	103		100	103	103
2001 [75]	Heller 1995	124		119	124	124
	Ramsay 1992	136		136	0	125
	Turnbull 1985	140		140	140	140
CBZ:PHB	De Silva 1996	64	59%	63	64	64
Tudur Smith *et al*	Heller 1995	119		115	119	119
2003 [72]	Mattson 1985	309		309	309	302
	Placencia 1993	192		189	192	192
CBZ:PHT	De Silva 1996	108	61%	106	108	108
Tudur Smith *et al*	Heller 1995	124		121	124	124
2002[73]	Mattson 1985	319		319	319	313

Comparison examined	Trials included with IPD	Total number randomised	Percentage of IPD obtained from eligible trials[1]	Time to withdrawal	Time to 12 month remission	Time to first seizure
PHB:PHT	De Silva 1996	64	70%	63	64	64
Taylor *et al*	Heller 1995	121		116	121	121
2001[74]	Mattson 1985	320		320	320	313
	Pal 1998	94		0	57	94
CBZ:LTG	Brodie 1995a	136	100%	136	0	123
Gamble *et al*	Brodie 1995b	124		124	0	119
2004 [78]	Reunanen 1996	352		0	0	349
	Brodie 1999	150		150	0	150
	Barrera 2001	622		622	0	0
PHT:OXC	Bill 1997	287	100%	287	173	282
Muller *et al*	Guerreiro 1997	193		193	135	190
2004 [77]						
PHB:VPS	De Silva 1996	59	53%	57	59	59
Tudur Smith *et al*	Heller 1995	119		113	119	119
2004 [76]						

[1] Number of patients for which IPD are available divided by the total number of patients in all eligible and identified trials for that particular comparison

Table 4.2. Characteristics of trials included in eight IPD systematic reviews of anti-epileptic drugs

Trial	Age (years)	Seizure requirements	Previous AED	Blinding
1.Heller 1995	≥16	≥ 2 TC seizures or partial seizures ± secondary generalisation in preceding year	Untreated	Open
2.De Silva 1996	3-16	≥ 2 TC seizures or partial seizures ± secondary generalisation in preceding year	Untreated	Open
3.Mattson 1985	18-70	Simple or complex partial or secondarily generalised TC seizures	Untreated or under treated	Double blind
4.Mattson 1992	18-70	Complex partial, secondarily generalised TC seizures, or both	Untreated or under treated	Double blind
5.Richens 1994	>16	≥ 2 generalised TC seizures or partial seizures ± secondary generalisation in previous 6m	Untreated (98%)	Open
6.Verity 1995	5-16	≥ 2 generalised TC seizures or partial seizures ± secondary generalisation in previous 6m	Untreated or seizures had recurred	Open
7.Brodie 1995a	≥13	≥ 2 partial seizures or generalised TC seizures in previous 6 months	Untreated	Double blind
8.Brodie 1995b	≥13	≥ 2 partial seizures or generalised TC seizures in previous 6 months	Untreated	Double blind
9.Reunanen 1996	>12	≥ 2 partial ± generalised TC seizures in previous 6 months	Untreated or recurrent epilepsy	Open

Trial	Age (years)	Seizure requirements	Previous AED	Blinding
10.Ramsay 1992		Newly diagnosed primary generalised TC seizures with ≥ 2 seizures within 14 days of starting study	Untreated, or seizure free without AED for 2 years before recent seizures	Open
11.Craig 1994	>60	≥1 unprovoked generalised TC seizure or ≥2 partial seizures	-	Open
12.Turnbull 1985	>16	≥2 TC and partials seizures in previous 3 years and last seizure within 3 months.	Untreated	Open
13.Placencia 1993	2-60	≥2 afebrile seizures (excluding generalised absence or myoclonus) in previous year	Untreated or previously treated	Open
14.Brodie 1999	≥65	≥ 2 seizures of any type in previous year	Untreated	Double blind
15.Bill 1997	15-91	≥2 partial seizures or generalised TC seizures in preceding 6 months	Untreated	Double blind
16.Guerreiro 1997	5-17	≥2 partial seizures or generalised TC seizures in preceding 6 months	Untreated	Double blind
17.Pal 1998	2-18	≥2 unprovoked seizures in preceding year	Untreated in previous 3 months	Open
18.Barrera 2001	2-83	≥ 2 seizures in previous 6 months with at least one partial seizure or secondarily generalised TC seizure in previous 3 months	Newly diagnosed or currently untreated	Open

Table 4.3. Availability of patient covariate data across trials

Trial	Number randomised	Age at entry n mean(sd)	Gender n % female	Epilepsy type n % partial	Number of seizures[1] n, median (25th,75th centile)	Time from 1st seizure[2] (years) n, median (25th,75th centile)	EEG n % normal	CT-EMI SCAN n % normal	Neurological signs n % yes
1.Heller 1995	243	240 **32.3(14.8)**	243 **52 %**	243 **42 %**	240 **2(2,6)**	239 **2(1,4)**	NA	NA	241 **92 %**
2.De Silva 1996	167	167 **9.9(3.6)**	167 **49 %**	167 **53 %**	167 **3(2,10)**	167 **1(1,2)**	NA	NA	167 **91 %**
3.Mattson 1985	474	471 **41.0(15.5)**	471 **12 %**	474 **100 %**	468 **1(1,3)**	470 **2(0,8)**	469 **27 %**	427 **72 %**	NA
4.Mattson 1992	480	480 **47.1(16.1)**	480 **7 %**	480 **100 %**	442 **12(4,96)**	461 **3(1,14)**	NA	NA	NA
5.Richens 1994	300	298 **33.0(14.9)**	300 **49 %**	300 **49 %**	295 **4(2,10)**	NA	NA	NA	NA
6.Verity 1995	260	247 **10.1(2.9)**	260 **53 %**	260 **42 %**	248 **3(2,6)**	229 **1(1,1)**	NA	NA	NA
7. Brodie 1995a	136	136 **34.0(15.8)**	136 **59 %**	131 **59 %**	136 **4(2,17)**	136 **1(0,3.5)**	134 **46 %**	94 **87 %**	136 **10 %**
8. Brodie 1995b	124	124 **30.0(14.1)**	124 **55 %**	115 **46 %**	124 **3(2,7.5)**	124 **1(0,1)**	118 **64 %**	92 **78 %**	124 **13 %**
9. Reunanen 1996	351	349 **32.1(14.2)**	351 **46 %**	322 **66 %**	350 **3(2,9)**	348 **1(0,4)**	26 **50 %**	21 **76 %**	351 **13 %**
10.Ramsay 1992	136	136 **20.9(14.2)**	136 **46 %**	136 **0 %**	NA	121 **1(1,1)**	NA	NA	NA
11.Craig 1994	166	163 **78.2(7.1)**	163 **56 %**	166 **48 %**	163 **3(2,6)**	NA	102 **27 %**	NA	NA
12.Turnbull 1985	140	140 **35.2(16.1)**	140 **48 %**	140 **45 %**	140 **2(1,5)**	140 **1(1,2)**	140 **50 %**	27 **63 %**	NA
13.Placencia 1993	192	192 **29.0(17.6)**	192 **65 %**	191 **70 %**	192 **2(1,4)**	192 **5(2,13)**	192 **5 %**	NA	NA

Trial	Number randomised	Age at entry n mean(sd)	Gender n % female	Epilepsy type n % partial	Number of seizures[1] n, median (25th,75th centile)	Time from 1st seizure[2] (years) n, median (25th,75th centile)	EEG n % normal	CT-EMI SCAN n % normal	Neurological signs n % yes
14.Brodie 1999	150	150 **76.9(6.0)**	150 **45 %**	138 **70 %**	150 **3(2,10)**	NA	NA	149 **42 %**	NA
15.Bill 1997	287	286 **26.8(10.7)**	287 **39 %**	286 **64 %**	287 **3(2,8)**	165 **1(0,3)**	278 **45 %**	242 **71 %**	NA
16.Guerreiro 1997	193	193 **10.5(3.1)**	193 **50 %**	190 **79 %**	193 **2(2,4)**	73 **0(0,2)**	191 **48 %**	138 **91 %**	NA
17. Pal 1998	94	94 **11.4(5.0)**	92 **49 %**	94 **64 %**	NA	92 **2.5(1,7)**	NA	NA	94 **26 %**
18.Barrera 2001	622	621 **27.2(21.4)**	622 **47 %**	622 **98 %**	622 **3(2,13)**	NA	NA	NA	NA

NA: Data not available or recorded in original trial

[1] Number of seizures in 6 months before randomisation

[2] Time from first ever seizure to randomisation (years)

4.3. Meta-analyses of epilepsy data

A log-rank analysis, stratified by trial, was the method used for the meta-analysis of each outcome within each systematic review of anti-epileptic drugs described in the previous section. The pooled hazard ratio and its 95% confidence interval and the test for homogeneity in treatment effect for each outcome within each review are presented in Table 4.4 with individual graphical displays for each comparison and each outcome summarised in Figure 4.1 (CBZ:VPS), Figure 4.2 (PHT:VPS), Figure 4.3 (CBZ:PHB), Figure 4.4 (CBZ:PHT), Figure 4.5 (PHB:PHT), Figure 4.6 (PHB:VPS), Figure 4.7 (CBZ:LTG), and Figure 4.8 (PHT:OXC).

Standard AED versus standard AED

For the comparison between standard AEDs CBZ and VPS, the log-rank analyses suggest that the drugs could have similar effectiveness for the outcome time to withdrawal (Figure 4.1), with HR and 95% CI 1.03(0.84 to 1.25), but clinically important values for hazard ratios in favour of either drug are included within the confidence interval and equivalence cannot be established. There are non-significant trends to suggest that CBZ may be more effective at reducing the time taken to achieve a period of 12 month remission and prolonging the time taken to experience first seizure, with hazard ratios and 95% CI of 1.14(0.98 to 1.34) and 0.92(0.81 to 1.06) respectively. The two standard drugs VPS and PHT appear to have similar effectiveness for all outcomes but the 95% CI are too wide to establish equivalence as clinically important differences cannot be excluded (Figure 4.2). For the comparison between CBZ and PHB (Figure 4.3), there is evidence to suggest that time to withdrawal due to adverse effects or poor seizure control is significantly shorter for PHB with HR and 95% CI of 0.68(0.52 to 0.89). The drugs have similar effectiveness for time to 12 month remission and a non-significant trend suggesting that PHB may be better at reducing time to first seizure with hazard ratios and 95% CI of 1.16(0.86 to 1.55) and 1.17(0.95 to 1.45) respectively. The two standard drugs CBZ and PHT appear to have similar effectiveness for all outcomes but the 95% CI are again too wide to establish equivalence as clinically important differences have not been excluded (Figure 4.4). For the comparison between PHB and PHT (Figure 4.5), time to withdrawal due to adverse effects or poor seizure control is significantly shorter for PHB with HR and 95% CI of 1.62(1.22 to 2.15). The drugs have similar effectiveness for time to 12 month remission and a non-significant trend

suggesting that PHB may be better at reducing time to first seizure with hazard ratios and 95% CI of 0.90(0.68 to 1.19) and 0.84(0.68 to 1.05) respectively. For the final comparison between 'standard' AEDs, time to withdrawal due to adverse effects or poor seizure control is again significantly shorter for PHB when compared to VPS with HR and 95% CI of 1.79(1.04 to 3.07). The two drugs PHB and VPS have similar effectiveness in terms of time taken to achieve a period of 12 month remission or first seizure with hazard ratios and 95% CI of 0.89(0.60 to 1.31) and 1.04(0.72 to 1.52) respectively (Figure 4.6).

New AED versus standard AED

Two new AEDs, LTG and OXC, were compared with a standard AED in two separate reviews [78], [77]. The results for the comparison between CBZ and LTG (Figure 4.7) suggest that withdrawal is significantly more likely with CBZ with a non-significant trend suggesting that time to first seizure is shorter for LTG. The hazard ratios and 95% CIs for these two outcomes are 1.72(1.29 to 2.29) and 0.86(0.69 to 1.08) respectively. These results indicate that LTG is generally better tolerated than CBZ but may not be as effective in terms of controlling seizures. For the comparison between PHT and OXC (Figure 4.8), time to withdrawal is significantly shorter with PHT (HR and 95% CI of 1.64(1.09 to 2.47)) indicating a clinical advantage for the newer drug OXC. In terms of achieving a period of 12 month remission or time to first seizure, the CIs for hazard ratios are wide and equivalence cannot be inferred with hazard ratios and 95% CIs of 0.92(0.62 to 1.37) and 1.08(0.83 to 1.40) respectively. Both comparisons of standard versus new AEDs suggest improved tolerability with newer treatments.

Heterogeneity

The chi-square test (Table 4.4 and relevant figures) suggests evidence against the null hypothesis of homogeneity in treatment effect across studies for the comparisons between CBZ and VPS (time to 12 month remission), CBZ and PHB (time to withdrawal), PHB and PHT (time to withdrawal), CBZ and LTG (time to withdrawal), and PHB compared to VPS (time to withdrawal).

For the comparison between CBZ and VPS (time to 12 month remission) there is evidence of qualitative heterogeneity with inconsistency in the direction of treatment effect between trials. Due to *a prior* beliefs about the potential for an interaction between treatment and epilepsy type for this comparison, an investigation of patient level covariates to provide a potential explanation for heterogeneity are explored in more detail within Chapter 5.

For the outcome time to withdrawal, there is evidence for heterogeneity in every comparison involving the drug PHB. Discussions with epilepsy neurologists suggest there is likely to be a strong clinical bias against PHB due to concerns over side effects and its tolerability. These clinical prejudices could have an important influence on the patient/clinicians decision to withdraw from PHB, particularly in trials that involve children. Consequently, blinding could play an important role in meta-analysis of time-to-withdrawal with more extreme treatment effects (that do not favour PHB) anticipated in unblinded studies. For the three comparisons involving PHB, only one trial (Mattson 1985 [86]) used double-blinding (Table 4.2). The heterogeneity for these comparisons appears to be caused by the De Silva 1996 [45] trial, the only trial that recruited children only. The publication for this trial report states the following "Because of the well-known behavioural and cognitive side-effects of PHB in children, the inclusion of this drug was the subject of much deliberation. We eventually decided to include the drug because it is used extensively world wide and because comparative data with other anti-epileptic drugs are lacking. However, six of the first ten children assigned this drug had unacceptable side-effects, so no further children were assigned PHB." Further exploratory analyses of factors causing heterogeneity in these particular reviews are not undertaken in this thesis primarily due to limitations of the data.

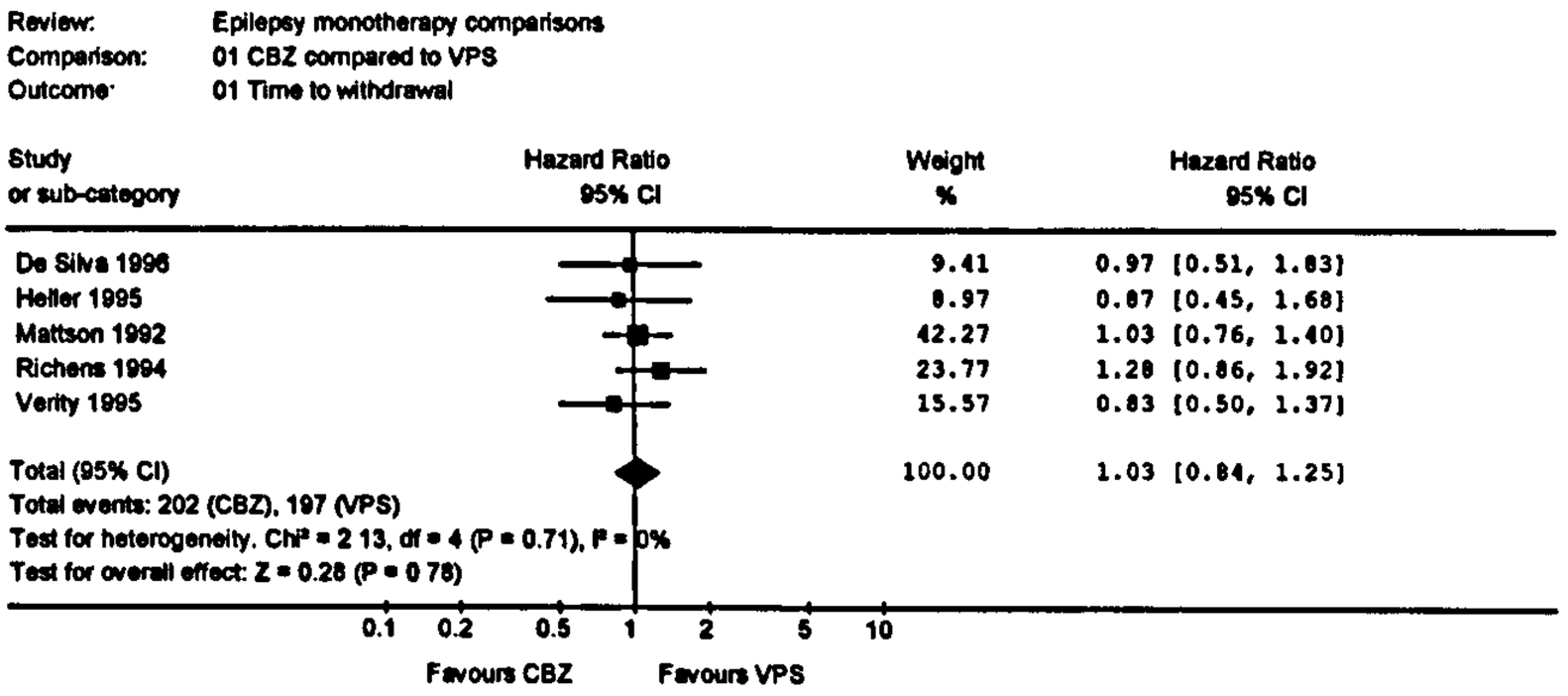
For the outcome time to withdrawal and the comparison between CBZ and LTG, there is mild evidence for heterogeneity (chi-square=6.54 (3), p=0.09, Figure 4.7) with the most extreme result in favour of LTG seen in a trial conducted in elderly patients (Brodie 1999 [87]). An age by treatment interaction or interaction effects between study drugs and other prescribed treatments could provide an explanation for heterogeneity. However, as trial results agree in terms of favouring LTG with overlapping confidence intervals across trials, factors for heterogeneity are not explored further in this thesis.

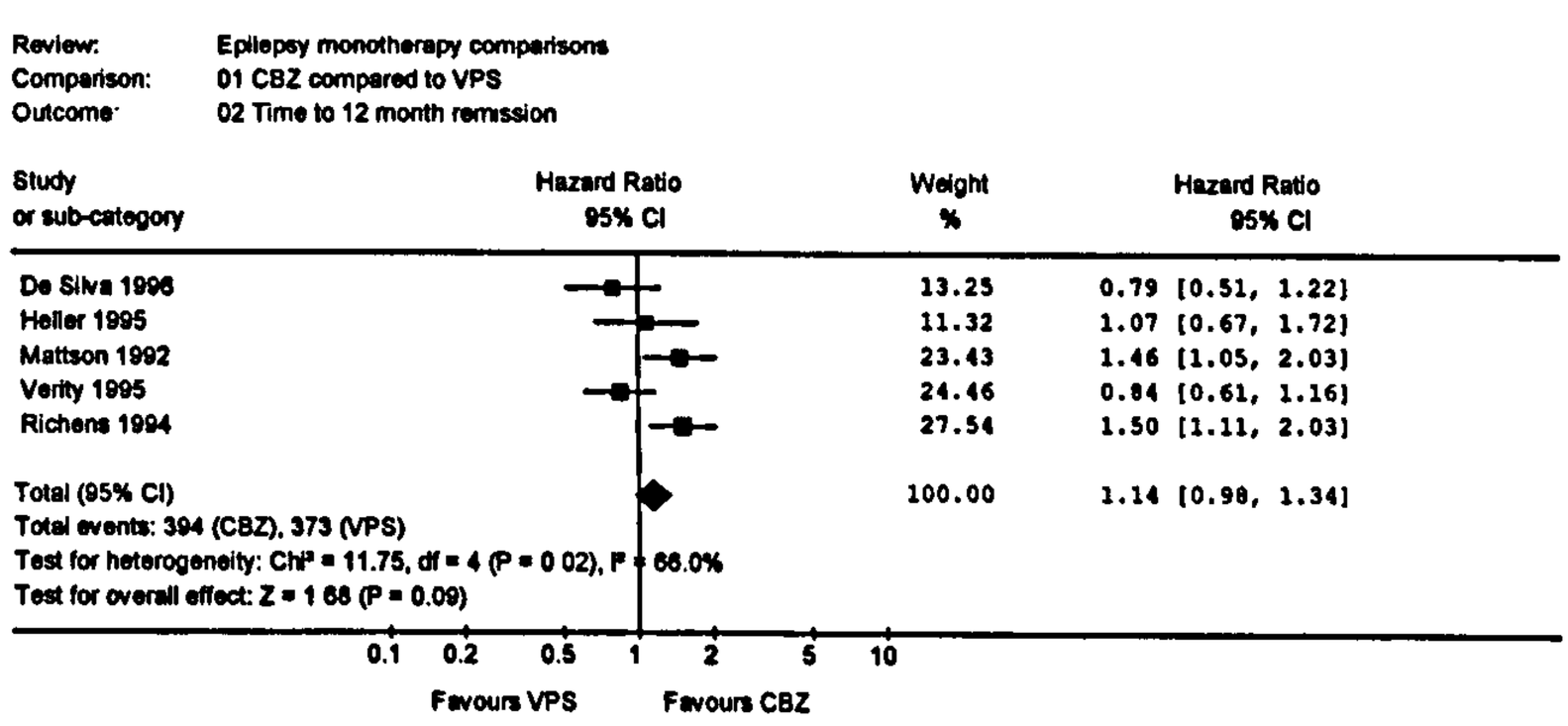
Table 4.4. Pooled hazard ratio and 95% CI from stratified log-rank analysis, and test of homogeneity of treatment effect

Comparison Examined*	Outcome					
	Time to withdrawal		Time to 12 month remission		Time to first seizure	
	Events / Total HR (95% CI)	Heterogeneity χ^2 (df) p-value	Events / Total HR (95% CI)	Heterogeneity χ^2 (df) p-value	Events / Total HR (95% CI)	Heterogeneity χ^2 (df) p-value
CBZ:VPS [39]	399 / 1200 1.03 (0.84 to 1.25)	2.13 (4) p=0.71	767 / 1225 1.14 (0.98 to 1.34)	11.75 (4) p=0.02	864 / 1225 0.92 (0.81 to 1.06)	5.89 (4) p=0.21
PHT:VPS [75]	137 / 495 1.05 (0.74 to 1.47)	5.03 (3) p=0.17	303 / 514 1.03 (0.78 to 1.36)	0.52 (3) p=0.92	371 / 639 0.96 (0.78 to 1.18)	4.24 (4) p=0.38
CBZ:PHB [72]	235 / 676 0.68 (0.52 to 0.89)	9.20 (3) p=0.03	280 / 684 1.16 (0.86 to 1.55)	3.62 (3) p=0.31	365 / 677 1.17 (0.95 to 1.45)	1.08 (3) p=0.78
CBZ:PHT [73]	196 / 546 0.99 (0.75 to 1.31)	2.27 (2) p=0.32	289 / 551 1.00 (0.77 to 1.28)	1.38 (2) p=0.50	362 / 545 1.10 (0.89 to 1.35)	2.83 (2) p=0.24
PHB:PHT [74]	211 / 499 1.62 (1.22 to 2.15)	9.33 (2) p=0.01	260 / 562 0.90 (0.68 to 1.19)	3.93 (3) p=0.27	351 / 592 0.84 (0.68 to 1.05)	2.70 (3) p=0.43
PHB:VPS [76]	66 / 170 1.79 (1.04 to 3.07)	3.70 (1) p=0.05	130 / 178 0.89 (0.60 to 1.31)	0.02 (1) p=0.90	134 / 178 1.04 (0.72 to 1.52)	0.14 (1) p=0.70
CBZ:LTG [78]	209 / 1032 1.72 (1.29, 2.29)	6.54 (3) p=0.09	DATA NOT AVAILABLE		344 / 741 0.86 (0.69 to 1.08)	2.01 (3) p=0.57
PHT:OXC[77]	91 / 480 1.64 (1.09, 2.47)	0.24 (1) p=0.62	170 / 308 0.92 (0.62, 1.37)	0.32 (1) p=0.57	229 / 472 1.08 (0.83, 1.40)	0.36 (1) p=0.55

* First AED compared to second AED with HR>1 indicating that the event of interest is more common for the first AED.

Figure 4.1. Graphical display of meta-analysis comparing CBZ to VPS





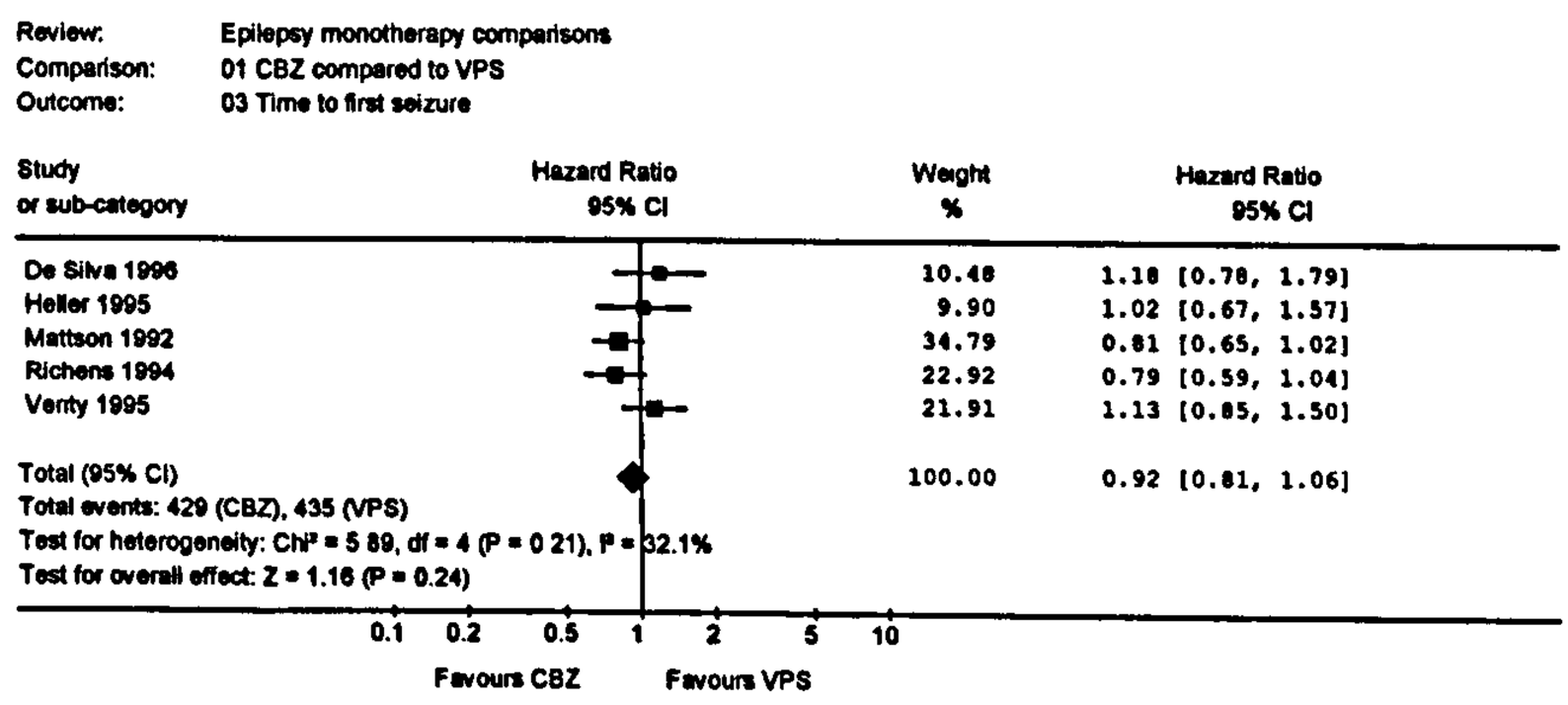


Figure 4.2. Graphical display of meta-analysis comparing PHT to VPS

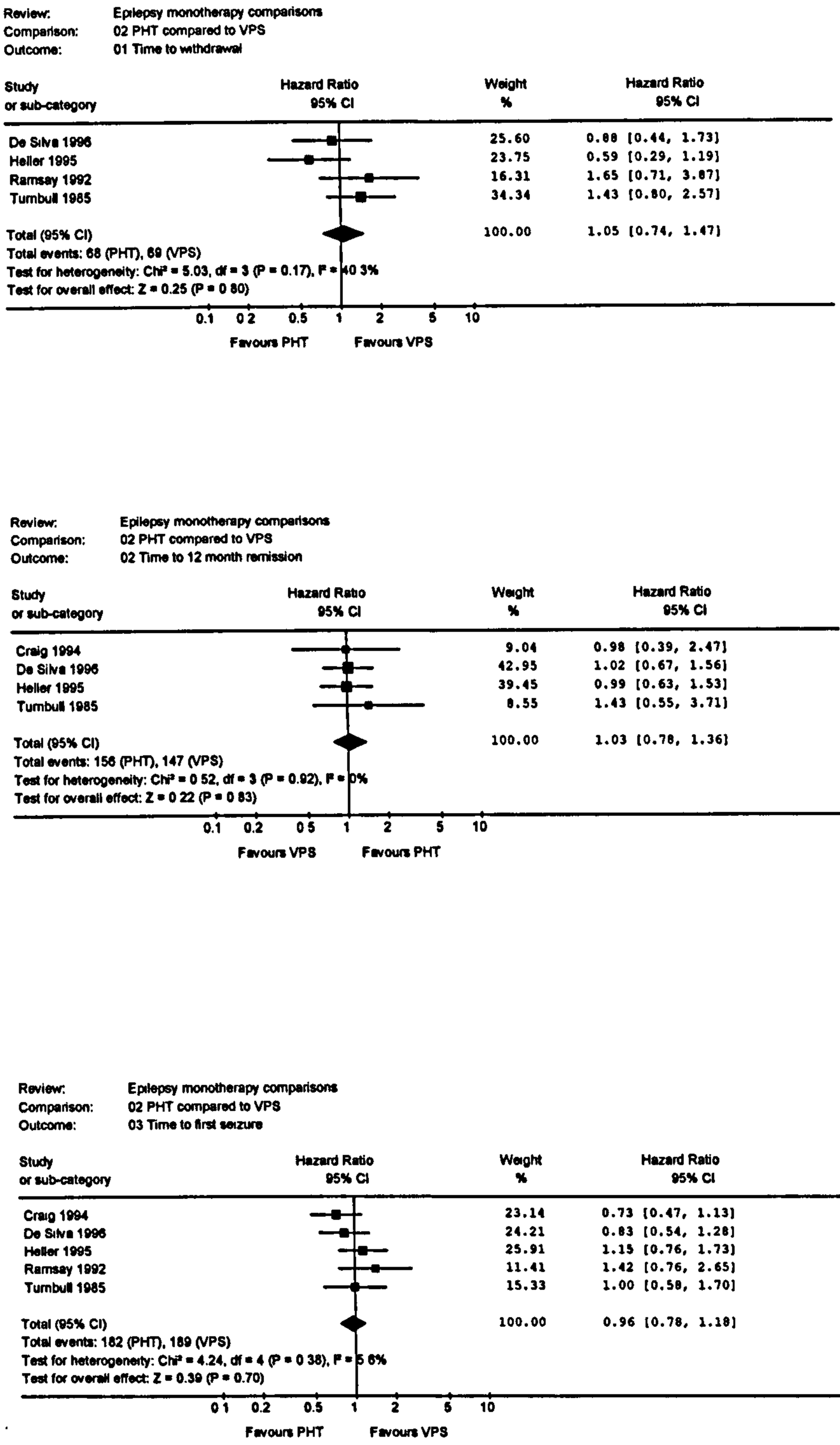


Figure 4.3. Graphical display of meta-analysis comparing CBZ to PHB

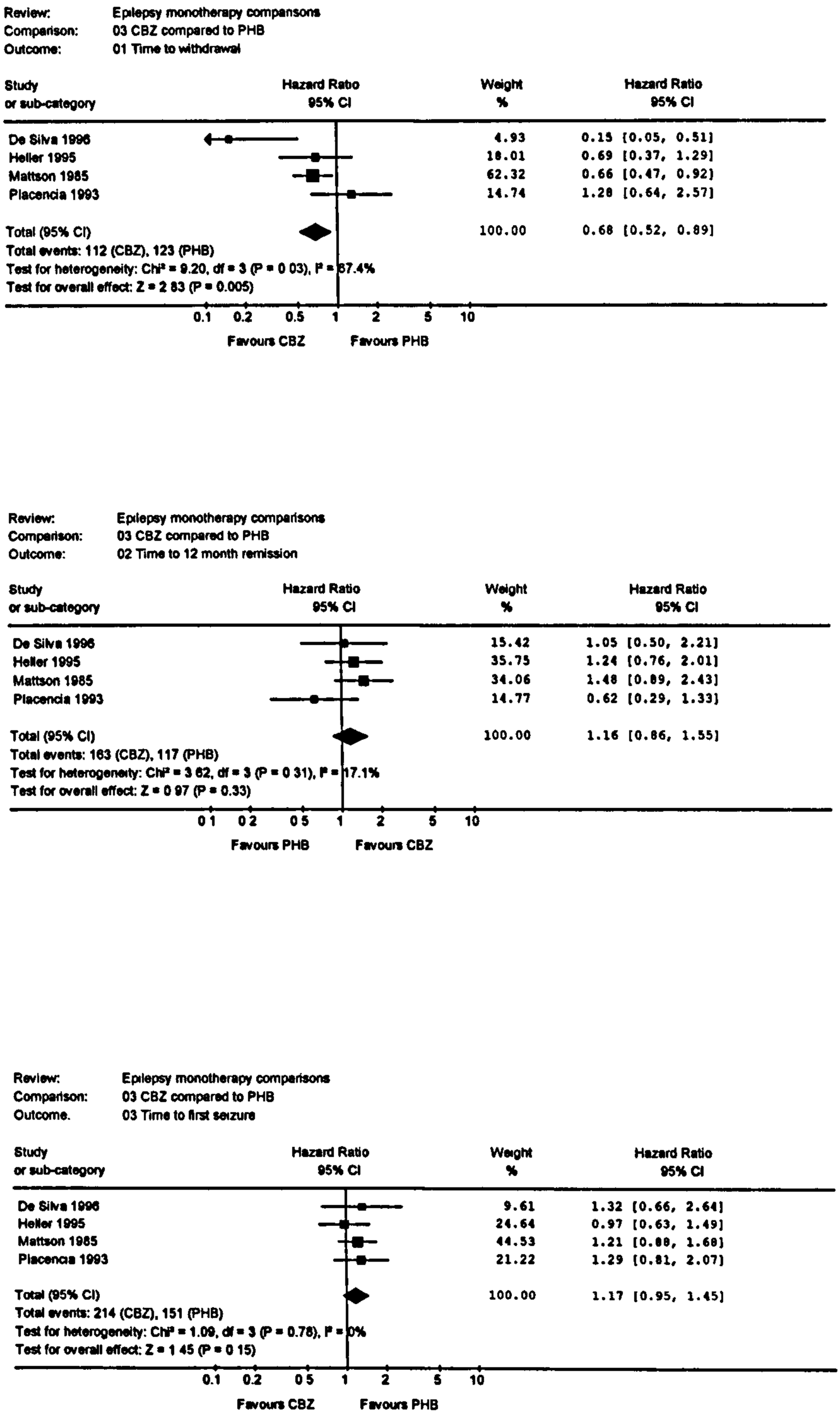


Figure 4.4. Graphical display of meta-analysis comparing CBZ to PHT

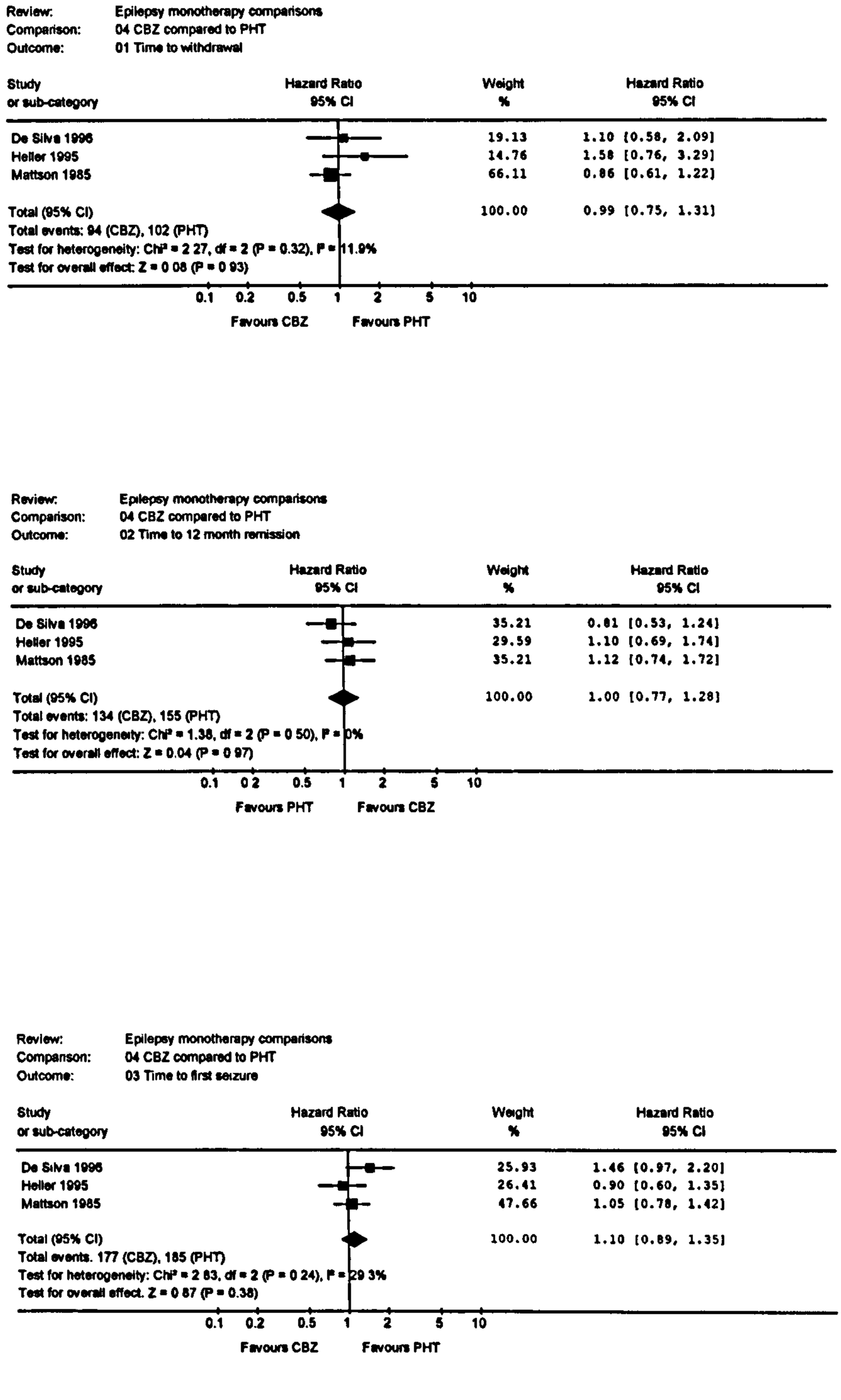


Figure 4.5. Graphical display of meta-analysis comparing PHB to PHT

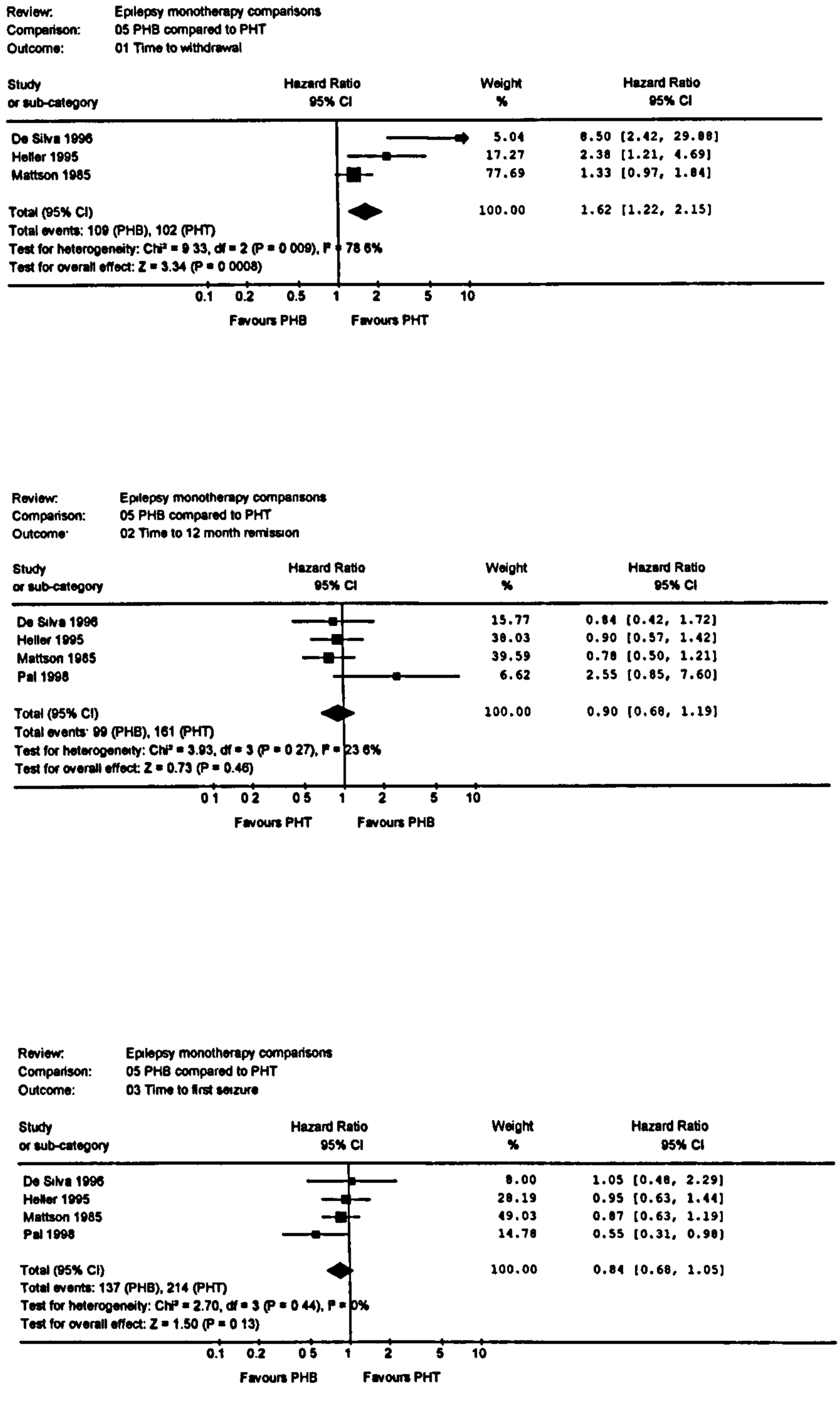


Figure 4.6. Graphical display of meta-analysis comparing PHB to VPS

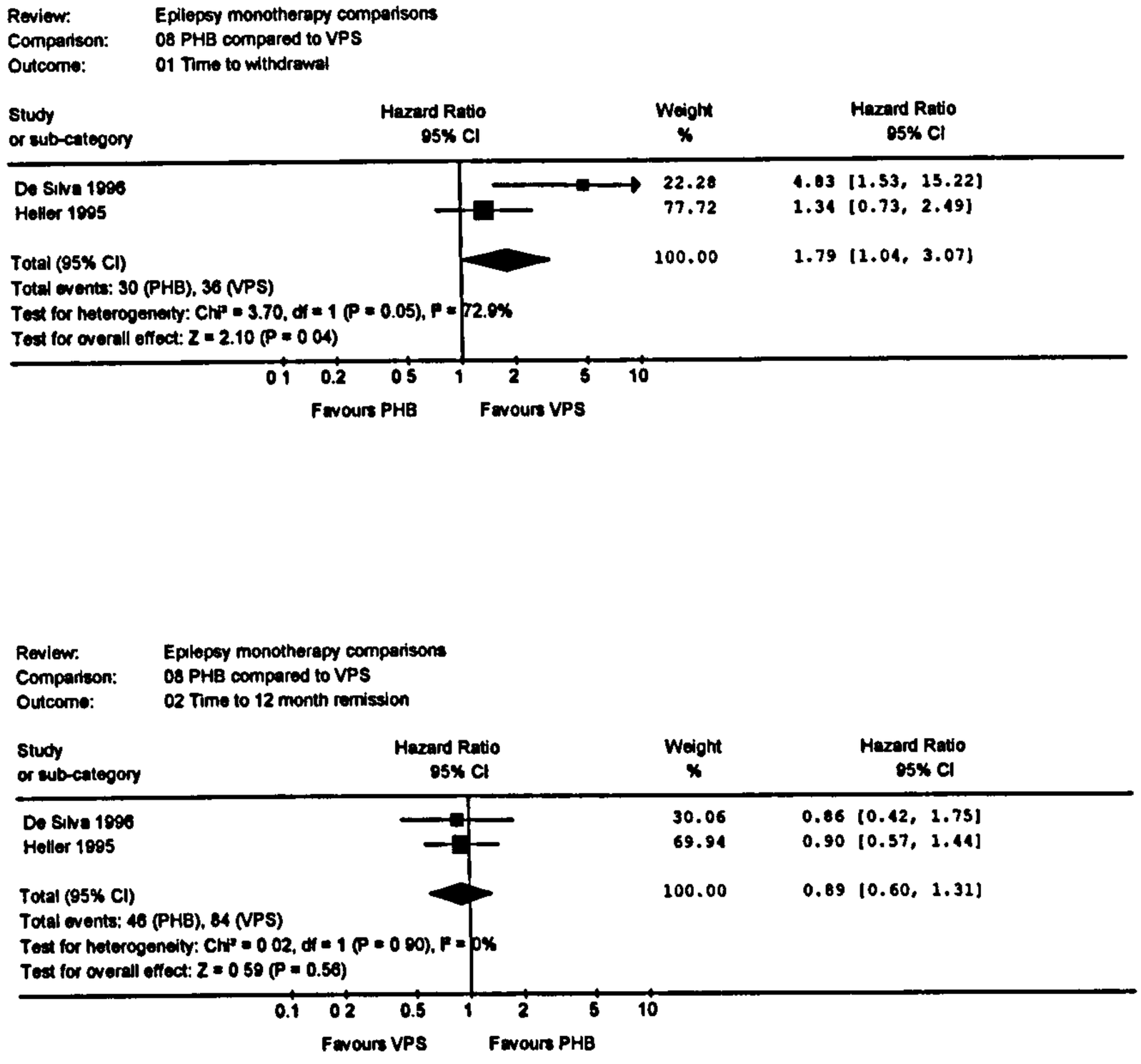
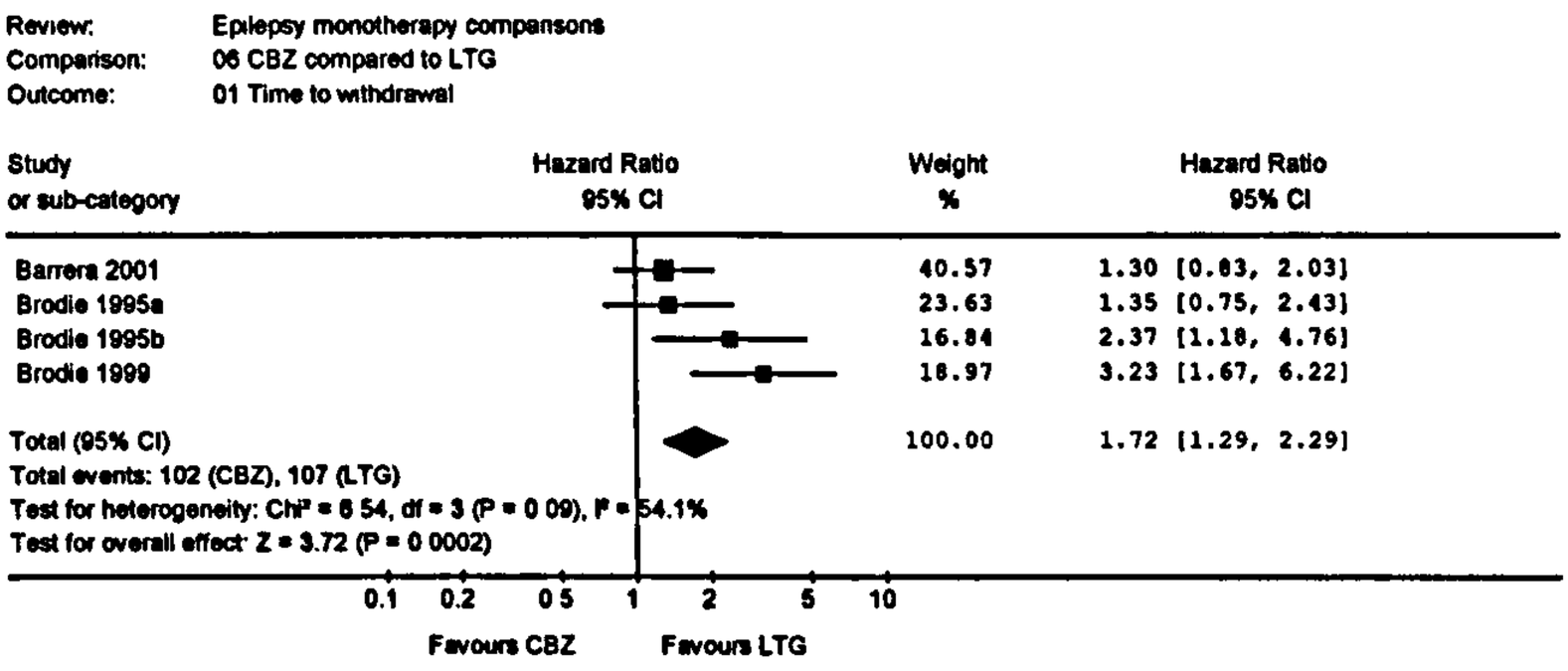


Figure 4.7. Graphical display of meta-analysis comparing CBZ to LTG



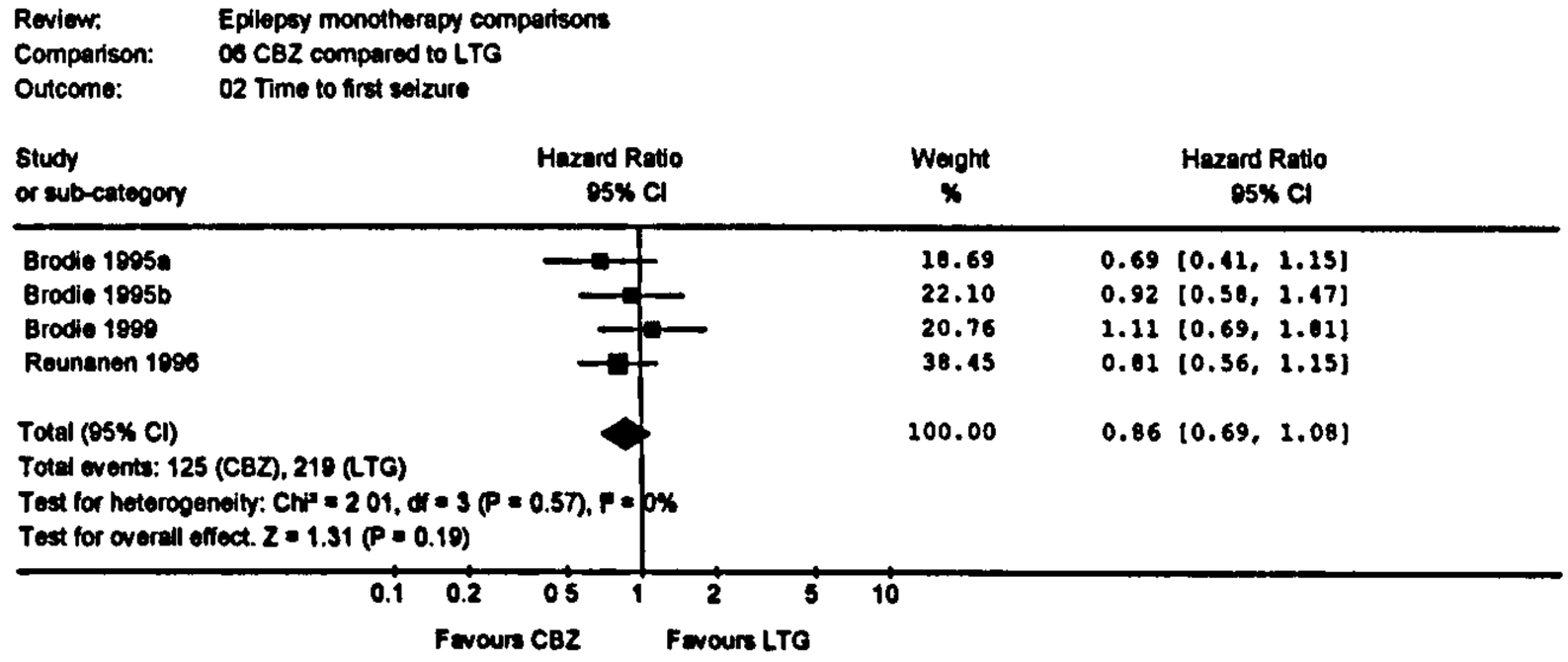
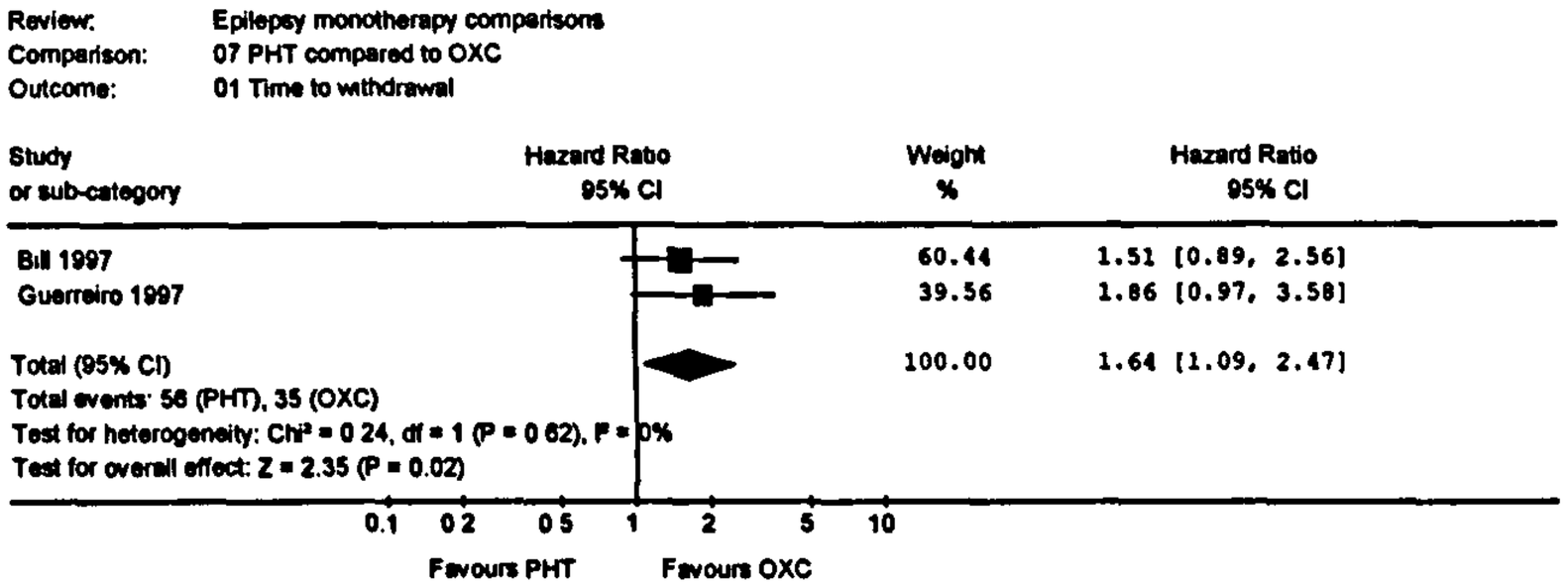
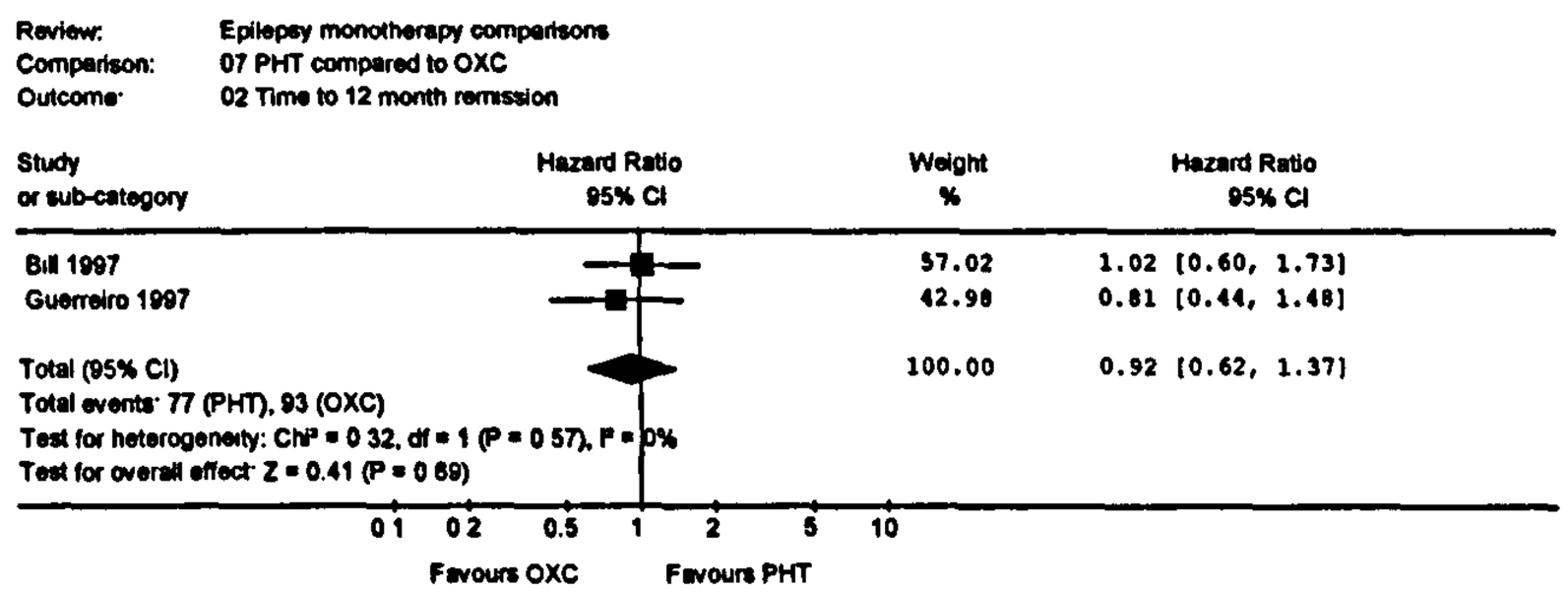
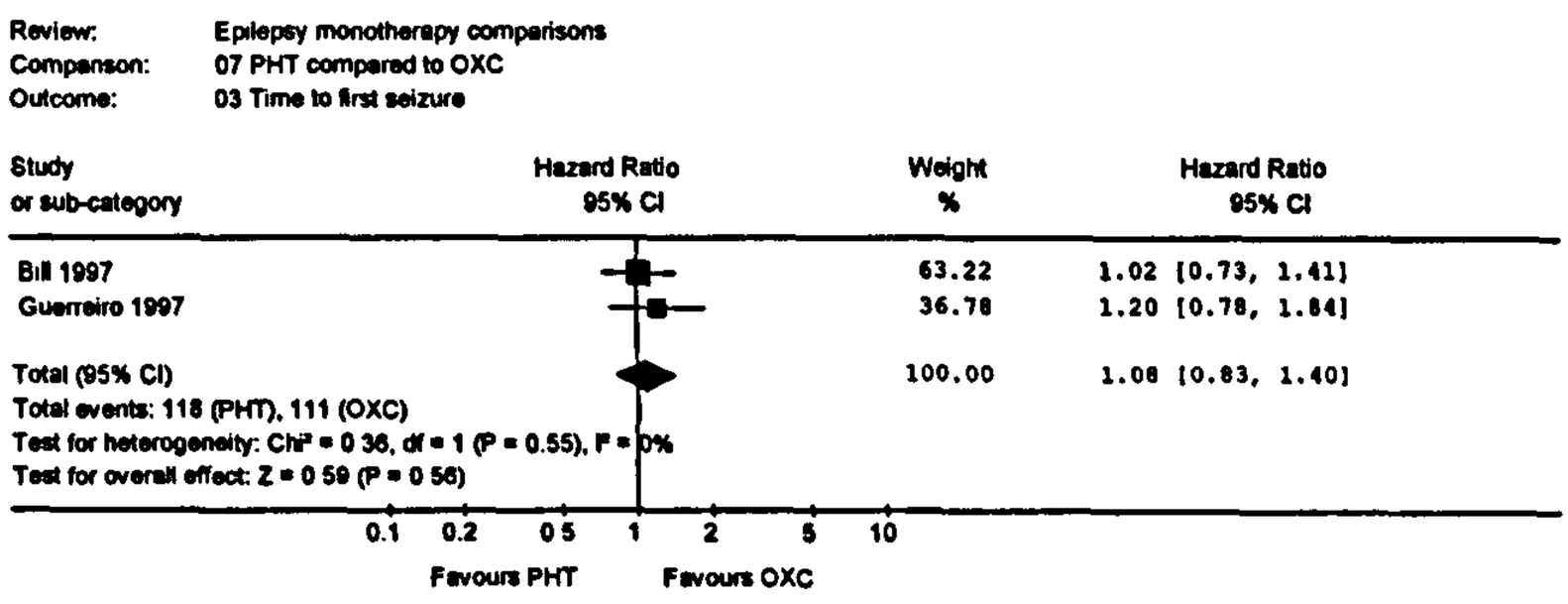


Figure 4.8. Graphical display of meta-analysis comparing PHT to OXC







4.4. Comparison of stratified log-rank analysis with Cox model

The original Cochrane protocol for each review of epilepsy monotherapy trials stated that log-rank analyses stratified by trial were to be used for primary analyses. As described using simulated data in Chapter 3 of this thesis, the results of failure time meta-analysis with IPD may differ according to the method of analysis adopted. To assess robustness of results and provide empirical evidence, a comparison of stratified log-rank and stratified Cox model analyses are undertaken for each outcome from each review in the current section. The stratified Cox model results are displayed in Table 4.5.

For the outcome time to withdrawal there is agreement in estimates of hazard ratio and 95% confidence interval limits for comparisons CBZ:VPS, PHT:VPS, CBZ:PHB, and CBZ:PHT. Values of the absolute log hazard ratio and 95% confidence limits are close to the value of 0.5 examined in the simulation study and the agreement between estimates is in keeping with results of the simulation study for this parameter value. There are discrepancies between the two approaches to meta-analysis for the comparisons PHB:PHT, PHB:VPS, CBZ:LTG, and PHT:OXC. The magnitude of treatment effect from a stratified log-rank analysis (Table 4.4) is greater for each of these comparisons, except PHT:OXC, and the degree of discrepancy between stratified log-rank and stratified Cox model results increases as the magnitude of effect (or confidence interval limits) increases as suggested by the simulation study.

For the outcome time to 12 month remission, there is much less agreement between results obtained from the two approaches with narrower confidence intervals from stratified Cox model analyses (Table 4.5). Since hazard ratios are close to 1 and there is only evidence for heterogeneity across trials within one comparison (CBZ:VPS), the discrepancy observed between methods are not entirely consistent with the simulation study results of Chapter 3. One factor not considered in the simulation study was the possibility of tied event times which can frequently occur with time-to-event data. Tied event times are an important issue for analysis of the outcome time to 12 month remission since a large proportion of patients achieve an immediate period of 12 months remission from seizures, hence the time to event for these individuals is 365 days. This may provide an explanation for the observed discrepancy between stratified log-rank and stratified Cox model analyses and should be explored in more detail by

Table 4.5. Pooled hazard ratio and 95% CI from Cox proportional hazards model stratified by trial (using Efron method for handling ties)

	Outcome		
	Time to withdrawal	**Time to 12 month remission**	**Time to first seizure**
Comparison examined*	Events / Total HR (95% CI)	Events / Total HR (95% CI)	Events / Total HR (95% CI)
CBZ:VPS	399 / 1200 1.03 (0.84, 1.25)	767 / 1225 1.14 (0.99, 1.32)	864 / 1225 0.92(0.81, 1.06)
PHT:VPS	137 / 495 1.05 (0.75, 1.47)	303 / 514 1.04 (0.83, 1.31)	371 / 639 0.96 (0.78, 1.18)
CBZ:PHB	235 / 676 0.68 (0.52, 0.89)	280 / 684 1.08 (0.84, 1.38)	365 / 677 1.17 (0.94, 1.45)
CBZ:PHT	196 / 546 0.99 (0.75, 1.31)	289 / 551 1.01 (0.80, 1.27)	362 / 545 1.10 (0.89, 1.35)
PHB:PHT	211 / 499 1.61 (1.21, 2.12)	260 / 562 0.94 (0.73, 1.22)	351 / 592 0.84 (0.68, 1.05)
PHB:VPS	66 / 170 1.75 (1.03, 2.95)	130 / 178 0.91 (0.62, 1.33)	134 / 178 1.05 (0.72, 1.52)
CBZ:LTG	209 / 1032 1.68 (1.27, 2.21)	DATA NOT AVAILABLE	344 / 741 0.86 (0.69, 1.08)
PHT:OXC	91 / 480 1.65 (1.08, 2.52)	170 / 308 0.92 (0.68, 1.24)	229 / 472 1.08 (0.83, 1.40)

* First AED compared to second AED with HR>1 indicating that the event of interest is more common for the first AED

extending the previously described simulation study. The issue of tied event times will be discussed in more detail in Chapter 5.

For the outcome time to first seizure, the point estimate of hazard ratio and limits for the 95% confidence interval are identical for each comparison using stratified log-rank and stratified Cox regression model analyses. For this outcome, there is no evidence for heterogeneity across individual trial results within each comparison (Table 4.4) and the estimates of hazard ratio do not deviate substantially from the null value. These results are consistent with the simulation study results of Chapter 3.

4.5. Missing Data

For various reasons IPD may not be available from all trials identified as eligible. For the epilepsy monotherapy reviews, the percentage of IPD (number of patients for whom data are available as a percentage of total number of eligible patients) obtained varies between 53% and 100% across these reviews (Table 4.1). The most common reasons why IPD were not available for a particular trial included lost or destroyed data, no seizure data recorded, and unwillingness of the trialist to collaborate. For two trials ([48] and [45]), data for reason and date of withdrawal from treatment were not computerised. The data were however extracted independently from study case report forms by two reviewers (Paula Williamson, Anthony Marson). The five reviews that obtained less than 90% of IPD are likely to be more prone to bias compared to the reviews which include over 90% of the data available for eligible patients. Missing data for individual patients could threaten the validity of meta-analysis results and conclusions particularly if trialists are more willing to provide IPD based on the statistical significance of original results or their personal concerns about quality of the trial methodology.

For the epilepsy reviews, eligible trials for which IPD were not provided were examined to determine whether aggregate data could be extracted to allow estimation of summary statistics required to undertake meta-analyses for the time-to-event outcomes of interest. For the majority of trials there were insufficient reported data and these trials did not contribute to overall analyses. As insufficient aggregate data is often a reason for the initial decision to collect IPD, this situation is likely to be common in systematic reviews

of IPD. Some level of data could be extracted for two small trials but sensitivity analyses including these results with trials providing IPD made very little difference to meta-analysis results and conclusions. Further details may be found in the full report of each systematic review.

The author would recommend that in the absence of IPD, aggregate data should be extracted and sensitivity analyses undertaken to check that the trial results for which IPD are not available are consistent with trial results where IPD are available. If inconsistencies are identified, this may indicate the potential for bias within the set of IPD trials. If sufficient AD for the outcome of interest are not available to allow such sensitivity analyses, it may be possible to assess robustness of results in terms of other outcomes. However, due to the potential for within-study selective reporting of outcomes this approach may not be particularly helpful. Missing data in IPD reviews is certainly an area that requires further research. A systematic review of IPD reviews to assess factors such as reasons for missing IPD, strategies for dealing with missing IPD and an assessment of the potential for bias introduced by missing data would be particularly helpful.

4.6. Discussion

For the epilepsy reviews explored within this chapter, individual patient data were essential due to the unavailability of sufficient aggregate data and inconsistencies across trials in defining the time-to-event outcomes of interest. The overall result for each comparison explored in each review are presented in this thesis with further detail regarding clinical interpretation and implications provided in the following Chapters and within the full text of each review which are obtainable from the Cochrane Library.

The entire process of collecting IPD generally involves greater resources and time compared to the process of extracting aggregate data. The time involved is likely to depend on a number of factors. Clearly, as the number of eligible trials increases so will the resources required. The publication date of included trials could have some impact as the authors and trialists of older trials may be harder to trace, and may be more likely to have lost or destroyed the data. The chances of obtaining data in a suitable computerised format may also be less likely for older trials. For some of the epilepsy monotherapy trials the original individual patient data were only available as hard copy

and required data to be computerised by the review team. For two further trials, computerised data were only available in the format of a non-standard computerised database that required additional software and time for database tuition. Data cleaning, and the process of defining each time-to-event outcome, involved a substantial length of time due to the complexities of the data involved and different methods for recording seizures across trials. This aspect of the IPD process may be more time consuming than in other therapeutic areas, such as cancer, where the outcome is time to death. However, for the latter example, the collection of additional follow-up information is likely to introduce further demands on resources. Additional follow-up data collected after trial closure was rarely an issue in the epilepsy monotherapy trials. As the time and resources required to undertake a systematic review with IPD is likely to vary according to example, this issue should be considered in the overall evaluation of whether IPD is worthwhile. As a future research project, a simple survey involving authors of completed IPD reviews to enquire about costs, length of time taken, and where the majority of time was required would be particularly useful.

It is possible, and quite likely, that data are not available for every individual from every trial in a systematic review and meta-analysis based on IPD. For example, data for an entire trial may have been lost or destroyed, or follow-up information for the outcome of interest may not have been collected for a subset of individuals within a particular trial. Data may therefore be missing at the individual patient level or at the trial level. Analyses based on such incomplete data may be biased in a similar way to analyses that are based only on published information (publication bias) or non-intention to treat analyses. The available individual patient data as a proportion of patients from eligible trials varied in each of the epilepsy reviews examined and could threaten the validity of results if the proportion of missing data are extensive. However, since aggregate data were also not available for the outcomes of interest in the trials that did not provide IPD, the IPD based analyses encompass the entire currently available data. Further research is required to explore these issues in more detail.

For some systematic reviews, both IPD and AD may be available and an overall meta-analysis required. On a simple level, this may be achieved by combining IPD and AD estimates using an inverse variance weighted average. This approach is appealing as it is simple to implement and is useful for comparing the degree of contribution of each trial

by examining respective weights. Furthermore, trial estimates based on potentially less accurate AD approaches, such as the survival curve approach, may be down-weighted in the overall meta-analysis, or sensitivity analyses undertaken if there are particular concerns regarding any of the AD based estimates. A more complex approach may be possible by including both levels of data in a single multi-level model or by adopting a hierarchical Bayesian approach.

CHAPTER 5

Modelling Heterogeneity

Studies included in a meta-analysis will usually differ in terms of methodological characteristics, design features, clinical procedures and characteristics of included patients. These factors, as well as some unknown characteristics, can contribute to variability in the treatment effect between studies in a meta-analysis. This variation in treatment effect is termed *statistical heterogeneity* but is more commonly referred to as just *heterogeneity*. As an example, suppose there is variation in the participant age groupings of studies in a meta-analysis comparing an experimental treatment with a standard control. If the experimental treatment has increased benefit for older participants, the underlying age by treatment interaction could manifest itself as a treatment by study interaction (statistical heterogeneity) such that the treatment effect is different across trials. Exploring heterogeneity during the meta-analysis process is extremely important as an interpretation of overall results in the presence of statistical heterogeneity can be misleading as the incompatibility of effects could indicate incompatible populations from which relating an overall combined analysis to specific populations would be difficult. A sensible investigation of sources of heterogeneity should increase both the scientific and the clinical relevance of the results of meta-analyses [13] and discovering potential explanations for statistical heterogeneity can be clinically informative. In the example described above, detecting the presence of heterogeneity and exploring potential causes may result in discovering some evidence for an age by treatment

interaction to be confirmed in further trials. This information may be valuable for deciding whether future patients from certain age groups would benefit more from the experimental treatment than others; valuable information for treatment policy decision making and for the individual patient.

Regression modelling is a popular approach for investigating heterogeneity between studies and to assess the effect of important characteristics. If IPD are available, patient as well as study level characteristics may be included in the regression model and relationships with treatment and the impact on heterogeneity can be investigated [88]. With AD, study level characteristics are included in the model and relationships with the relative treatment effect in a trial are examined. Alternative model structures and approaches for parameter estimation are available and differ depending on whether fixed or random treatment effects are assumed and whether IPD or AD are used in the model. In the meta-analysis literature the term meta-regression is used to refer to this general class of models. Although investigating heterogeneity is possible with both AD and IPD, a simulation study undertaken by Lambert *et al* [89] showed that the statistical power of meta-regression using aggregate binary data was dramatically and consistently lower than that of the corresponding IPD analysis with little agreement between the parameter estimates obtained from the two methods. They conclude that IPD is required for a reliable exploration of heterogeneity. Although meta-regression with aggregate data can be useful, the approach has limitations for exploring patterns between treatment effects and patient characteristics. In this case, the meta-regression model is based on using trial-level averages of individual characteristics such as the mean age of participants in a trial, rather than age of individuals in the trial. This has implications for interpretation because relationships with patient averages across trials may not be the same as the relationship for patients within trials [126] and can only be reliably explored with IPD.

General aggregate data meta-regression models are suitable for time-to-event outcomes if an estimate of the log hazard ratio and its variance and suitable aggregate covariate data are available for each trial. With IPD, the fixed effect Cox regression models described in Chapter 3 can be used to detect and explore heterogeneity in meta-analysis. Potential causes of heterogeneity in the meta-analysis of epilepsy trials comparing CBZ and VPS were investigated using a Cox regression model with fixed treatment effect and

fixed trial indicator variables by Williamson *et al* [90]. The basic aim of the modelling approach to exploring heterogeneity is to try and explain what may cause incompatible treatment effects across trials by incorporating covariates (trial or patient level) into the model. However, the included covariates may not provide an adequate explanation and there may be variation left unexplained, called residual heterogeneity which can be allowed for by adopting a random treatment effects approach. For aggregate data, random effects meta-regression models are already developed and can be easily fitted with appropriate software [91]. Individual patient data offer greater flexibility for meta-analysis and improve the potential to investigate and explain heterogeneity thoroughly. When IPD are available, meta-analysis can be undertaken using a hierarchical framework. Several such models have been described previously for the meta-analysis of binary [59], continuous [60], and ordinal outcomes [61]. The gap identified in the literature describing suitable models for analysing time-to-event data and modelling heterogeneity motivated much of the work undertaken in the current chapter.

In the general assessment of how meta-analyses based on AD and IPD compare, there is clearly a need to include a comparison between results and conclusions obtained from exploring heterogeneity using these two data types. However, many factors could potentially contribute to an overall difference between analyses using AD and IPD [49], [52]. For instance, IPD may include additional follow-up information for some patients, authors may have excluded certain patients in the AD presented in a study report, or AD may not be available for the patient characteristics of interest thus limiting the investigation into sources of heterogeneity. Such potential differences within the overall patient-level data set could introduce a further level of complexity to the comparison between AD and IPD heterogeneity investigations. To enable a comparison of methods would require that the models be based on exactly the same data, which could be accomplished by generating suitable AD from the IPD. An alternative more pragmatic comparison would examine explorations based on available IPD compared with AD extracted from trial reports.

This chapter has two aims. The first aim is to develop and extend the current methodology for fitting alternative random effects Cox proportional hazards models that would be appropriate for meta-analysis and exploring heterogeneity with IPD. These alternative models are compared using simulated and empirical data. The second

aim is to apply existing methodology for exploring heterogeneity with AD in order to compare empirically results and conclusions drawn from IPD and AD based investigations.

5.1. Exploring heterogeneity with individual patient data using the Cox model

The 'fixed effect' Cox models described in Chapter 3 (model 3.7 and 3.8) can be used for undertaking meta-analysis and for detecting and exploring possible sources of heterogeneity. A large body of literature is available discussing the general issue of 'heterogeneity' (in terms of variability between individuals or groups of individuals) and random effects models for the analysis of failure time data. See for example [92], [93], [94], [95]. In this setting, the random effect is a continuous variable that describes excess risk or frailty for distinct categories, such as individuals or families [96]. The principle behind these frailty models is that individuals have different frailties, and that those who are most frail will die (if the event of interest is death) earlier than the others [92]. In the context of meta-analysis, since interest is usually focused on heterogeneity in treatment effects across trials rather than heterogeneity at the individual level, standard frailty models are not entirely appropriate. However, since the data structure of meta-analysis is similar to that of a multi-centre clinical trial the literature in this latter area is relevant to meta-analysis. Alternative random effect formulations of the Cox proportional hazards model suitable for exploring heterogeneity in meta-analysis are proposed in the following sections. The models differ in how they accommodate for the effects of both trial and treatment.

5.1.1. Fixed trial and treatment effect (FE/FE)

The Cox model with fixed trial effects represented by indicator variables and fixed treatment effect described in Chapter 3 (model 3.7) can be used to assess the assumption of homogeneity in treatment effect across trials by including treatment-trial interaction indicator variables such that

$$\lambda_{ij}(t) = \lambda_0(t)\exp(\beta_{0j} + \beta_1 x_{1ij} + \beta_{2j} x_{2ij}) \qquad \textbf{(5.1)}$$

where $x_{2ij} = 1$ if the *ith* individual belongs to the experimental treatment group of the *jth* trial (with β_{21} constrained to equal zero) and $x_{2ij} = 0$ otherwise. A formal test for treatment-trial interaction (i.e. statistical heterogeneity) can be obtained by comparing the value *–2*log(likelihood)* of models (3.7) without interaction terms, and (5.1). Under the null hypothesis of no heterogeneity, this statistic follows approximately a chi-square distribution on *J-1* degrees of freedom where *J* denotes the total number of trials. Further covariates may also be included in the linear predictor of models (3.7) and (5.1) and potential explanations for heterogeneity in treatment effect can be explored by inspecting the importance of the treatment-trial interaction after inclusion of important clinical factors which may include treatment-covariate interactions.

5.1.2. Stratified by trial with fixed treatment effect (SFE/FE)

Model (3.8), described in the third chapter, is also suitable for assessing the evidence for heterogeneity in the treatment effect across trials by fitting a stratified model with trial specific treatment effects such that

$$\lambda_{ij}(t) = \lambda_{0j}(t)\exp(\beta_{1j}x_{1ij}) \qquad (5.2)$$

A formal test for heterogeneity can be obtained by comparing the value *-2*log(likelihood)* of models (3.8) and (5.2). Under the null hypothesis of no heterogeneity, this statistic follows approximately a chi-square distribution on *J-1* degrees of freedom.

Residual heterogeneity is not allowed for in the above models that assume treatment effect is fixed across trials.

5.1.3. Fixed trial effect and random treatment effects (FE/RE)

The third model considered appropriate for modelling heterogeneity (or undertaking meta-analysis) includes random treatment effects and does make allowance for residual heterogeneity. The hazard function for the *ith* individual in the *jth* trial is written as

$$\lambda_{ij}(t) = \lambda_0(t)\exp(\beta_{0j} + \beta_{1j}x_{1ij}) \qquad (5.3)$$

$$\beta_{1j} = \beta_1 + b_{1j}$$

$$b_{1j} \sim N(0,\tau^2)$$

where the fixed parameters β_{0j} (with β_{01} constrained to equal zero) indicate the trial membership for all individuals in the *jth* trial. The coefficient β_1 can be interpreted as the average log hazard ratio for a population of possible treatment effects, and b_{1j} is the deviation of the log hazard ratio in the *jth* trial from this population average. The random quantities b_{1j} are assumed to follow a Normal distribution with mean zero and variance τ^2 which is a measure of the between trial variability in treatment effect i.e. a measure of the degree of statistical heterogeneity. This model suffers from the same limitation as model (5.1), that the hazards are assumed to be proportional to the same common baseline hazard function λ_0. An alternative formulation that does not require such an assumption is given by model (5.4).

5.1.4. Stratified by trial with random treatment effects (SFE/RE)

In this stratified version of model (5.3) the hazard function for the *ith* individual in the *jth* trial is written as

$$\lambda_{ij}(t) = \lambda_{0j}(t)\exp(\beta_{1j}x_{1ij}) \qquad (5.4)$$

$$\beta_{1j} = \beta_1 + b_{1j}$$

$$b_{1j} \sim N(0,\tau^2)$$

where λ_{0j} is the baseline hazard function in the *jth* trial. As with model (3.8), the stratified model with fixed treatment effect, the hazards are only assumed to be proportional within each trial. The parameters β_1 and b_{1j} are interpreted in a similar way to parameters of model (5.3).

5.1.5. Random trial effects and random treatment effects (RE/RE)

In the final model considered in this thesis the hazard function for the *ith* individual in the *jth* trial is written

$$\lambda_{ij}(t) = \lambda_0(t)\exp(b_{0j} + \beta_{1j}x_{1ij}) \qquad (5.5)$$

$$\beta_{1j} = \beta_1 + b_{1j}$$

$$b_{0j} \sim N(0, \sigma^2)$$

$$b_{1j} \sim N(0, \tau^2)$$

$$\mathrm{cov}(b_{0j}, b_{1j}) = 0$$

The coefficient β_1 is interpreted as the average log hazard ratio for a population of possible treatment effects, and b_{1j} is the deviation of the relative treatment effect in the *jth* trial from this population average. The random quantities b_{0j} represent the deviation in the *jth* trial from the overall underlying baseline risk and these random effects are assumed to follow a Normal distribution with mean zero and variance σ^2 representing the variation in baseline risk across trials. In the example used for illustration in later sections of this Chapter, the meta-analysis includes only 5 trials, and the covariance between b_{0j} and b_{1j} is therefore assumed to be zero. A coding structure of ± ½ is used for the treatment indicator variable x_{1ij} in order to impose equal between-trial variances in the hazard function for the experimental and control group. Under this treatment coding structure, the random quantities b_{0j} represent the deviation in the *jth* trial of the risk from the average of the two treatments. Further discussion regarding the covariance structure in models with random trial and treatment effects for binary outcomes is given by Turner *et al* [59] and general remarks about treatment coding in random effects models are given in section 5.3 of this thesis. The inclusion of random trial effects assumes that the trials included in the meta-analysis are a random sample from a larger population of trials. Such an assumption may not be appropriate in the context of meta-analysis due to the underlying principle that all relevant trials are identified and included in the meta-analysis.

The FE/FE model makes an assumption of proportional hazards across trials such that the hazards of event in each trial are proportional to the same common underlying hazard function $\lambda_0(t)$. In the authors opinion this assumption that baseline hazards should have the same pattern with time is too restrictive and unnecessary for meta-analysis. The SFE/FE model only assumes proportional hazards within each trial which is a more realistic assumption for meta-analysis particularly if dealing with an active control group where variation in factors such as dose and timing could introduce variation in the baseline hazards. For the assumption of a fixed treatment effect the current author would advocate the use of a SFE/FE model for meta-analysis. In terms of models assuming random treatment effects the FE/RE model assumes hazards are fixed and proportional across trials, the SFE/RE model assumes hazards are fixed and proportional within trial and the RE/RE model assumes hazards are random effects but proportional across all trials. Following the same argument as for models assuming fixed treatment effects, the restrictive assumption of proportional hazards across all trials would lead the author to advocate use of the SFE/RE model for meta-analysis assuming random treatment effects.

5.2. Parameter estimation

For the unstratified fixed effect proportional hazards model FE/FE (model 3.7), the maximum likelihood estimates of the parameters are obtained, using numerical methods, by maximizing the usual partial log-likelihood function given by

$$\log L(\beta) = \sum_{i=1}^{n_j} \sum_{j=1}^{J} \delta_{ij} \left\{ \eta_{ij} - \log \sum_{l \in \Re_{ij}} \exp(\eta_l) \right\}$$

for individuals $i\,(i = 1,\ldots,n_j)$ within trial $j\,(j = 1,\ldots,J)$ where $\eta_{ij} = \sum_{k=1}^{p} \beta_k x_{kij}$, δ_{ij} is a censoring indicator taking the value unity if a failure occurs at time t_{ij} and zero otherwise, and $\Re_{ij}$ represents the risk set at time t_{ij}. For the stratified model SFE/FE (model 3.8) the overall log-likelihood is equal to the sum of within trial partial log-likelihood terms such that

$$\log L(\beta) = \sum_{j=1}^{J} \left\{ \sum_{i=1}^{n_j} \delta_i \left\{ \eta_i - \log \sum_{l \in \Re_i} \exp(\eta_l) \right\} \right\}$$

Standard statistical software packages (e.g. SAS, S-plus, STATA) capable of analysing time to event data can be used to fit the fixed effect models FE/FE (model 3.7 and 5.1) and SFE/FE (model 3.8 and 5.2). A Cox regression model with random effects can be fitted using the *frailty* option of the *coxph* function within the S-plus (or R) software package. This function uses the penalized partial likelihood approach with variance of random effects based on an approximate REML equation. Further discussion regarding the computational algorithm is given in the technical report written by the authors of the function [97]. Such a model with random effects defined by shared frailty for each trial equates to a Cox model with random trial effects in the context of the models described in section 5.1 (model with random trial effects and fixed treatment effect not shown). However, to the authors' knowledge, no facility is currently available to extend this model to include random treatment effects. As the model with random trial effects and fixed treatment effect is not examined in this thesis the *frailty* option of *coxph* is not considered further. A brief examination of alternative software packages that fit a Cox proportional hazards model with some form of random effect revealed that none could accommodate fitting a stratified Cox model including a random treatment effect (SFE/RE).

Yamaguchi and Ohashi [62] describe an approach to estimate the parameters of a proportional hazards regression model including random centre and random treatment effects (equivalent to model RE/RE (5.5) of the present chapter with trial defined as centre) to analyse data from a multi-centre clinical trial. The approach they outline (summarised below using similar notation) is an extension of the method described by McGilchrist and Aisbett [98] and McGilchrist [99]. The approach involves firstly maximising the penalized partial likelihood $l = l_1 + l_2$ to obtain Best Linear Unbiased Predictor (BLUP) estimators of β_k and b_{uj}, $u = 0,1$. The two likelihood components l_1 and l_2 are given by

$$l_1 = \sum_{i=1}^{n_j} \sum_{j=1}^{J} \delta_{ij} \left\{ \eta_{ij} - \log \sum_{l \in \Re_{ij}} \exp(\eta_l) \right\}$$

$$l_2 = -\frac{1}{2} \sum_{u=0,1} \left(J \log 2\pi\theta_u + \sum_{j=1}^{J} \frac{b_{uj}^2}{\theta_u} \right)$$

where θ_0, θ_1 are equivalent to σ^2 and τ^2 in the notation used for model (5.5) and $\eta_{ij} = \sum_{k=1}^{p} \beta_k x_{kij} + b_{0j} + b_{1j} x_{1ij}$ in the random trial, random treatment effect model. Letting

$$\beta = (\beta_1, \beta_2, \ldots, \beta_p)', \quad b_0 = (b_{01}, b_{02}, \ldots, b_{0J})', \quad b_1 = (b_{11}, b_{12}, \ldots, b_{1J})'$$
$$\eta = (\eta_1', \eta_2', \ldots, \eta_J'), \quad \eta_j = (\eta_{1j}', \eta_{2j}', \ldots, \eta_{n_j J}')'$$
$$\text{and} \quad \eta = X\beta + b_0 Z_0 + b_1 Z_1$$

where X, Z_0 and Z_1 are design matrices for β, b_0 and b_1. The Newton-Raphson iterative procedure for maximizing $l_1 + l_2$ is

$$\begin{bmatrix} \beta^{(s+1)} \\ b_0^{(s+1)} \\ b_1^{(s+1)} \end{bmatrix} = \begin{bmatrix} \beta^{(s)} \\ b_0^{(s)} \\ b_1^{(s)} \end{bmatrix} - V^{-1} \begin{bmatrix} 0 \\ \theta_0^{(s)-1} b_0^{(s)} \\ \theta_1^{(s)-1} b_1^{(s)} \end{bmatrix} + V^{-1} \begin{bmatrix} X' \\ Z_0' \\ Z_1' \end{bmatrix} \frac{dl_1}{d\eta}$$

$$V = \begin{bmatrix} V_{11} & V_{12} & V_{13} \\ V_{21} & V_{22} & V_{23} \\ V_{31} & V_{32} & V_{33} \end{bmatrix} = \begin{bmatrix} X' \\ Z_0' \\ Z_1' \end{bmatrix} \left(-\frac{d^2 l_1}{d\eta d\eta'} \right) [X Z_0 Z_1] + \begin{bmatrix} 0 & 0 & 0 \\ 0 & \theta_0^{(s)-1} I & 0 \\ 0 & 0 & \theta_1^{(s)-1} I \end{bmatrix}$$

$$V^{-1} = \begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{bmatrix}$$

where: I is a 1x1 identity matrix; V_{11} and A_{11} are p x p; V_{12}, V_{13}, A_{12} and A_{13} are p x 1; V_{21}, V_{31}, A_{21} and A_{31} are 1 x p; $V_{22}, V_{23}, V_{32}, V_{33}, A_{22}, A_{23}, A_{32}$ and A_{33} are 1 x 1 matrices.

The superscript (s) indicates the solution after s iterations. The restricted maximum likelihood (REML) estimator θ_u (u=0,1) given the estimated β, b_0, b_1 is given by

$$\theta_u^{(s+1)} = \frac{b_u^{(s+1)'} b_u^{(s+1)}}{J - \theta_u^{(s)^{-1}} trace(A_{u+2.u+2})}$$

with covariance matrix for the estimated β given by A_{11} and the asymptotic variance of estimated θ_u is $2\theta_u^2\left[J - 2\theta_u^{-1} trace\,(A_{u+2.u+2}) + \theta_u^{-2} trace\,(A_{u+2.u+2}^2)\right]^{-1}$. The estimation process for fitting the above model was programmed by [62] using the SAS IML procedure and the code was kindly made available by Professor Takuhiro Yamaguchi. Similar code is presented in Appendix A.3 for the analysis of five RCTs comparing carbamazepine and valproate for the treatment of epilepsy.

5.2.1. Software Development

As suitable software is currently unavailable for fitting the suite of Cox models with random effects that are of interest here (FE/RE (model 5.3), SFE/RE (model 5.4)) and to meta-analysis in general, the approach described above and SAS IML code is extended and adapted in this thesis to allow estimation of required parameters. For model FE/RE (5.3), the modifications required relate to the linear component η_{ij}, such that $\eta_{ij} = \sum_{k=1}^{p} \beta_k x_{kij} + b_{1j} x_{1ij}$, and the second log likelihood component becomes,

$$l_2 = -\frac{1}{2}\left(J \log 2\pi\theta_1 + \sum_{j=1}^{J} \frac{b_{1j}^2}{\theta_1}\right)$$

$$\begin{bmatrix} \beta^{(s+1)} \\ b_1^{(s+1)} \end{bmatrix} = \begin{bmatrix} \beta^{(s)} \\ b_1^{(s)} \end{bmatrix} - V^{-1} \begin{bmatrix} 0 \\ \theta_1^{(s)^{-1}} b_1^{(s)} \end{bmatrix} + V^{-1} \begin{bmatrix} X' \\ Z_1' \end{bmatrix} \frac{dl_1}{d\eta}$$

$$V = \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix} = \begin{bmatrix} X' \\ Z_1' \end{bmatrix} \left(-\frac{d^2 l_1}{d\eta d\eta'}\right) [X \; Z_1] + \begin{bmatrix} 0 & 0 \\ 0 & \theta_1^{(s)^{-1}} I \end{bmatrix}$$

$$V^{-1} = \begin{bmatrix} A_{11} A_{12} \\ A_{21} A_{22} \end{bmatrix}$$

where: I is a 1x1 identity matrix; V_{11} and A_{11} are p x p; V_{12} and A_{12} are p x 1; V_{21} and A_{21} are 1 x p; V_{22} and A_{22} are 1 x 1 matrices. The REML estimator θ_1 given the estimated β and b_1 is

$$\theta_1^{(s+1)} = \frac{b_1^{(s+1)'} b_1^{(s+1)}}{J - \theta_1^{(s)^{-1}} trace(A_{22})}$$

with covariance matrix for the estimated β given by A_{11} and the asymptotic variance of estimated θ_1 is $2\theta_1^2 \left[J - 2\theta_1^{-1} trace\,(A_{22}) + \theta_1^{-2} trace\,(A_{22}^2) \right]^{-1}$.

In the above formulation of the likelihood for the FE/RE model, the l_1 component is constructed over ordered event times for all participants in all trials. For the corresponding model stratified by trial, SFE/RE (5.4), only ordered event times within a trial are used to define the partial log likelihood for that particular trial therefore the parameter estimation procedure for model FE/RE (5.3) is further modified by redefining

$$l_1 = \sum_{j=1}^{J} \left\{ \sum_{i=1}^{n_j} \delta_i \left\{ \eta_i - \log \sum_{l \in \Re_i} \exp(\eta_l) \right\} \right\}$$

where, within the *jth* trial, $\eta_i = \sum_{k=1}^{p} \beta_k x_{ki}$ is the linear component for the *i'th* patient, δ_i is a censoring indicator taking the value unity if a failure occurs at time t_i and zero otherwise, and $\Re_i$ represents the risk set at time t_i and is constructed from the set of individuals still at risk of the event in the *jth* trial.

The adapted SAS IML code for fitting these random effects models is included in Appendix A (A.1 for FE/RE and A.2 for SFE/RE). As already mentioned, the original code supplied by Professor Yamaguchi was used to fit a random centre, random treatment (RE/RE) model in their 1999 paper. The example they used for illustration was a cancer clinical trial of 174 patients from 16 centres with the number of patients

within a centre ranging from 4 to 24. In the meta-analysis examples of interest, we have few 'centres' with much larger numbers of patients within each centre and consequently the estimation procedure can take a large amount of computing time.

5.3. Treatment coding

In trials comparing experimental and control treatments, the covariate representing experimental treatment group membership is traditionally coded as 1 whilst the control group takes the value 0. For trials comparing two active treatments, one treatment may be taken as a reference group and coded 0 whilst the other might take the value 1. In either case, if the treatment codes are switched, the value of the relative treatment effect remains unchanged in a model fitting both trial and treatment as fixed effects. However, for models containing random treatment effects, the interpretation of variance components will depend on the treatment coding values adopted and could potentially alter conclusions made regarding the overall treatment effect. The log hazard rate for the *ith* individual in the *jth* trial for each treatment group assuming two alternative treatment coding approaches within each trial are summarised below.

Group	FE/RE Model	SFE/RE Model	RE/RE Model
A $x_{1ij}=1$	$\log(\lambda_{ij}(t))=\log(\lambda_0(t))$ $+\beta_{0j}+\beta_1+b_{1j}$	$\log(\lambda_{ij}(t))=\log(\lambda_{0j}(t))$ $+\beta_1+b_{1j}$	$\log(\lambda_{ij}(t))=\log(\lambda_0(t))$ $+b_{0j}+\beta_1+b_{1j}$
B $x_{1ij}=0$	$\log(\lambda_{ij}(t))=\log(\lambda_0(t))$ $+\beta_{0j}$	$\log(\lambda_{ij}(t))=\log(\lambda_{0j}(t))$	$\log(\lambda_{ij}(t))=\log(\lambda_0(t))$ $+b_{0j}$
A $x_{1ij}=1/2$	$\log(\lambda_{ij}(t))=\log(\lambda_0(t))$ $+\beta_{0j}+1/2\beta_1+1/2b_{1j}$	$\log(\lambda_{ij}(t))=\log(\lambda_{0j}(t))$ $+1/2\beta_1+1/2b_{1j}$	$\log(\lambda_{ij}(t))=\log(\lambda_0(t))$ $+b_{0j}+1/2\beta_1+1/2b_{1j}$
B $x_{1ij}=-1/2$	$\log(\lambda_{ij}(t))=\log(\lambda_0(t))$ $+\beta_{0j}-1/2\beta_1-1/2b_{1j}$	$\log(\lambda_{ij}(t))=\log(\lambda_{0j}(t))$ $-1/2\beta_1-1/2b_{1j}$	$\log(\lambda_{ij}(t))=\log(\lambda_0(t))$ $+b_{0j}-1/2\beta_1-1/2b_{1j}$

For each model, under the assumption of normality for the b_{1j}'s, the variability in log hazard rate across trials for treatment A is greater than the variability in log hazard rate across trials for treatment B if treatment A is coded as $x_{1ij}=1$ and treatment B as

x_{1ij} =0. If treatment codes are switched, the variability across trials for treatment B is greater than the variability across trials for treatment A.

For each model, under the assumption of normality for the b_{1j}'s, the variability in log hazard rate across trials for treatment A is equal to the variability in log hazard rate across trials for treatment B if treatment A is coded as x_{1ij} =1/2 and B as x_{1ij} =-1/2. If treatment codes are switched, the variability across trials remains unchanged.

The decision of which treatment coding structure to adopt should ideally be made prior to analysis as the interpretation of the between trial variability parameter will differ accordingly. Provided the trials included are clinically and methodologically similar, for meta-analyses comparing an active treatment with placebo, the assumption of greater between trial variability in log hazard rate for the active treatment group may be appropriate. For meta-analyses of trials comparing two active treatments, the assumption of equal variability in log hazard rates for each group would seem more appropriate.

In some situations, depending on the approach used for parameter estimation, the treatment coding adopted could theoretically alter estimates and conclusions. For example, the between trial variability in log hazard rates for the treatment group coded as zero is not modelled in the 0/1 coding approach but is modelled using values of ±1/2. Parameter estimates could alter between these two approaches if the between trial variability in log hazard rates is non-negligible for this treatment group. A brief examination of meta-analysis of binary outcomes using the random treatment effect logistic regression model (fitted using MLWIN) revealed that the relative treatment estimates and their standard errors changed with different treatment coding structures. Similar treatment coding issues in the random trial, random treatment effect model have been discussed previously [59],[60].

The estimation approach for the fixed trial, random treatment effect Cox models (FE/RE, SFE/RE) adopted in this thesis is unaffected by the choice of treatment coding. This is because, the procedure consists of maximizing the sum of two components ($l = l_1 + l_2$), the first of which is the standard log partial likelihood of the

Cox model conditional on fixing the random effect (l_1), and the second component (l_2) is the log likelihood of a normally distributed variable. As the second component does not depend upon the value of x_{1ij}, the parameter estimates are identical regardless of treatment coding used.

5.4. Handling Ties

In the analysis of time-to-event outcomes, tied event times occur when the time to some event is identical for two or more individuals. The partial likelihood for the Cox model described in previous sections is developed under the assumption of continuous (i.e. untied) data [96]. For situations where there are a number of tied event times, the partial likelihood requires some adjustment. Two commonly used approaches for handling ties are the Breslow [100] and Efron [101] approximations.

The original SAS IML code supplied by Takuhiro Yamaguchi used the Breslow method for handling ties. As the Efron approximation generally produces results that are much closer to the exact partial likelihood in the presence of tied event times [29], the SAS IML code was adapted to facilitate the use of Efron's approximation for handling ties. The basic procedure is described below and the adapted SAS IML code for each random effects (FE/RE, SFE/RE, RE/RE) model using Efron's approximation is given in Appendix B (B.1, B.2, B.3).

For an individual trial with untied data, i.e. distinct event time for each individual, the partial likelihood function is a product over r ordered event times denoted by subscript $k = 1,2,...r$. Each individual who experiences an event at some time t_k, contributes a term

$$\frac{\exp(\eta_k)}{\sum_{l \in \Re_k} \exp(\eta_l)}$$

to the partial likelihood function. Letting $\psi(m) = \exp(\eta_m)$, the situation of tied event times is illustrated by Therneau and Grambsch [96] as follows. Suppose five individuals are included in a trial and the first two individuals have tied event times, assumed to have occurred because the time to an event is often not recorded precisely enough, e.g.

time to death recorded in days instead of days, hours and minutes. If the time data had been recorded more precisely, the first two terms would be either

$$\left(\frac{\psi(1)}{\psi(1)+\psi(2)+\psi(3)+\psi(4)+\psi(5)}\right)\left(\frac{\psi(2)}{\psi(2)+\psi(3)+\psi(4)+\psi(5)}\right)$$

or

$$\left(\frac{\psi(2)}{\psi(1)+\psi(2)+\psi(3)+\psi(4)+\psi(5)}\right)\left(\frac{\psi(1)}{\psi(2)+\psi(3)+\psi(4)+\psi(5)}\right)$$

but we do not know which.

The two approaches to handle tied event times differ in how they accommodate these alternative possibilities for defining the risk set. The approximation proposed by Breslow uses the complete sum for both denominators and the first two terms of the partial likelihood would therefore be

$$\left(\frac{\psi(1)}{\psi(1)+\psi(2)+\psi(3)+\psi(4)+\psi(5)}\right)\left(\frac{\psi(2)}{\psi(1)+\psi(2)+\psi(3)+\psi(4)+\psi(5)}\right)$$

The Breslow approximation is the simplest approach to program and the most efficient method when there are no ties. However, the approach produces a conservative bias with estimated regression coefficients that are too close to zero in absolute value [96] because individuals experiencing an event are included more than once in the denominator of the partial likelihood.

The approximation proposed by Efron uses the average denominator in the second term giving a likelihood contribution of

$$\left(\frac{\psi(1)}{\psi(1)+\psi(2)+\psi(3)+\psi(4)+\psi(5)}\right)\left(\frac{\psi(2)}{0.5\psi(1)+0.5\psi(2)+\psi(3)+\psi(4)+\psi(5)}\right)$$

which is to be preferred particularly as the number of tied event times increases.

5.5. Example: CBZ-VPS monotherapy trials using IPD

The effects of two anti-epileptic drugs, Carbamazepine (CBZ) and Sodium Valproate (VPS) when used as monotherapy in patients with partial onset seizures or generalized onset seizures were assessed in a systematic review and meta-analysis of randomised controlled trials [39]. As described in previous chapters, due to the lack of uniformity in reporting of outcome measures, the desire to investigate time to event outcomes and the need to examine treatment-covariate interactions, IPD were requested and obtained for 1225 patients from 5 trials accounting for 83% of individuals randomised in 8 eligible trials. Information for five clinically important patient characteristics (age at randomisation, sex, seizure type (generalized onset or partial onset), number of seizures prior to randomisation and time from first seizure to randomisation) are also available for each trial. The effect of three further covariates (EEG, CT/MRI scan, Neurological signs) cannot be investigated due to an unacceptable amount of missing data (Table 4.3, Chapter 4).

Illustration of the models described in section 5.1 is undertaken with the meta-analysis of two outcomes, time to 12 month remission and time to first seizure post-randomisation. The former outcome relates to the time taken to reach a period of 12 months free from seizures. Due to the fact that many individuals (29%) achieved an immediate seizure free period of 12 months (i.e. time to 12 month remission is 12 months), the occurrence of tied event times is a further issue for consideration for this outcome that is dealt with by using Efron's approximation for handling ties. The Kaplan-Meier survival curves summarising the time-to-event experience for each treatment group within each trial are displayed in Appendix C.

The original Log-rank analysis stratified by trial suggested statistically significant evidence for heterogeneity in the treatment effect between trials for the outcome time to 12 month remission (χ^2_4=11.75, p=0.02, Table 4.4) but evidence for statistical heterogeneity was not detected for time to first seizure (χ^2_4=5.89, p=0.21, Table 4.4). Parameter estimates and standard errors obtained from applying the models described in section 5.1 for both outcomes are summarised in Table 5.1. Since tied event times are not of concern for the outcome time to first seizure, the Breslow method for handling ties is used for all models. Similar parameter estimates are expected if the Efron method

is adopted. For the outcome time to 12 month remission, where event times are frequently tied, the Efron approximation is adopted throughout. For both outcomes, parameter estimates from FE/FE and SFE/FE models agree well with each other.

For time to 12 month remission, all five models agree in the direction of treatment effect and suggest a trend toward favouring CBZ. Compared with the appropriate fixed treatment effect model, the models with random treatment effects (FE/RE, SFE/RE, RE/RE) have a different parameter estimate of β_1 and larger standard error due to the allowance made for heterogeneity in these models. The parameter τ^2 is intuitively not very well estimated in this example due to the small number of trials and this is reflected by the large standard error for estimates of τ^2. The point estimates of this parameter are close to zero and do not appear to reflect the evidence found for significant heterogeneity obtained from the first two models (p=0.021 and p=0.023 for FE/FE and SFE/FE respectively). However, the increase in standard error for β_1 observed by introducing random treatment effects suggests an important amount of heterogeneity is present.

For time to first seizure, all five models agree in the direction of treatment effect. There is insufficient evidence for heterogeneity (p=0.18 and p=0.21 for FE/FE and SFE/FE respectively). On comparison with the results for time to 12 month remission, this is reflected in smaller estimates of τ^2 in the models that include random treatment effects (FE/RE, SFE/RE, RE/RE). For this reason, making allowance for heterogeneity has less of an impact and similar estimates of β_1 are obtained for all models with less of an increase in standard errors.

Table 5.1. Parameter estimates (standard error) for time to 12 month remission and time to first seizure obtained using alternative hierarchical formulations of the Cox regression model (5 trials, 1225 individuals)

Model	β_1 (VPS)	τ^2	σ^2
*Time to 12 month remission**			
FE/FE	-0.137 (0.073)		
SFE/FE	-0.132 (0.073)		
FE/RE	-0.103 (0.126)	0.0498 (0.056)	
SFE/RE	-0.098 (0.125)	0.0484 (0.055)	
RE/RE	-0.099 (0.126)	0.0510 (0.056)	0.0155 (0.016)
Time to first seizure$^{\varphi}$*			
FE/FE	0.087 (0.068)		
SFE/FE	0.079 (0.068)		
FE/RE	0.067 (0.089)	0.0146 (0.028)	
SFE/RE	0.061 (0.087)	0.0131 (0.027)	
RE/RE	0.067 (0.090)	0.015 (0.028)	0.019 (0.019)

FE/FE: Fixed trial indicators and fixed treatment effect, **SFE/FE**: stratified by trial with fixed treatment effect, **FE/RE**: Fixed trial indicators with random treatment effects, **SFE/RE**: Stratified by trial with random treatment effects, **RE/RE**: Random trial and random treatment effects. *Using Efron's Approximation for handling ties $^{\varphi}$ Using Breslow's Approximation for handling ties

As evidence for heterogeneity is identified for the outcome time to 12 month remission, potential explanations are explored by investigating the importance of clinically relevant patient characteristics. To do this covariates which are identified as statistically significant from univariate analyses are included in a model with treatment. All interaction with treatment terms are subsequently included and finally, any non-significant terms are omitted from the model. A model which includes a term describing type of epilepsy, age, log(number of seizures) and an age-treatment interaction is obtained for the outcome time to 12 month remission. Alternative model selection strategies (backward elimination, forward selection, stepwise selection) were also adopted and each produced the same final model. The parameter estimates and standard

errors obtained from including these terms in each of the models under investigation are displayed in Table 5.2.

Table 5.2. Parameter estimates (standard errors) for 'time to 12 month remission' using alternative hierarchical formulations of the Cox regression model (using Efron's approximation) allowing for patient-level covariates (5 trials, 1183 individuals)

Model	Estimates (SE)						
	β_1 (VPS)	τ^2	σ^2	Age	Epilepsy type (partial)	Log(sez)	Age *VPS
FE/FE	0.151 (0.129)	-	-	0.005 (0.003)	-0.189 (0.088)	-0.192 (0.031)	-0.009 (0.004)
SFE/FE	0.162 (0.129)	-	-	0.005 (0.003)	-0.186 (0.088)	-0.192 (0.031)	-0.009 (0.004)
FE/RE	0.150 (0.141)	0.008 (0.028)	-	0.005 (0.003)	-0.188 (0.088)	-0.192 (0.031)	-0.008 (0.004)
SFE/RE	0.163 (0.139)	0.006 (0.027)	-	0.005 (0.003)	-0.185 (0.088)	-0.192 (0.031)	-0.009 (0.004)
RE/RE	0.158 (0.140)	0.007 (0.028)	0.026 (0.026)	0.006 (0.003)	-0.167 (0.086)	-0.188 (0.031)	-0.009 (0.004)

The effect of age is included as the actual age (in years) of each individual in all models considered. For the models that assume fixed treatment effects (FE/FE and SFE/FE), parameter β_1 is interpreted as the log hazard ratio comparing VPS to CBZ for two individuals aged zero with the same epilepsy type and same number of seizures before randomisation. For the models that assume random treatment effects (FE/RE, SFE/RE and RE/RE), parameter β_1 is interpreted as the average log hazard ratio comparing VPS to CBZ for a population of possible log hazard ratios for two individuals aged zero with the same epilepsy type and number of seizures before randomisation. A more relevant clinical interpretation could be achieved by using centering. For example, the average age (across all trials) of 31 years could be subtracted and this centered covariate included instead of actual age. Parameter β_1 would then be

interpreted as the log hazard ratio comparing VPS to CBZ for two individuals aged 31 with the same epilepsy type and number of seizures before randomisation.

In all models the log hazard ratio is allowed to vary according to age due to the inclusion of an age by treatment interaction term. However, for the fixed treatment effect models (FE/FE and SFE/FE) the pattern according to age is assumed to be identical across trials whereas for the random treatment effects models, the pattern according to age is allowed to vary across trials with this additional level of variability incorporated into the model. In models that include an age by treatment interaction term, the random quantities b_{1j} represent deviations in the *jth* trial from the average log hazard ratio according to age whilst τ^2 is a measure of between trial variability in the treatment effect according to age.

Having allowed for patient-level covariates (covariate main effects modelling the variability in baseline hazards and the interaction with treatment effect modelling variation in treatment effects), the test for heterogeneity in models with fixed treatment effect FE/FE and SFE/FE, are no longer statistically significant (p=0.40 and p=0.44 respectively) and suggest that the included covariates explain the heterogeneity in treatment effect across trials. Furthermore, in support of these results, the estimate of τ^2 for FE/RE, SFE/RE, and RE/RE models has decreased after allowing for covariates. As the variables age at randomisation and number of seizures before randomisation were not recorded for 42 out of 1225 individuals, the parameter estimates displayed in Table 5.2 are based on a subset of 1183 individuals. To enable a comparison of parameters before and after inclusion of these covariates, each corresponding model excluding patient level covariates were fitted to the same subset of 1183 individuals (Table 5.3).

Table 5.3. Parameter estimates (standard error) for time to 12 month remission fitted to a subset of data (5 trials, 750 events, 1183 individuals)

Model	β_1 (VPS)	τ^2	σ^2
*Time to 12 month remission**			
FE/FE	-0.115(0.073)		
SFE/FE	-0.112(0.074)		
FE/RE	-0.085(0.121)	0.044(0.051)	
SFE/RE	-0.081(0.120)	0.043(0.051)	
RE/RE	-0.082(0.121)	0.044(0.052)	0.015(0.016)

*Using Efron's Approximation for handling ties

On comparison of parameter estimates from Table 5.3 and 5.2, the percentage change in τ^2 suggests that inclusion of age, epilepsy type, log(number of seizures), and an age by treatment interaction has explained 82% of the heterogeneity (FE/RE), 86% (SFE/RE) and 84% (RE/RE) for each model respectively. The estimate of σ^2 in the model with random trial and treatment effects (RE/RE) has increased following inclusion of patient-level covariates. From a clinical perspective, the age by treatment interaction suggests that older patients taking CBZ are more likely to experience a period of 12 month remission from seizures, hence a better clinical outcome, whilst younger patients fare better on VPS. In general, this change in direction of effect occurs at around the age of 18. There is evidence that individuals are unlikely to be diagnosed with generalized epilepsy beyond the age of between 25 and 30 years [102]. The interaction between age and treatment may be viewed as a surrogate for an interaction between type of epilepsy and treatment since there is strong clinical belief, which is unsupported by the data for this outcome, that CBZ is better for partial seizures whilst VPS is better for generalized [69]. Further detail regarding the clinical implication of these results are given by Marson *et al* [39] and Williamson *et al* [90].

5.6. Simulation study

In order to examine the reliability of the estimates from the different models for a meta-analysis with only 5 trials, and to gain insight into the behaviour of the models investigated, a small simulation study was undertaken. To allow different levels of

between trial variation, the data are simulated under a model with random trial and treatment effects (RE/RE) as described in section 3.6 of the third chapter. For each set of simulation parameters, the mean of the estimated log hazard ratios was calculated along with the coverage over all 100 simulations. Coverage is defined as the percentage of 95% confidence intervals that contain the true underlying value of log hazard ratio.

The simulation study results (Table 5.4) indicate that the average absolute bias and spread of the β_1 parameter increases for all models as the underlying degree of heterogeneity increases, but is similar across models for a given value of τ^2. The average absolute bias and spread of the estimate of β_1 are marginally smaller for the stratified models compared to corresponding un-stratified models (SFE/FE compared to FE/FE and model SFE/RE compared to FE/RE) for no treatment effect. For a fixed value of underlying heterogeneity ($\tau^2 = 0.1$), as the true value of β_1 increases, the average bias in the estimate of β_1 tends to decrease with little effect on coverage values for all models. For all models coverage for the log hazard ratio is 95% when there is no heterogeneity and the true underlying log hazard ratio is zero. As the degree of heterogeneity increases, coverage values for the two fixed treatment effect models (FE/FE and SFE/FE) decreases quite substantially. For the random treatment effect models (FE/RE, SFE/RE and RE/RE), coverage values are smaller when there is some degree of heterogeneity but values remain fairly stable for increasing heterogeneity. In conclusion, the estimate of the log hazard ratio is more likely to be biased for larger values of underlying heterogeneity with extremely poor coverage for fixed treatment effect models.

For a fixed underlying value of β_1, the average absolute bias in the estimate of τ^2 is minimal for all models (FE/RE, SFE/RE, RE/RE) with no systematic pattern for increasing values of true underlying heterogeneity. However, the increasing spread as the true value of τ^2 increases indicates that estimates of τ^2 are more likely to be biased when the true value is greater. Coverage for this parameter is poor when the underlying value is zero indicating poor variance estimation when $\tau^2 = 0$. For a fixed underlying value of τ^2 the bias in estimating this parameter and coverage values do not appear to be affected by increasing underlying values for β_1 regardless of model chosen.

Finally, for the model including random trial and random treatment effects (RE/RE), the bias in estimating σ^2 is unaffected by increasing underlying values of β_1 or τ^2. Coverage values are less than 75% for all scenarios.

The epilepsy example involves larger patient numbers than this simulation study. As the estimate of τ^2 in the example is fairly close to zero, particularly when covariates are included in the model, the simulation study provides additional reassurance that the parameter estimates obtained from the meta-analysis of the five epilepsy trials are reasonably reliable.

Table 5.4. Mean (standard deviation) and % coverage values of parameter estimates from 100 simulated meta-analyses of 5 trials (40 individuals per group)

Parameter	Underlying values	Model				
		FE/FE	SFE/FE	FE/RE	SFE/RE	RE/RE
β_1	0	-0.003 (0.117) 95	-0.001(0.115) 95	-0.003 (0.117) 95	-0.002 (0.115) 95	-0.004 (0.116) 95
τ^2	0			0.013 (0.028) 56	0.013 (0.029) 52	0.013 (0.028) 55
σ^2	0					0.006 (0.010) 73
β_1	0	-0.020 (0.189) 77	-0.017 (0.189) 76	-0.021 (0.190) 89	-0.019 (0.190) 88	-0.020 (0.189) 89
τ^2	0.1			0.096 (0.107) 80	0.094 (0.104) 80	0.095 (0.108) 80
σ^2	0					0.005(0.009) 73
β_1	0	-0.032 (0.336) 55	-0.029 (0.328) 55	-0.035 (0.346) 87	-0.033 (0.343) 86	-0.035 (0.344) 87
τ^2	0.5			0.503 (0.411) 79	0.492 (0.393) 79	0.496 (0.411) 79
σ^2	0					0.005 (0.009) 74
β_1	0	-0.041 (0.424) 47	-0.035 (0.410) 47	-0.045 (0.450) 87	-0.041 (0.444) 88	-0.047 (0.446) 87
τ^2	0.9			0.917 (0.708) 77	0.897 (0.677) 78	0.903 (0.704) 76
σ^2	0					0.005 (0.009) 74

Parameter	Underlying values	Model				
		FE/FE	SFE/FE	FE/RE	SFE/RE	RE/RE
β_1	0.1	0.082 (0.188) 77	0.083 (0.186) 77	0.082 (0.188) 88	0.082 (0.187) 88	0.082 (0.187) 86
τ^2	0.1			0.096 (0.109) 79	0.094 (0.105) 80	0.095 (0.110) 80
σ^2	0					0.005 (0.009) 69
β_1	0.5	0.490 (0.185) 75	0.487 (0.183) 76	0.491 (0.186) 89	0.490 (0.184) 89	0.489 (0.185) 88
τ^2	0.1			0.095 (0.110) 80	0.093 (0.106) 76	0.093 (0.111) 76
σ^2	0					0.005 (0.010) 70
β_1	0.9	0.895 (0.193) 77	0.888 (0.193) 79	0.900 (0.192) 89	0.894 (0.194) 90	0.894 (0.191) 89
τ^2	0.1			0.096 (0.109) 79	0.097 (0.111) 78	0.092 (0.108) 78
σ^2	0					0.006 (0.010) 68

5.7. Exploring heterogeneity with aggregate data

Investigating heterogeneity and the influence of prognostic factors in a regression framework is commonly undertaken using AD and the phrase *meta-regression* used to describe these models. Associations between the relative treatment effect and study-level characteristics are examined by fitting a regression model with study-level estimates of treatment effect as the response and summary measures of study-level characteristics e.g. average age, as explanatory variables. Consequently, although individual studies may have used randomization, the strength of this process is lost and these approaches are subject to many of the biases that occur with observational studies. Furthermore, if aggregated patient-level characteristics, such as mean age, are under investigation such approaches are subject to *ecological bias* which arises when results based on aggregate data are incorrectly assumed to apply at the individual level. Nevertheless, meta-regression models with AD can be useful for investigating the extent to which study-level characteristics might explain heterogeneity but there is a need to investigate how these results might compare to models based on IPD. A brief outline of the AD methods is given below but a discussion of assumptions and further details of the models are described in depth by Thompson and Sharp [91]. The methods are illustrated by Thompson and Sharp using binary data, but the general concepts can be extended to situations where continuous or time-to-event outcomes are of interest.

5.7.1. Weighted regression with no allowance for residual heterogeneity

The first model using AD assumes the observed log hazard ratio in each trial ($\log HR_j$) is independently normally distributed such that

$$\log HR_j \sim N(\alpha + \beta x_j, v_j) \qquad (5.6)$$

where v_j is the variance of the log hazard ratio in the *jth* trial. The parameter β is interpreted as the change in log hazard ratio per unit change in covariate x_j and α is the log hazard ratio for a covariate value of zero. To account for the assumption that the variance of the log hazard ratio from each trial are not equal (sampling variability), a weighted regression with weights defined by the reciprocal of the variance is used. This model is fitted using standard statistical software for weighted regression with the

standard errors of regression coefficients corrected through division with the square root of the mean squared error (MSE). This correction is necessary as standard statistical software packages usually fit the model $\log HR_j \sim N(\alpha + \beta x_j, v_j\sigma^2)$ rather than model (5.6) as required, where σ^2 is the mean squared error.

5.7.2. Weighted regression incorporating residual heterogeneity

The first meta-regression model (5.6) corresponds to a fixed effect model as between trial variability is not accounted for. This model can be extended to incorporate heterogeneity that remains unexplained by the covariates fitted through inclusion of an additive between study variance component τ^2. This random effects meta-regression model may be written as follows

$$\log HR_j \sim N(\alpha + \beta x_j, v_j + \tau^2) \qquad (5.7)$$

An explicit estimate of τ^2 is required as the weights used in the regression are given by the inverse of the sum of the within and between study variance components. The Moment Estimator (MM), Maximum Likelihood estimate (ML), Restricted Maximum Likelihood estimate (REML) and Empirical Bayes estimate (EB) methods of estimating τ^2 have been proposed and are described in detail by Thompson and Sharp [91].

5.8. Example: CBZ-VPS monotherapy trials using AD

For the CBZ-VPS time to 12 month remission meta-analysis examined in section 5.5 the original potential explanation of heterogeneity obtained using IPD (Table 5.2) motivated the question as to whether the same explanation of heterogeneity and clinical interpretations would have been obtained had the IPD been unavailable. However, for this particular example appropriate AD to allow such an assessment could not be extracted from trial reports and the IPD were therefore used to generate AD. The resulting AD for the outcome 'time to 12 month remission' and the 5 patient factors of interest, are summarized in Table 5.5.

Table 5.5. CBZ-VPS example: Aggregate data generated from IPD for each trial

Trial	Log hazard ratio[1] (SE)	Mean Age	Proportion Female	Proportion Partial epilepsy	Mean (log(number of seizures))	Mean (log(time from first ever seizure))
1. Heller 1995 [48]	-0.086 (0.223) N=122	30.65 N=119	0.52 N=122	0.40 N=122	1.27 N=119	0.81 N=119
2. De Silva 1996 [45]	0.223 (0.214) N=103	10.19 N=103	0.53 N=103	0.52 N=103	1.79 N=103	0.49 N=103
3. Richens 1994 [47]	-0.371 (0.134) N=288	33.32 N=286	0.50 N=288	0.49 N=288	1.67 N=288	N/A
4. Verity 1995 [46]	0.158 (0.148) N=246	10.09 N=244	0.53 N=246	0.42 N=246	1.56 N=245	0.08 N=226
5. Mattson 1992 [38]	-0.312 (0.145) N=466	47.21 N=466	0.08 N=466	1.00 N=466	3.01 N=431	1.46 N=450

[1] The log hazard ratio for VPS compared to CBZ calculated using individual Cox regression models with Efron method for handling ties for each trial (a positive log hazard ratio indicates a clinical advantage for VPS)

N/A=not available since covariate not measured in this trial.

Table 5.6. Parameter estimates (SE) from univariate meta-regression models using AD

	Covariate	α^* (SE)	β (SE)	Z, p-value	τ^2
Model (5.6)	Null	-0.132(0.073)	-	-	-
	Mean Age	0.290(0.158)	-0.015 (0.005)	-3.0, p=0.003	-
	Proportion female	-0.385(0.172)	0.621 (0.382)	1.63, p=0.104	-
	Proportion Partial	0.154(0.196)	-0.481 (0.306)	-1.57, p=0.116	-
	Mean (log(number of seizures))	0.192(0.237)	-0.166 (0.116)	-1.43, p=0.152	-
	Mean (log(time from first ever seizure))	0.237(0.139)	-0.365 (0.147)	-2.48, p=0.013	-
Model (5.7) MM	Null	-0.098(0.126)	-	-	0.05
	Mean Age	0.290(0.158)	-0.015 (0.005)	-3.01, p=0.003	0
	Proportion female	-0.392(0.329)	0.700 (0.717)	0.98, p=0.329	0.0563
	Proportion Partial	0.199(0.359)	-0.512 (0.577)	-0.89, p=0.376	0.0578
	Mean (log(number of seizures))	0.228(0.438)	-0.171 (0.220)	-0.78, p=0.437	0.0612
	Mean (log(time from first ever seizure))	0.237(0.139)	-0.365 (0.147)	-2.47, p=0.013	0
Model (5.7) ML	Null	-0.103(0.112)	-	-	0.034
	Mean Age	0.290(0.158)	-0.015 (0.005)	-3.01, p=0.003	0
	Proportion female	-0.389(0.243)	0.673 (0.533)	1.26, p=0.207	0.0210
	Proportion Partial	0.185(0.269)	-0.504 (0.428)	-1.18, p=0.239	0.0213
	Mean (log(number of seizures))	0.220(0.330)	-0.171 (0.164)	-1.04, p=0.297	0.0232
	Mean (log(time from first ever seizure))	0.237(0.139)	-0.365 (0.147)	-2.47, p=0.013	0

	Covariate	α^* (SE)	β (SE)	Z, p-value	τ^2
Model (5.7) REML	Null	-0.099(0.124)	-	-	0.048
	Mean Age	0.290(0.158)	-0.015 (0.005)	-3.01, p=0.003	0
	Proportion female	-0.391(0.310)	0.696 (0.676)	1.03, p=0.303	0.0475
	Proportion Partial	0.198(0.341)	-0.511 (0.547)	-0.93, p=0.351	0.0497
	Mean (log(number of seizures))	0.227(0.418)	-0.171 (0.209)	-0.82, p=0.415	0.0534
	Mean (log(time from first ever seizure))	0.237(0.139)	-0.365 (0.147)	-2.47, p=0.013	0
Model (5.7) EB	Null	-0.099(0.122)	-	-	0.045
	Mean Age	0.290(0.158)	-0.015 (0.005)	-3.01, p=0.003	0
	Proportion female	-0.391(0.299)	0.693 (0.652)	1.06, p=0.288	0.0427
	Proportion Partial	0.197(0.334)	-0.510 (0.536)	-0.95, p=0.341	0.0466
	Mean (log(number of seizures))	0.227(0.411)	-0.171 (0.206)	-0.83, p=0.407	0.0508
	Mean (log(time from first ever seizure))	0.237(0.139)	-0.365 (0.147)	-2.47, p=0.013	0

MM: Moment estimator, ML: Maximum Likelihood, REML: Restricted Maximum Likelihood, EB: Empirical Bayes

* Comparison is VPS to CBZ therefore VPS is better if HR>1

As only 5 trials are available multivariate models using AD are not investigated. Furthermore, with so few trials, there is an increased probability of finding a statistically significant result as further covariates are explored and analyses should therefore be interpreted with caution. Model (5.6) was fitted using the regress command and (5.7) using the metareg command of STATA version 8.2. The parameter estimates and corresponding standard errors obtained from fitting each univariate model are summarised in Table 5.6.

As residual heterogeneity is not accounted for in model (5.6) the SEs of each regression coefficient are smaller compared to model (5.7). There is generally reasonable agreement between the estimated regression coefficients from alternative models using AD.

All models provide strong evidence for a significant effect of mean age and mean(log(time from first ever seizure)) with identical regression coefficients and standard error across all models. The results suggest that VPS may be more effective in trials with younger patients on average, whilst CBZ may be more effective in trials with an older average age with the change in direction of effect occurring at around mean age of 19 years. VPS appears more effective in trials recruiting patients with a shorter interval between first seizure and randomisation on average whilst CBZ more effective in trials with larger intervals on average. For model (5.7), the between trial variability parameter τ^2 is estimated to be zero for all estimation procedures when either of these variables are included suggesting that all of the heterogeneity may be explained by considering these trial-level averages of patient level covariates. This is further supported by noting that the parameter estimates and standard error for random effects models (5.7) are precisely equal to those of the fixed effect model (5.6) after including these covariates.

Evidence to suggest a relationship between treatment effect and any other aggregated covariate examined in this investigation is much weaker, with non-zero estimates of τ^2 indicating that some residual variability remains unexplained by the effect of each aggregate level covariate. For model (5.7) comparing the change in τ^2 with corresponding null models as a measure of the proportion of variation explained gives quite different values depending on the estimation approach. However, as estimation of τ^2 is poor when the number of included trials is small, as in this case, the reliability of

these results is questionable and care is required for interpretation. In this example with 5 trials, estimates of τ^2 for all covariates are smaller using ML compared with MM, REML and EB approaches. Thompson *et al.* [91] suggest that the maximum likelihood approaches are preferable but due to the downward bias of the estimate of τ^2 using ML, they suggest that an REML estimate will be most appropriate in practice.

5.9. Exploring heterogeneity with IPD or AD

Parameter estimates obtained from models using IPD (section 5.5) or AD (section 5.8) may be compared to provide an empirical evaluation of meta-regression analyses using both types of data. In the CBZ-VPS example, AD were generated from IPD and the comparison is therefore slightly artificial and reflects a comparison of methods rather than results that might be seen in reality.

Since meta-regression models with AD are used to explore relationships between trial level covariates and trial level relative treatment effects, comparisons can only be made with IPD based models in terms of treatment by covariate interactions rather than covariate main effects. The ability to examine and adjust for covariate main effects is one advantage of having IPD. To compare models using IPD and AD the following approach is taken. For each IPD model with fixed treatment effect (FE/FE, SFE/FE) and the stratified Cox model with random treatment effects (SFE/RE), the main effects of treatment and the covariate of interest are included in a model with the corresponding treatment by covariate interaction variable (results of these models are displayed in Table 5.7).

Results

The first point to note is the AD null model result assuming fixed treatment effect (model 5.6, table 5.6) gives exactly the same parameter estimate and standard error (-0.132(0.073)) as the SFE/FE Cox model using IPD (Table 5.7). This agreement occurs because the AD estimates of log hazard ratio and SE (Table 5.5) have been generated from IPD using a separate Cox model for each trial. The fixed effect AD meta-regression model (5.6) without covariates is equivalent to a simple Inverse Variance weighted average of trial level estimates. As stated in Chapter 3, assuming a fixed

treatment effect, the IV weighted average of Cox model estimates can give very similar pooled results to the those of the stratified Cox model under certain conditions. For conditions similar to the current example, the simulation study with underlying values of τ^2=0 or 0.1 and log hazard ratio close to 0.1 (Figures 3.1, 3.2, 3.5, 3.6 in Chapter 3) indicate particularly good agreement between these two approaches.

For models that assume random effects without considering the effect of covariates, parameter estimates and standard errors for meta-regression analyses using AD (model 5.7, Table 5.6) are similar to the three random effects models based on IPD (Table 5.1) which use REML for estimating τ^2. The AD model results based on a ML approach for estimating τ^2 are most similar to IPD estimates obtained from the FE/RE Cox model although the AD standard error and estimate of τ^2 are smaller than corresponding IPD estimates. The three remaining AD approaches (MM, REML, EB) agree well with IPD estimates from SFE/RE and RE/RE Cox model results.

The only treatment by covariate interaction identified as statistically significant by the Cox models using IPD appears to be between treatment and age (Table 5.7). The results suggest that individuals over the age of approximately 18 years have a better clinical outcome with CBZ. Very similar numerical results and clinical conclusions are drawn from the AD models (Table 5.6) after considering this particular covariate with both approaches estimating τ^2 to be equal to zero. The AD models measure the relationship between log hazard ratio and mean age across trials whereas the IPD models measure within trial relationships (averaged across trials) between treatment and age and are therefore measuring different quantities.

Based on the IPD Cox model results, there is insufficient evidence to suggest that any further interactions exist between treatment and each of the covariates under consideration (Table 5.7). Estimates of τ^2 that are close to that of the model without any covariate effects (null model in table 5.7) suggest that inclusion of these variables do not provide a sufficiently adequate explanation for heterogeneity. The AD model results agree to some extent in terms of the statistical significance for three covariates proportion female, proportional partial, and mean(log(number of seizures)). However, for mean(log(time from first ever seizure)), the AD models suggest evidence of a relationship (p=0.013) with an estimate for τ^2 equal to zero which might suggest that

this aggregate level variable can provide an explanation for statistical heterogeneity. As the safer IPD approaches failed to detect an overall within study relationship, the AD result is likely to be spurious. This highlights the potential for misinterpretation that can arise when associations with averages across trials are examined.

Assuming the AD generated from IPD had been available for each trial, one could have reached the conclusion that statistical heterogeneity could be explained by either the effect of age or time from first ever seizure. These two aggregate variables are highly correlated with longer average intervals from first ever seizure observed in trials with a greater average age. As data are available for a maximum of five trials, models including more than one covariate were not examined. Furthermore, as the false-positive rate increases as more characteristics are explored, these models that examine 5 characteristics with only 5 trials should be interpreted very cautiously. The availability of IPD allowed a thorough investigation into the main effects of each covariate (Table 5.2) which was not possible using meta-regression of AD. However, for the full SFE/RE model described in section 5.5 (Table 5.2, based on 750 events and 1183 individuals due to missing covariate values) there is a small amount of residual heterogeneity (τ^2=0.006(0.027)) with 86% of the heterogeneity explained by including the main effects of age, epilepsy type, log(number of seizures) and an interaction between treatment and age. Let this model be referred to as model (1). The IPD model (SFE/RE) that includes the main effect of age and an interaction with treatment term (Table 5.7, based on 764 events and 1218 individuals) suggests that 100% of the heterogeneity can be explained by these variables alone. Due to a small amount of missing covariate data these two alternative models are based on different subsets of the original data for 1225 individuals which makes a comparison of models difficult. As a sensitivity analysis, the variables treatment, age and their interaction term were fitted to the same subset of IPD for 1183 individuals used in model (1). The results based on this subset of data (Table 5.8) are not substantially different to the original (Table 5.7) but do suggest that a small amount of residual heterogeneity remains unexplained (τ^2=0.003(0.024)) by these variables using this data. In summary, the age by treatment interaction appears to explain the heterogeneity across trials but the variables epilepsy type and log(number of seizures) are also clinically important.

Table 5.7. Parameter estimates (SE) from Cox proportional hazards models with main effect of treatment, covariate and corresponding interaction term using IPD

Model	Events/ Total	Covariate	Treat (VPS) (SE)	Covariate (SE)	Treat*covariate (SE)	change, df, p-value	τ^2(SE)
FE/FE	767/1225	Null	-0.137(0.073)	-	-		
	764/1218	Age	0.210(0.129)	0.004(0.003)	-0.012(0.004)	9.74, 1, p=0.002	
	767/1225	Sex (Female)	-0.088(0.095)	-0.025(0.080)	-0.115(0.148)	0.61, 1, p=0.44	
	767/1225	Epilepsy Type (Partial)	-0.055(0.113)	-0.354(0.084)	-0.139(0.147)	0.89, 1, p=0.34	
	752/1186	log(number of seizures))	-0.058(0.121)	-0.214(0.030)	-0.026(0.057)	0.20, 1, p=0.65	
	525/898	log(time from first ever seizure)	0.029(0.107)	-0.055(0.050)	-0.077(0.086)	0.80, 1, p=0.37	
SFE/FE	767/1225	Null	-0.132(0.073)	-	-		
	764/1218	Age	0.210(0.129)	0.004(0.003)	-0.012(0.004)	9.99, 1, p=0.002	
	767/1225	Sex (Female)	-0.086(0.095)	-0.016(0.081)	-0.110(0.148)	0.55, 1, p=0.46	
	767/1225	Epilepsy Type (Partial)	-0.043(0.113)	-0.351(0.085)	-0.150(0.147)	1.04, 1, p=0.31	
	752/1186	log(number of seizures))	-0.048(0.122)	-0.214(0.030)	-0.028(0.057)	0.25, 1, p=0.62	
	525/898	log(time from first ever seizure)	0.037(0.107)	-0.057(0.050)	-0.081(0.086)	0.89, 1, p=0.35	
SFE/RE	767/1225	Null	-0.098(0.125)	-	-		0.0484(0.055)
	764/1218	Age	0.210(0.129)	0.004(0.003)	-0.012(0.004)		0
	767/1225	Sex (Female)	-0.006(0.152)	-0.030(0.081)	-0.206(0.158)		0.064(0.067)
	767/1225	Epilepsy Type (Partial)	-0.049(0.151)	-0.347(0.085)	-0.088(0.160)		0.043(0.052)
	752/1186	log(number of seizures))	-0.035(0.147)	-0.213(0.030)	-0.019(0.058)		0.031(0.043)
	525/898	log(time from first ever seizure)	0.005(0.150)	-0.057(0.051)	-0.013(0.095)		0.040(0.063)

The issue of missing data are not confined to meta-analysis and are often faced in individual trials if trying to develop the prognostic models where covariate values may be missing for some individuals. Issues of missing data are not considered in detail here but could present potential problems for investigating heterogeneity with IPD. The degree of missing data for the five covariates of interest is not considered a significant problem for this example.

For this particular empirical comparison involving a small number of trials, but still reflective of many meta-analyses in practice, the results suggest that conclusions from a meta-regression using AD can agree with results from IPD models if there is evidence for a within study treatment by covariate interaction and sufficient between trial variation for the aggregate value of the covariate. Departures from this condition could mean that meta-regression results using AD are unreliable. Furthermore, failure to find an effect in an AD based analysis is not evidence of a lack of effect [89]. Berlin *et al* [103] have undertaken similar comparisons of meta-regression analyses based on IPD or AD with an empirical example of 5 trials. Their investigations revealed that the AD meta-regression analyses failed to detect the importance of a particular covariate, panel reactive antibodies (PRA), included as the percentage above or below a particular cut-off value. In contrast, the IPD based models revealed a clinically important and statistically significant difference between the effect of treatment among patients whose PRA value was above or below the chosen cut-off. These results show another means by which AD and IPD based meta-regression analyses could potentially differ. The authors recommend that IPD should be used whenever feasible [103].

A meta-regression using AD from trial reports was not possible as suitable data were not available therefore undertaking a meta-analysis and investigation of potential sources of heterogeneity would simply not have been feasible. In this example, the IPD have proved to be extremely valuable.

Table 5.8. Sensitivity analysis: Parameter estimates (SE) from Cox proportional hazards models with main effect of treatment, covariate and corresponding interaction term using IPD using subset of data for 1183 individuals

Model	Events/Total	Covariate	Treat (VPS) (SE)	Covariate (SE)	Treat*covariate (SE)	change, df, p-value	τ^2 (SE)
FE/FE	750/1183	Null	-0.115(0.073)	-	-		
	750/1183	Age	0.195(0.130)	0.004(0.003)	-0.011(0.004)	8.37, 1, p=0.004	
SFE/FE	750/1183	Null	-0.112(0.074)	-	-		
	750/1183	Age	0.206(0.130)	0.004(0.003)	-0.011(0.004)	8.79, 1, p=0.003	
SFE/RE	750/1183	Null	-0.081(0.120)	-	-		0.0428(0.051)
	750/1183	Age	0.204(0.134)	0.004(0.003)	-0.011(0.004)		0.003(0.024)

5.10. Discussion

In any meta-analysis it is important to evaluate heterogeneity. The availability of IPD allows patient level covariates to be evaluated as potential causes of heterogeneity (treatment effect modifiers) using a regression framework with either fixed or random effects. Current literature in the area of meta-analysis with IPD have not addressed the analysis of time-to-event IPD using random effects Cox regression models. In this chapter a number of hierarchical formulations of the Cox model potentially suitable for undertaking meta-analysis of individual patient failure time data have been described and developed. In particular, random effects Cox regression models with suitable SAS programs for fitting these models have been developed and explored using empirical and simulated data.

A semi-parametric Cox regression model, which does not require making any parametric assumptions regarding the distribution for the survival times or baseline hazard function, has been assumed throughout. This attractive feature of the semi-parametric Cox model makes it a flexible approach that is commonly undertaken for the analysis of failure time data. However, if the assumption of a particular probability distribution for the data is valid, a more powerful analysis may be achieved by considering parametric models. Furthermore, parametric models may offer particular computational advantages for fitting models that include random effects and have made a noticeable contribution to the genetic epidemiology literature. See for example Scurrah *et al* [128] for a discussion of Generalized Linear Mixed Models, or Zahl and Harris [129] for an application of shared frailty models for the analysis of cancer incidence rates in twins. The use of different parametric models will be explored as a future research project.

Trial effects can be allowed for either by the inclusion of fixed effects using indicator variables, by stratification, or through the inclusion of random effects. For the fixed trial effect model the likelihood is constructed using ordered event times from all trials, whereas the likelihood from the stratified model is a summation of likelihood terms from each individual trial. The latter model is therefore more appropriate for meta-analysis as the within trial structure is maintained in the likelihood construction. However, if many trials are included in the meta-analysis, unstable estimates may be produced using fixed trial effect or stratified models. The model with random trial effects assumes that the trials in the meta-analysis are a random sample of trials from a

larger population of trials. This may not be reasonable since an assumption that underpins meta-analysis undertaken within a systematic review is that all relevant trials are identified and included in the analysis. Although this has not been investigated in detail here, O'Quigley and Stare [104] have recently concluded using simulations that the random effects model (random trial effects) only provides modest efficiency gains for group sizes (i.e. number of individuals per trial in the meta-analysis context) of five or more compared to the stratified model. For moderate to large numbers of very small groups, of sizes two or three, they conclude that the efficiency gains of the random effects model is far from negligible and there is a strong case for using this model rather than a stratified model. In the epilepsy example, with at least 100 individuals in each of five trials, the current author would consider the stratified models (SFE/FE or SFE/RE) to be the most appropriate. In particular, the stratified Cox model with random treatment effects (SFE/RE) is the least computer intensive of the random treatment effect models making this an attractive approach.

The EM algorithm can be used to estimate parameters in the Cox models that include random effects. However, the partial penalized likelihood approach has been the focus in this chapter as it has been noted that the EM algorithm is slow and variance estimates require further computation [97]. Alternative software packages (e.g. MLWIN and STATA) may provide the ability to fit the unstratified random effects models examined. However, earlier investigations using these software packages indicated that the computing time required was likely to be problematic for the analysis of the epilepsy data set. Attention has therefore focused on extending the estimation approach and SAS IML code originally described by Yamaguchi & Ohashi [62] to fit a random trial, random treatment effect model. Further modifications were also made to allow use of the Efron approximation which proved to be important in the example from epilepsy where tied event times occurred for the remission outcome. The resulting collection of programs for fitting the random effects models presented can potentially be extended further to incorporate additional random effects and non-zero random effect covariance structures. Unfortunately, the computing time required to fit the random effect models is likely to be restrictive in many situations and will depend on the number of trials, patients and events in the meta-analysis. For the two epilepsy examples, the analysis of time to first seizure outcome with 864 (71%) events took considerably longer than the analysis of time to 12 month remission with 767 (63%) events.

Choosing suitable values to represent the treatment covariate is a further important consideration required for the random treatment effect models to ensure that variability across trials in the hazard rate of both treatment groups is incorporated if appropriate. Although the estimation approach adopted here was unaffected by choice of treatment coding, the issue is likely to be a problem for other estimation procedures or other outcome types and further work is required to investigate the implications and extent of this problem.

Although computationally intensive, the simulation study provided an insight into the behaviour of the models investigated. Further work could be undertaken to examine other parameter values, such as non-zero values of the underlying between trial variability parameter σ^2, and alternative factors such as censoring pattern and proportionality of hazards that may influence the behaviour of these models.

Formal statistical tests of the evidence for heterogeneity in treatment effect across trials were examined in models which include fixed trial and treatment effects only. Although formal statistical tests are available for random effects failure time models, the performance of such tests was not the primary focus here and they were therefore not examined. Further details are given by Walker and Babiker [105], Gray [106] and Therneau and Grambsch [96].

The example from epilepsy provided the original motivation to investigate and apply alternative models to undertake a meta-analysis and explore heterogeneity using individual patient failure time data. In some situations where trials agree in outcome definition and the reporting of suitable data, aggregate approaches are likely to be less resource intensive but potentially more restricted. A pragmatic comparison of results using IPD versus results using extracted AD was not possible for this example as sufficient data were unavailable directly from trial reports. Such a limitation commonly arises in meta-analysis and often prevents any reasonable investigation into sources of heterogeneity. As the AD used for comparison were constructed from the IPD, the AD results represent the "best possible" results obtainable using this data type. One advantage of having IPD is the ability to examine main effects of covariates. For the epilepsy example, the clinical interpretation obtained from the final Cox regression

models would not have been discovered without IPD. The current author would recommend that for investigating heterogeneity, the model selection strategy should involve examination of all pre-specified clinically important treatment by covariate interactions, rather than exploring interactions only if the corresponding main effect is found to be significant. However, although unlikely in many meta-analysis situations, independent validation of interaction effects is ideally required therefore inferences should be made cautiously.

Further comparisons between IPD and AD and a systematic assessment of the empirical evidence are needed in order to provide guidelines of how, and in which situations, IPD is most beneficial for meta-analysis and meta-regression. A systematic review of empirical comparisons for the main treatment effect [57] and an international collaborative effort to perform empirical comparisons of meta-regressions (Jesse Berlin, personal communication, ESTEEM project) are currently being planned. The comparison and discussion presented in this paper can be added to this growing body of empirical evidence evaluating the benefits of using IPD or AD. The current author agrees with the recommendations of Berlin *et al* [103] and Lambert *et al* [89] that IPD should be used whenever possible to reliably study patient characteristics and investigate heterogeneity. This recommendation is especially important when the number of trials in the meta-analysis is small and AD approaches are likely to become increasingly more uncertain. Furthermore, if time-to-event outcomes are of interest, IPD can be extremely valuable due to limitations reporting appropriate summary data.

CHAPTER 6

External evidence and indirect comparisons

6.1. Introduction

Meta-analyses using small numbers of trials that directly compare two treatments of interest are common and estimates of the relative efficacy parameter and heterogeneity parameter are often imprecise in such situations. Incorporating external evidence, defined here as relevant evidence relating to a particular comparison of treatments that can be obtained from outside the usual source of evidence (i.e. randomised controlled trials that provide a direct comparison of treatments), could potentially improve the precision in estimating the parameters of interest. In some cases there may be no evidence available from randomised controlled trials that directly compare the treatments of interest. For example, trials may have been undertaken to compare a number of active drugs with placebo but no trials comparing the active drugs with each other. The examination of external evidence may be the only viable option to gain some knowledge of how the active treatments of interest compare.

Although there are many types of external evidence, this thesis will only consider external evidence from randomised controlled trials that indirectly compare the treatments of interest. As an example, suppose that the treatment effect for a

comparison of two drugs A versus B is of primary interest. An estimate of this treatment effect can be obtained indirectly by using external evidence from randomised trials that compare drug A versus C, and trials that compare drug B versus C. Such an indirect comparison can be valuable in situations where direct comparisons either do not exist (i.e. no trials directly compare A versus B), comprise a limited amount of data, or are unlikely to ever be examined in future trials.

Several drugs or therapies are often available in clinical practice to treat a particular medical condition. Although a meta-analysis within a systematic review provides a summary of the available evidence, the results conventionally relate to a direct comparison of only two drugs, commonly referred to as a pair-wise or head-to-head comparison. To accommodate an overall summary of the evidence concerning multiple drugs, several meta-analyses of pair-wise comparisons may be examined alongside each other within the same systematic review. For example, Marson *et al* [107] conducted a systematic review to examine a series of comparisons of active drugs used as add-on therapy versus placebo. Six individual pair-wise comparisons were analysed (Gabapentin (GBP) versus placebo, Lamotrigine (LTG) versus placebo, Tigabatrin (TIG) versus placebo, Topiramate (TPM) versus placebo, Vigabatrin (VGB) versus placebo and Zonisimide (ZNS) versus placebo) with an estimate of treatment effect and 95% confidence interval presented alongside each other. Although 'formal' analyses comparing individual active drugs e.g. GBP versus LMT, were not undertaken, the overlapping CIs within each comparison led to the conclusion that there was insufficient evidence to support a difference between any of the active drugs. Informal indirect comparisons are often made by authors or researchers interpreting the results for themselves. However, as the indirect comparisons between active drugs are not themselves based on randomised evidence, there may be substantial diversity across trials and between patient populations that could contribute to misleading conclusions if inappropriate methods are used. There is a need to recognise such limitations and either present formal analyses of indirect evidence or indicate reasons against making such comparisons.

The indirect comparison of treatments using external evidence can be formally examined and, if appropriate, incorporated into the meta-analysis using both frequentist and bayesian frameworks. Relevant methods based on both aggregate and individual

patient data will be reviewed and described in following sections with application of some frequentist methods to examples from epilepsy. The novel IPD methods for estimating indirect comparisons with time-to-event outcomes presented in this chapter are further compared with aggregate data based results where appropriate. Examples taken from the systematic reviews of epilepsy data described in Chapter 4 are used to illustrate relevant methods. In order to highlight different aspects of indirect comparisons and how they may be valuable, some examples used for illustration have been altered slightly and may not reflect true clinical results.

The eight reviews of monotherapy trials in epilepsy each consider a direct pair-wise comparison of two AEDs of interest. The total of six AEDs examined across eighteen RCTs in these reviews provide a network of interrelated direct and indirect evidence. Furthermore, the availability of IPD for each of these trials presents a unique opportunity to fully explore this entire body of evidence which is referred to in this thesis as the *totality of evidence*. The models for exploring indirect comparisons with individual patient failure time data proposed in this chapter are further extended to allow the totality of evidence analysis to be undertaken. To the author's knowledge, models for analyses involving indirect evidence, IPD and time-to-event outcomes have not been proposed elsewhere. Furthermore, the 'new' clinical results presented will be valuable to the epilepsy clinician in practice.

6.2. Overview of methods for indirect comparisons

A project commissioned by the HTA has been undertaken by Glenny, Altman *et al* (personal communication). The project aims, outlined in the report, were to survey the frequency of use of indirect comparisons in systematic reviews and evaluate the methods used in their analysis and interpretation, identify alternative statistical approaches for the analysis of indirect comparisons, assess the properties of different statistical methods used for performing indirect comparisons, carry out empirical work comparing direct and indirect estimates of the same effects within reviews. They identified 349 meta-analyses from electronic searching of the Database of Abstracts of Reviews of Effectiveness (DARE), a database containing abstracts of quality assessed systematic reviews. Thirty-six (10%) of these reviews included indirect comparisons, thirteen of which also included a direct comparison of the interventions of interest. The

method used for indirectly comparing interventions was classified by the report authors as 'naïve' or 'adjusted'. They define the naïve approach as pooling data across treatment arms, thereby ignoring the fact that the studies are RCTs. The adjusted indirect comparison is a general term they use to describe a number of approaches that adjust the comparison of the interventions of primary interest by the results of their direct comparison with a common intervention (or control group). This approach is preferred to the naïve approach as the advantages of randomisation within a trial are preserved. Of the reviews they identified that examine indirect comparisons, 25 (69%) used an adjusted approach and 11 (31%) used the naïve approach to make such comparisons. Of the 13 reviews that examined both direct and indirect comparisons, the two approaches gave similar results in terms of the direction of effect (but not necessarily the magnitude) in 8 meta-analyses but different results were obtained in 3 meta-analyses, and uncertainty regarding agreement in the last 2 reviews.

Within the HTA report, a separate systematic review of indirect comparison methodology identified ten publications [108], [109], [110], [111], [112], [113], [114], [115], [116], [117] describing some aspect of methodology for undertaking indirect comparisons with data from randomised controlled trials and a further 3 publications [118], [119], [120] addressing similar issues for uncontrolled studies.

The general method proposed by Bucher *et al* [108] for the indirect comparison of binary outcome data was employed by Fisher *et al* [114] to estimate the effect of a new drug compared to placebo when placebo controlled trials were considered unethical. The same approach was applied by Packer *et al* [117] whilst Hasselblad and Kong [116] describe the application of the approach proposed by Bucher *et al* [108] to other effect measures (risk differences, relative risk and hazard ratios). Hirotsu and Yamada [113] discuss a method for estimating odds ratios from direct and indirect evidence, a method which Glenny *et al.* (personal communication) note is equivalent to the Bucher approach using inverse variance weighting. The Berkey *et al* [109] publication describes a generalised least squares model for the joint meta-analysis of more than one outcome such that not all trials included need to have reported each outcome. The method also facilitates inclusion of multiple treatments and does not require that all trials assess all treatments. Hasselblad [111] describes the use of a logistic regression model for binary outcomes. Gleser and Olkin [115] describe a fixed effect regression model approach for

the simultaneous analysis of trials comparing one or more treatments with a control. They describe how relative treatment effects (risk difference and odds ratio) and their confidence intervals can be calculated by fitting each treatment effect as the response in a weighted least squares analysis with appropriate allowance made for covariance if the control group within a trial is used more than once. The final two publications identified [112], [110] discussed the analysis of indirect evidence with or without direct comparison within a Bayesian framework.

The HTA report concludes that only two basic valid approaches can be applied using standard software, the adjusted indirect comparison proposed by Bucher *et al* [108] and multiple (logistic) regression. They note that the regression approach has the ability to make adjustments for other variables that might help explain some of the heterogeneity within and between groups of trials making the same comparisons. They also note that for the adjusted indirect method, the random effects analysis is a safer option to allow for the potential for heterogeneity in at least one of the sets of trials used in the indirect comparison.

The existing methodology for indirect comparisons has mostly focused on estimation using aggregate binary data. One publication [116] describes the Bucher *et al* [108] approach in relation to estimates of the log hazard ratio from Cox proportional regression models (aggregate data) but results from trials involving more than two treatments are not considered. A separate publication by the same author [111] describes a regression approach (logistic regression) for undertaking indirect comparisons with individual patient binary data. A small section of a publication by Higgins *et al* [60] extends a random treatment effects regression model with IPD for the analysis of continuous outcome data from trials with 3 treatment arms. However, to the author's knowledge, there are no publications that address methods for undertaking indirect comparisons with IPD when time-to-event outcomes are of interest. In the current Chapter, existing approaches for aggregate data are described and a Cox regression model is adopted to allow appropriate analysis with IPD. The model is used to estimate the treatment effect for indirect comparisons, combine direct and indirect evidence and accommodate trials with more than two treatment groups. Where possible, corresponding analyses using aggregate data are presented for comparison.

6.3. Indirect aggregate data comparisons with time-to-event data

The approach described by Bucher *et al* [108] for indirect comparisons using aggregate binary data preserves randomisation within a trial by utilising the within-trial effect measure (log odds ratio in the 1997 publication) and its variance. If the comparison of treatment A versus B is of primary interest, an estimate of the indirect log odds ratio ($\log OR_{A:B\ indirect}$) and its variance can be obtained using external evidence from two separate trials, one that directly compares treatment A versus C ($\log OR_{A:C}$) and another that directly compares treatment B versus C ($\log OR_{B:C}$). This can be achieved using the following expressions,

$$\log OR_{A:B\ indirect} = \log OR_{A:C} - \log OR_{B:C} \qquad \textbf{(6.1)}$$

$$\text{var}(\log OR_{A:B\ indirect}) = \text{var}(\log OR_{A:C}) + \text{var}(\log OR_{B:C}) \qquad \textbf{(6.2)}$$

The indirect estimate is unbiased in large samples if there is no interaction between covariates defining subgroups of patients and the magnitude of the treatment effect [108]. The variance of the indirect estimate (A versus B) is equal to the sum of variances (6.2) since the two odds ratios (A versus C and B versus C) are estimated from separate studies and are thus statistically independent [108].

If multiple trials are available comparing A versus C and B versus C, standard meta-analyses can be undertaken for each comparison separately with $\log OR_{A:C}$, $\log OR_{B:C}$, $\text{var}(\log OR_{A:C})$, and $\text{var}(\log OR_{B:C})$ in (6.1) and (6.2) replaced by corresponding pooled estimates obtained from each meta-analysis. This method implicitly assumes homogenous treatment effects across trials within each comparison.

The approach described by Bucher *et al* [108] can be extended to time-to-event outcome data by utilising the log hazard ratio and its variance for each comparison such that

$$\log HR_{A:B\ indirect} = \log HR_{A:C} - \log HR_{B:C} \qquad \textbf{(6.3)}$$

$$\text{var}(\log HR_{A:B\ indirect}) = \text{var}(\log HR_{A:C}) + \text{var}(\log HR_{B:C}) \qquad (6.4)$$

This method will be referred to as the aggregate data indirect approach (AD-indirect) in remaining sections.

Hasselblad and Kong [116] write down expressions (6.3) and (6.4) assuming that the two log hazard ratios and their variances are estimated from a Cox proportional hazards model and this will also be assumed in remaining sections of this thesis. However, the log hazard ratio extracted from each trial report may well be estimated from alternative approaches such as Kaplan-Meier curves or a Log-rank analysis. In fact, the empirical results from Chapter 2 of this thesis revealed that the latter methods are more likely to be reported than Cox regression coefficients.

Glenny *et al* (personal communication) note that four times as many similar sized trials are needed for the indirect approach to have the same power as directly randomised comparisons. Suppose for one trial, the estimated treatment effect $\hat{\theta}$ has variance σ^2. For a meta-analysis of 2J trials of the same size and assuming a common true treatment effect, the variance for the pooled treatment effect (θ_{pooled}) using an inverse variance weighted average would be given by

$$\text{var}(\hat{\theta}_{pooled}) = \frac{1}{\sum_{j=1}^{2J} 1/\sigma_j^2} = \frac{1}{2J(1/\sigma^2)} = \frac{\sigma^2}{2J}$$

Now suppose that J trials compare treatment A to C and J different trials compare treatment B to C, with an assumption of equal variances. For each comparison, the variance for the pooled treatment effect using an inverse variance weighted average is given by σ^2 / J. The expected variance for the treatment effect comparing treatment A to B, estimated indirectly is given by the sum of the variances for the A versus C and B versus C comparisons,

$$\frac{\sigma^2}{J} + \frac{\sigma^2}{J} = \frac{2\sigma^2}{J}$$

Therefore, it can be seen that one directly randomised trial is as precise as an indirect comparison based on four randomised trials of the same size. They note that this relation will be approximately true when θ is estimated from J trials of varying sizes. The result also depends on the assumption of equal variances for the three pair-wise comparisons considered.

If multiple trials contribute to each pair-wise comparison, the pooled meta-analysis results for each comparison may be obtained from either a fixed or random effects model. The indirect comparison results based on pooled estimates from random effects models incorporates the additional variation across trials within each comparison. This may lead to a larger estimate of indirect variance if the within comparison trial estimates vary more than expected due to chance.

An alternative formulation of the approach described by Bucher *et al* [108] can be obtained within a meta-regression framework. Each relative treatment effect (log odds ratio for binary data or log hazard ratio for time-to-event data) is fitted as the dependent variable in a model with indicator variables representing each comparison. A weighted regression is used to account for sampling variability. A similar approach has been described previously by Gleser and Olkin [115] for the meta-analysis of categorical data from trials with multiple treatment groups and a common control arm. The general method could be used to analyse direct and indirect comparisons together as well as incorporating multiple related treatment effects from within the same study by appropriate allowance for covariances. Although they describe the model assuming common effects across studies, the model could theoretically be extended to assume random treatment effects. However, the standard random effects meta-regression framework for aggregate data described in Chapter 5 may not be entirely appropriate for this as the estimate of between trial variability from such an approach would relate to the variability across all trials from both sets of trials contributing to the indirect comparison. For this application the heterogeneity across trials would require estimation for each comparison separately as heterogeneity across all trials, regardless of treatment comparison, would be somewhat meaningless. A more complex multi-level model could accommodate such an analysis.

6.4. Indirect individual patient data comparisons with time-to-event data

Hasselblad [111] describes how a logistic regression model may be used for the meta-analysis of trials involving multiple treatment arms for categorical or continuous aggregate data. A multilevel model to accommodate trials with 3 treatment arms has been described by Higgins *et al* [60] for the analysis of continuous outcome data. The model assumes random treatment effects and could be applied to accommodate the analysis of indirect and direct data together. A similar model to that described by Higgins *et al* [60] can in principle be tailored to accommodate other types of outcome measures assuming either random or fixed treatment effects. Extending these principles to model time-to-event outcomes with individual patient data is undertaken in sections 6.4, 6.6.2 and 6.7. To the current authors knowledge this has not been undertaken elsewhere previously.

Consider three treatments A, B and C with no trials directly comparing A and B, but two sets of independent trials comparing A with C or B with C. For the indirect comparison of failure time data, the stratified Cox proportional hazards models described in Chapters 3 and 5 could be extended to include two treatment indicator variables x_{1ij} and x_{2ij} to represent treatments A, B and C. The treatment coding structure assumed below codes treatment A as $x_{1ij}=1, x_{2ij}=0$, treatment B as $x_{1ij}=0, x_{2ij}=1$ and treatment C as $x_{1ij}=0, x_{2ij}=0$.

Using notation described previously, the Cox proportional hazards model stratified by trial for a fixed treatment effect analysis may then be written

$$\lambda_{ij} = \lambda_{0j}(t)\exp(\beta_1 x_{1ij} + \beta_2 x_{2ij}) \qquad (6.5)$$

where β_1 is the common log hazard ratio among trials for the direct comparison of treatment A to C and β_2 for the direct comparison of treatment B to C.

Model (6.5) assumes a common relative effect across trials within each comparison but due to the trial stratification, the baseline hazard function corresponding to the hazard

of event for individuals on treatment C, is allowed to vary across trials. Following the same underlying principle as Bucher *et al* [108], the log hazard ratio for the indirect comparison of treatment A to B is given by the difference in relative effects between the direct comparison of treatment A to C and the direct comparison of treatment B to C,

$$\log \hat{HR}_{A:B\,indirect} = \hat{\beta}_1 - \hat{\beta}_2 \qquad (6.6)$$

The set of trials directly comparing A to C only contribute to the estimation of parameter β_1 in model (6.5) and trials directly comparing B to C only contribute to the estimation of β_2. In fact, due to the likelihood construction of the stratified model, if two separate Cox models, each with a single treatment indicator variable, were fitted using the two independent sets of trials, the corresponding estimates of log hazard ratios and standard errors would be expected to be identical to those obtained from model (6.5). As the two sets of trials are independent it follows that β_1 and β_2 are independent and the standard error for the indirect log hazard ratio comparing A to B is given by

$$SE(\log \hat{HR}_{A:B\,indirect}) = \sqrt{SE(\hat{\beta}_1)^2 + SE(\hat{\beta}_2)^2} \qquad (6.7)$$

This indirect estimate of standard error will, by definition, be greater than either of the individual direct log hazard ratio standard errors reflecting the fact that uncertainty surrounding the indirect estimate correctly incorporates uncertainty from both sources of evidence. Further discussion and implications for non-independent parameters is given in sections 6.6.2 and 6.7.2.

If the log hazard ratio for each comparison is allowed to vary across trials through a random effects analysis, the model may be written as

$$\lambda_{ij} = \lambda_{0j}(t)\exp(\beta_{1j}x_{1ij} + \beta_{2j}x_{2ij}) \qquad (6.8)$$

$$\beta_{1j} = \beta_1 + b_{1j}$$

$$\beta_{2j} = \beta_2 + b_{2j}$$

where β_1 and β_2 are the average log hazard ratio for a population of possible treatment effects of A versus C and B versus C respectively. The parameters b_{1j} and b_{2j} are deviations of the log hazard ratio in the *jth* trial from the relevant population average. We further assume that

$$\begin{pmatrix} b_{1j} \\ b_{2j} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_1^2 & 0 \\ 0 & \tau_2^2 \end{pmatrix}\right)$$

without allowance for covariance between b_{1j} and b_{2j} to reflect that independent sources of data are used for estimation of the two average log hazard ratios and the heterogeneity parameters. Therefore, for the case when only 2-arm trials are included, each trial would only contribute to the estimation of either β_{1j} or β_{2j}. It follows that the average log hazard ratio for the indirect comparison of treatment A to B is given by $\hat{\beta}_1$ - $\hat{\beta}_2$ with SE equal to $\sqrt{SE(\hat{\beta}_1)^2 + SE(\hat{\beta}_2)^2}$. The heterogeneity parameter associated with the indirect log hazard ratio is given by $\hat{\tau}_1^2 + \hat{\tau}_2^2$ which may be interpreted as a measure of variation in the population of indirect treatment effects.

To distinguish approaches for estimating the log hazard ratio and variance indirectly using aggregate data or IPD, the above two approaches based on IPD are referred to as IPD-indirect approaches.

Parameter estimation of model (6.8) is undertaken by extending the approach outlined in Chapter 5 for the estimation involving a single treatment indicator variable (model 5.4, SFE/RE in Chapter 5).

6.5. Assumption of no interaction between treatment and covariates

The AD-indirect approach is valid if it is reasonable to assume that the relative treatment effects used for computing the indirect estimate are consistent across different trial settings. Consider two separate independent trials that are used to estimate an indirect comparison of treatment A versus B. If treatments A and C were compared in trial setting 1, whilst treatments B and C were compared in trial setting 2, one would

be required to assume that the two relative effects would be similar in both trial settings. The indirect estimate could thus be applied to both settings. This assumption is similar to the frequently made assumption of common treatment effect across trials in a fixed effect meta-analysis. If considerable variation were identified between the two trial settings, the assumption may be less realistic. Furthermore, if the relative treatment effects were expected to differ according to a patient characteristic (i.e. treatment-covariate interaction), the indirect estimate may be invalid. Baker and Kramer [121] illustrate the dangers of interpreting a simple indirect comparison if there is an underlying interaction between treatment and a binary variable. They consider the case in which trial settings differ according to the variable. Similar graphical methods are used below to illustrate dangers, describe what could happen if trial settings involved similar values of the characteristic under consideration, and discuss how limitations of the simple AD-indirect method may be overcome using IPD.

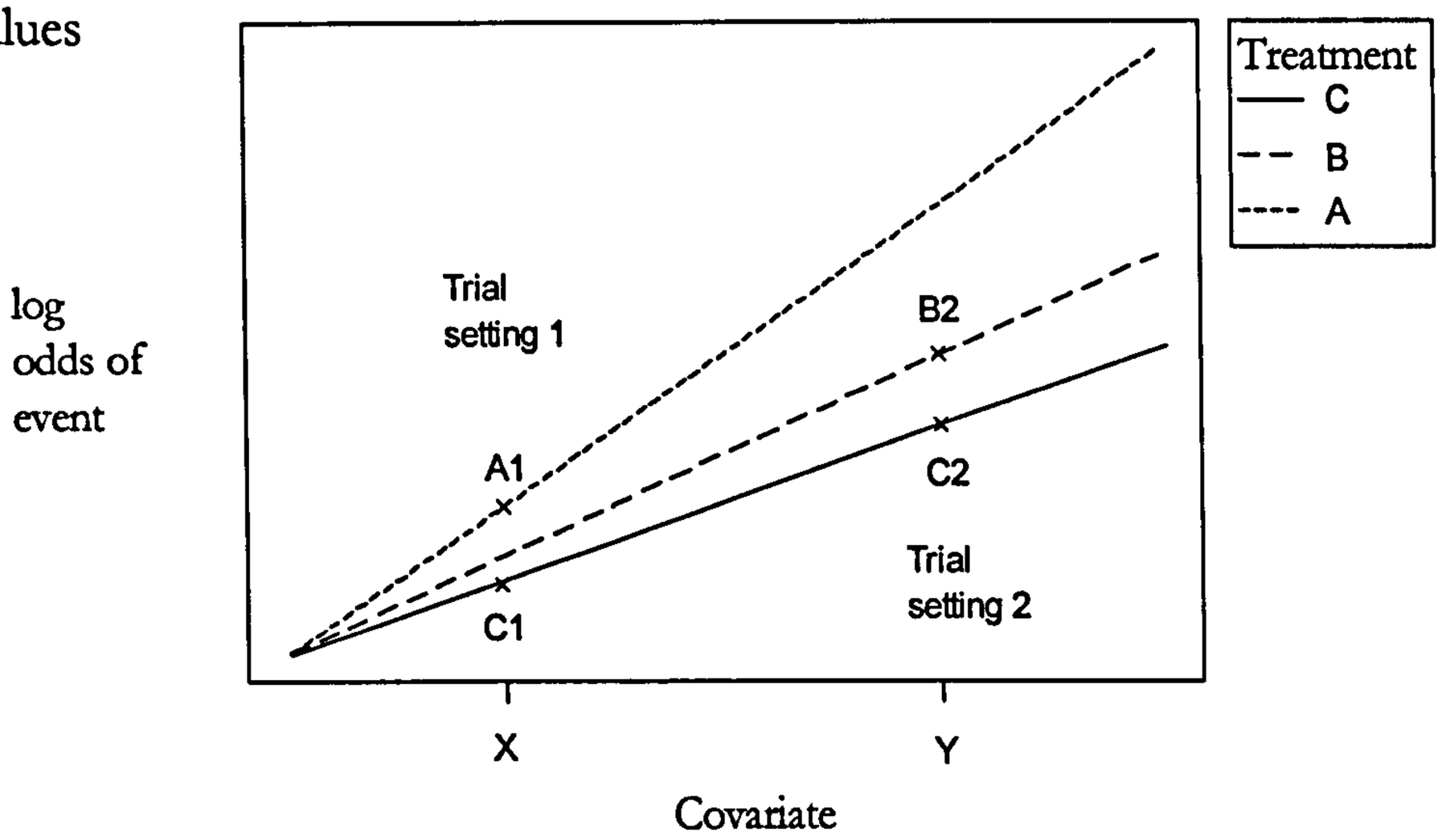
Example 1. Underlying treatment-covariate interaction and trial settings vary according to the covariate

Suppose that treatments A and C were compared in trial setting 1 (represented by A1 and C1, Figure 6.1) which included patients with covariate value X, and treatment B was compared to C in trial setting 2 (represented by B2 and C2, Figure 6.1) which included patients with covariate value Y. Figure 6.1 demonstrates that the log odds of a particular event and the relative treatment effects are greater for patients with covariate value Y compared to X (a quantitative interaction).

Since the relative treatment effect for A versus C can only be estimated for patients with covariate value X (trial setting 1) whilst the relative effect for B compared to C can only be estimated for patients with covariate value Y (trial setting 2), the underlying interaction cannot be identified in this example and an AD-indirect estimate of relative effect (A versus B) would be inaccurate and difficult to interpret as it cannot be applied to either setting. The corresponding IPD-indirect method would also suffer from the same problem, highlighting the need to examine comparability of characteristics and, if possible, assess whether the assumption of no interactions is reasonable. It may not be possible, as in this example, to formally assess the evidence for an interaction therefore it is important to discuss with clinicians and review literature regarding interactions in

previous related RCTs to consider whether there are any clinical reasons to expect relative effects to differ according to clinical factors.

Figure 6.1. Example 1: Illustration of treatment-covariate interaction with indirect comparison where trial settings involve different covariate values



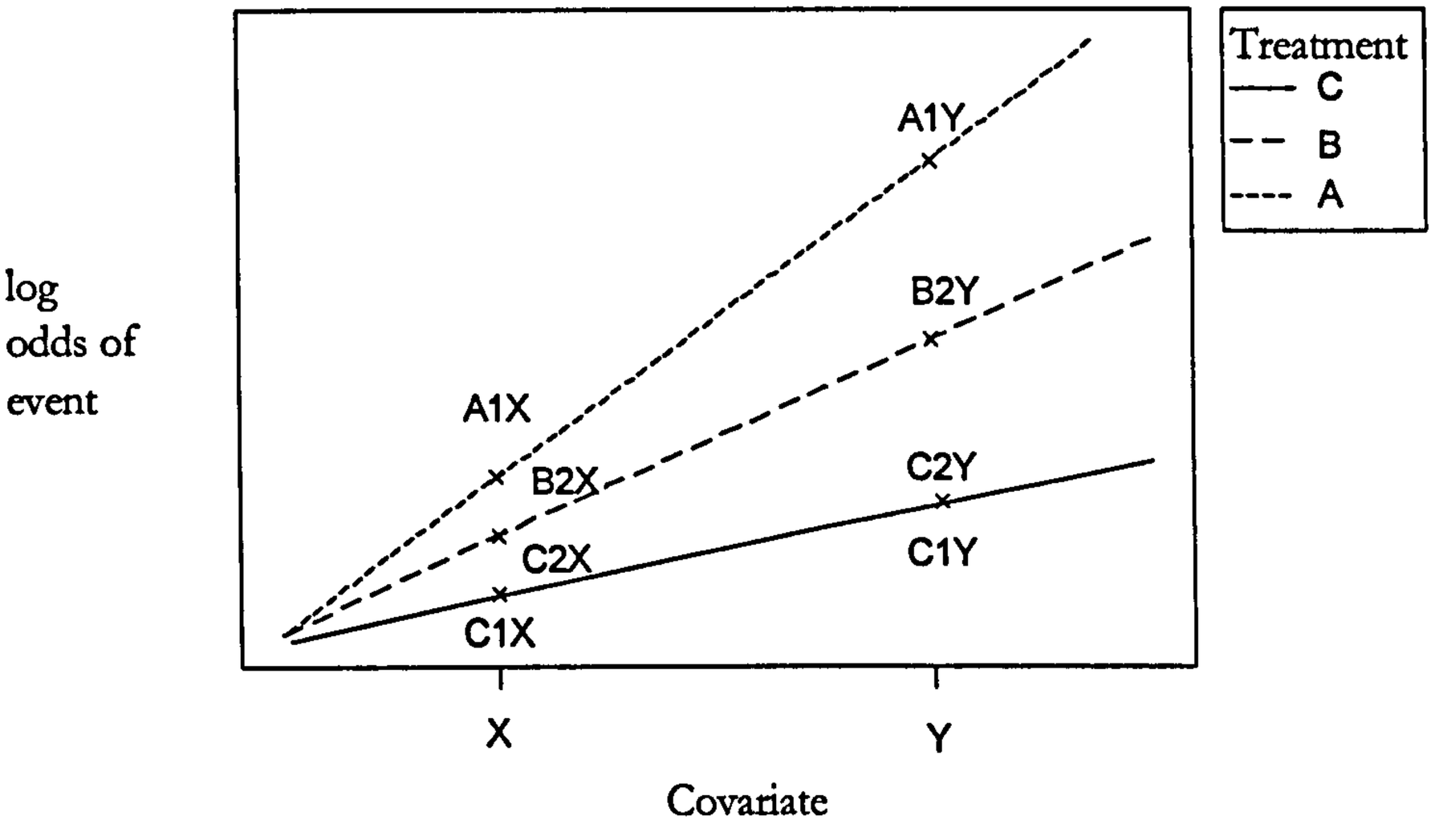
Example 2. Underlying treatment-covariate interaction with similar range of covariate value in different trial settings

Suppose treatment A and C were compared in trial setting 1 (represented by A1X, A1Y, C1X and C1Y, Figure 6.2) which included patients with covariate values X and Y, whilst treatment B was compared to C in trial setting 2 (represented by B2X, B2Y, C2X and C2Y, Figure 6.2) which also included patients with covariate values X and Y.

The relative treatment effects for A versus C and B versus C can be estimated for patients with covariate value X and Y. However, since the relative effects differ according to this covariate value, a simple indirect estimate unadjusted for the interaction should be avoided. For a binary covariate, the AD-indirect comparison of A versus B could be estimated within each subgroup defined by the covariate value. This would require suitable aggregate data to be presented for each subgroup within both trials. The AD-indirect comparison would be difficult to estimate for a continuous covariate. The IPD-indirect comparison could be estimated for both a binary or continuous covariate by including a suitable interaction term in the regression model. Furthermore, a more thorough investigation of the evidence for interactions between

treatment and covariate values, with the possibility to adjust for multiple interactions, can be undertaken with IPD. In situations where interactions exist, the AD-indirect method for indirect comparison is likely to be very limited.

Figure 6.2. Example 2: Illustration of treatment-covariate interaction with indirect comparison where trial settings involve similar covariate values



6.6. Combining indirect and direct evidence

Evidence may be available for the direct comparison of treatments A versus B in addition to indirect evidence estimated from the comparisons of A versus C and B versus C. It may be sufficient to summarise and estimate the indirect comparison and contrast with the direct comparison to explore consistency. In some situations, it may be appropriate to combine both sources of evidence to obtain an overall estimate of effect. This may be particularly attractive if there is limited or inconclusive direct evidence or if further clinical trials providing a direct comparison are unlikely. However, careful consideration should be given to determine whether a combined analysis is appropriate within a particular setting. Issues to consider, such as comparability of patients and clinical heterogeneity, should be similar to those examined prior to any meta-analysis.

6.6.1. Aggregate Data

The indirect treatment effect estimated using the AD-indirect method, denoted $\log HR_I$, may be combined with the direct treatment effect, denoted $\log HR_D$, by calculating an inverse variance weighted average given by

$$\log HR_{combined} = \frac{\dfrac{\log HR_D}{\text{var}(\log HR_D)} + \dfrac{\log HR_I}{\text{var}(\log HR_I)}}{\text{var}(\log HR_D)^{-1} + \text{var}(\log HR_I)^{-1}}$$

with variance for the combined estimate given by

$$\text{var}(\log HR_{combined}) = \frac{1}{\text{var}(\log HR_D)^{-1} + \text{var}(\log HR_I)^{-1}}$$

This approach will be referred to as AD-combined to distinguish from AD-indirect results. Either a fixed or random effects model may be assumed for combining the direct and indirect estimate, the latter making allowance for heterogeneity between sources of direct and indirect evidence. However, as there would only be two data points in a random effects analysis, the estimation of variability between sources of evidence is likely to be imprecise.

6.6.2. Individual patient data

Data from both sources of direct and indirect evidence may also be accommodated in an individual patient data regression model, referred to here as the IPD-combined approach. For the IPD-indirect approach, the regression coefficients and variance components of models (6.5) and (6.8) are independent when the two sets of trials contributing to the indirect comparison are 2-arm trials independently comparing either A with C or B with C. The additional incorporation of individual patient data from one or more trials that provide a direct comparison of treatments A and B introduces further complexity. Recall that β_1 and β_2 are the log hazard ratios comparing treatments A to C and treatments B to C respectively. The set of trials that directly compare A with B contribute to estimating both of these parameters which are therefore no longer independent. For the fixed effect model (6.5) the IPD-combined estimate of log hazard

ratio for treatment A compared to treatment B is given by $\hat{\beta}_1 - \hat{\beta}_2$. To account for non-independence of regression coefficients the covariance between these parameters $\text{cov}(\beta_1, \beta_2)$ is estimated and the standard error of the IPD-combined log hazard ratio is given by $\sqrt{SE(\hat{\beta}_1)^2 + SE(\hat{\beta}_2)^2 - 2 * \text{cov}(\hat{\beta}_1, \hat{\beta}_2)}$. For the random effects model (6.8), inclusion of direct evidence further creates dependence between the heterogeneity parameters τ_1^2 and τ_2^2 which can be modelled by assuming the following structure

$$\begin{pmatrix} b_{1j} \\ b_{2j} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_1^2 & \tau_{12} \\ \tau_{12} & \tau_2^2 \end{pmatrix} \right)$$

where τ_1^2 describes heterogeneity for the IPD-combined log hazard ratio comparing treatment A to C, τ_2^2 describes heterogeneity for the IPD-combined effect comparing treatment B to C, and $\tau_1^2 + \tau_2^2 - 2\tau_{12}$ for the IPD-combined effect comparing A to B.

Parameters of the IPD-combined model assuming fixed relative treatment effects are easily estimated using standard statistical software. For the IPD-combined model assuming random treatment effects, the parameter estimation procedure adopted for the IPD-indirect model (based on the approach outlined in Chapter 5) could theoretically be extended to accommodate the non-zero covariance structure for the correlated random effects. Further research is required to enable this procedure to be implemented but the computational aspects are beyond the scope of this thesis. An alternative IPD-combined approach that is more straightforward, particularly if assuming random treatment effects, involves analysing both sources of evidence (direct and indirect) using two separate stratified Cox regression models with the combined log hazard ratio and its variance estimated using an inverse variance weighted average. Similarly, results from trials where only AD are available may be combined with results of IPD based analyses using an inverse variance weighted average provided the AD and IPD sources of evidence are independent. Alternatively, a single multi-level model or hierarchical Bayesian framework may allow all data to be analysed simultaneously with the latter approach incorporating uncertainty surrounding parameters of the model which would be particularly beneficial for examples involving small numbers of trials.

As described in Chapter 3, the stratified Cox regression model for standard meta-analysis can produce different results to the inverse variance weighted average of within-trial estimates obtained from separate Cox models. There is therefore a possibility that the IPD-combined estimates based on a single stratified Cox model (including all direct and indirect IPD together) may not always be consistent with the inverse variance weighted average of direct and indirect evidence.

6.7. Incorporating trials with more than two treatments

Trials with multiple treatment groups provide an additional level of information in the analysis of indirect comparisons as they give an insight into the association between alternative pair-wise treatment effects from within the same population. As these trials provide both direct and indirect evidence from within the same population they increase confidence in the underlying assumption that different clinical questions relating to each pair-wise comparison are clinically relevant within the same population.

A trial with H treatment groups can provide up to $\binom{H}{2}$ pair-wise comparisons which are not independent if one treatment group is common to each comparison. For example, in a 3-arm trial comparing treatments A, B and C, the relative effects of A versus C and B versus C are not independent as patients included in group C are used in both calculations.

6.7.1. Aggregate Data

The AD-indirect method implicitly assumes that each trial contributes information from two treatment groups only as the dependence between relative treatment effects from within the same study would not be accommodated. The AD-indirect method could in theory be extended to incorporate multiple treatment arm trials, for example a 3-arm trial, by including the covariance between relative effects derived from within the same trial for estimating the variance of the indirect comparison. The indirect log hazard ratio would be estimated using expression (6.3) and the variance given by

$$\text{var}(\log HR_{A:B\ indirect}) = \text{var}(\log HR_{A:C}) + \text{var}(\log HR_{B:C}) - 2 * Cov(\log HR_{A:C}, \log HR_{B:C})$$

However, it seems highly unlikely that the results of a 3-arm trial would be presented as treatment effect parameters for only two pair-wise comparisons, their variances and the covariance between them. Extending the AD-indirect method to incorporate multiple treatment trials is most likely of limited value and will not be considered further.

6.7.2. Individual patient data

If IPD are available from a multiple treatment arm trial, inclusion of data for estimating the indirect comparison would, by definition, also include data for the direct comparison and these cannot be disentangled. The multiple treatment arm trial can be incorporated into the IPD-combined model described in section 6.6.2 as appropriate recognition is made for the dependence structure by estimating the covariance between regression parameters.

6.8. Illustration of methods

6.8.1. Indirect evidence from trials with two treatments

To illustrate methods described in preceding sections the comparison between VPS and PHT is chosen as the primary comparison for the outcome time to 12 month remission. Trials comparing VPS with CBZ and trials comparing PHT with CBZ are defined as the external indirect evidence for this primary comparison. Two trials (De Silva 1996 [45] and Heller 1995 [48]) compare CBZ, VPS and PHT within the same population but for simplicity and to illustrate results when only 2-arm trials are involved, the VPS arm of these two trials are excluded, thus data from both trials only contribute to the comparison between CBZ and PHT. The indirect evidence therefore comprises 3 trials, 551 patients, 289 events for the PHT versus CBZ comparison De Silva 1996 [45], Heller 1995 [48], Mattson 1985 [86] and 3 trials, 1000 patients, 598 events Mattson 1992 [38], Richens 1994 [47], Verity 1995 [46] for the VPS versus CBZ comparison.

Meta-analysis results, using a Cox model stratified by trial, for both pair-wise comparisons (Table 6.1) suggests that time to achieve a period of 12 month remission is similar for PHT when compared with CBZ (HR 95%CI: 0.99(0.79, 1.25)) with no evidence for heterogeneity across the 3 trials included in that comparison. For the comparison between VPS and CBZ, time to achieve 12 month remission is significantly shorter for CBZ (HR 95%CI: 0.83(0.70, 0.97)). However, evidence for heterogeneity across the 3 trials suggests that the fixed effect model may be unreasonable and after allowing for heterogeneity, the result is no longer statistically significant (HR 95%CI: 0.84(0.60, 1.16)).

Table 6.1. Pair-wise comparisons from separate stratified Cox models with fixed or random treatment effects

		Fixed effect model		Random effects model		
Comparison	**HR>1 favours**	logHR (SE)	HR (95%CI)	logHR (SE)	HR (95%CI)	τ^2 (SE)
PHT:CBZ 3 trials, 289 events, 551 patients	**PHT**	-0.0076 (0.1193)	0.99 (0.79,1.25)	-0.0076 (0.11931)	0.99 (0.79,1.25)	0 .
VPS:CBZ 3 trials, 598 events, 1000 patients	**VPS**	-0.1905 (0.0824)	0.83 (0.70,0.97)	-0.17827 (0.16617)	0.84 (0.60,1.16)	0.0625 (0.0828)

The aim for this illustration is to estimate the hazard ratio and 95% CI for the indirect comparison between VPS and PHT using the external evidence from the 3 trials comparing PHT to CBZ and 3 separate trials comparing VPS to CBZ using approaches described in sections 6.3 and 6.4.

Aggregate data

If sufficient aggregate data were available in original trial reports to allow estimation of log hazard ratios and their standard errors, the AD-indirect approach (expression 6.1 and 6.2 or the random effects alternative) could be used to estimate the log hazard ratio and standard error for the indirect comparison of VPS:PHT. In this example from

epilepsy, sufficient aggregate data were not available therefore the estimates generated from IPD (Table 6.1) are used to illustrate the AD-indirect calculations. Clearly, without IPD, further calculation and exploration of AD-indirect methods could not be undertaken for this example. Recall that the indirect evidence may be estimated by

$$\log \hat{HR}_{VPS:PHTindirect} = \log \hat{HR}_{VPS:CBZ} - \log \hat{HR}_{PHT:CBZ}$$

$$\text{var}(\log \hat{HR}_{VPS:PHTindirect}) = \text{var}(\log \hat{HR}_{VPS:CBZ}) + \text{var}(\log \hat{HR}_{PHT:CBZ})$$

Under the assumption of a fixed effect of treatment across trials within each pair-wise comparison,

AD-indirect fixed

$$\log \hat{HR}_{VPS:PHTindirect} = -0.1905 - (-0.0076) = -0.1829$$

$$\text{var}(\log \hat{HR}_{VPS:PHTindirect}) = 0.0824^2 + 0.1193^2 = 0.0210$$

whilst assuming random treatment effects across trials within each pair-wise comparison,

AD-indirect random

$$\log \hat{HR}_{VPS:PHTindirect} = -0.17827 - (-0.0076) = -0.17067$$

$$\text{var}(\log \hat{HR}_{VPS:PHTindirect}) = 0.16617^2 + 0.1193^2 = 0.0418$$

Individual patient data

Both fixed effect (model 6.5) and random effects (model 6.8) Cox models stratified by trial are adopted to estimate the indirect comparison of VPS versus PHT with individual patient data (IPD-indirect with fixed or random effects). In model (6.5) and (6.8), the treatment coding structure adopted is such that $x_{1ij} = 1$ for treatment group VPS, $x_{2ij} = 1$ for treatment group PHT, with CBZ taking value zero for both variables.

Parameter estimates and standard errors from the stratified Cox model based on the indirect evidence are shown below with an illustration of the calculations involved to estimate the treatment effect and standard error for the indirect comparison.

Under the assumption of a fixed effect of treatment across trials within each pair-wise comparison (model 6.5),

$$\log \hat{HR}_{VPS:CBZ} = \hat{\beta}_1 = -0.1910, \quad SE(\hat{\beta}_1) = 0.0824$$

$$\log \hat{HR}_{PHT:CBZ} = \hat{\beta}_2 = -0.0076, \quad SE(\hat{\beta}_2) = 0.1193$$

IPD-indirect fixed

$$\log \hat{HR}_{VPS:PHTindirect} = \hat{\beta}_1 - \hat{\beta}_2$$
$$= -0.1910 - (-0.0076) = -0.1834$$
$$\text{var}(\log \hat{HR}_{VPS:PHTindirect}) = \text{var}(\hat{\beta}_1) + \text{var}(\hat{\beta}_2)$$
$$= 0.0824^2 + 0.1193^2 = 0.0210$$

Under the assumption of random treatment effects across trials within each pair-wise comparison (model 6.8), the following results are obtained

$$\log \hat{HR}_{VPS:CBZ} = \hat{\beta}_1 = -0.17827, \quad SE(\hat{\beta}_1) = 0.16616,$$
$$\hat{\tau}_1^2 = 0.0625, \quad SE(\hat{\tau}_1^2) = 0.0828$$

$$\log \hat{HR}_{PHT:CBZ} = \hat{\beta}_2 = -0.0076, \quad SE(\hat{\beta}_2) = 0.1193,$$
$$\hat{\tau}_2^2 = 0$$

IPD-indirect random

$$\log \hat{HR}_{VPS:PHTindirect} = \hat{\beta}_1 - \hat{\beta}_2$$
$$= -0.17827 - (-0.0076) = -0.17067$$

$$\text{var}(\log \hat{HR}_{VPS:PHTindirect}) = \text{var}(\hat{\beta}_1) + \text{var}(\hat{\beta}_2)$$
$$= 0.16616^2 + 0.1193^2 = 0.0418$$

$$\hat{\tau}^2_{VPS:PHTindirect} = \hat{\tau}_1^2 + \hat{\tau}_2^2$$
$$= 0 + 0.0625 = 0.0625$$

Direct evidence comparing VPS and PHT

In this example so far, interest has focused on estimating the indirect comparison of VPS to PHT using external indirect evidence. For the outcome time to 12 month remission, 124 events and 287 patients from two additional independent trials Turnbull 1985 [122] and Craig 1994 [123]) provide direct evidence for the pair-wise comparison between VPS and PHT. Results from fitting a Cox regression model stratified by trial to this direct evidence are displayed in Table 6.2.

Table 6.2. Direct pair-wise comparison between VPS and PHT from a Cox model stratified by trial with fixed or random treatment effects

Comparison HR>1 favours	Fixed effect model		Random effects model		
	logHR (SE)	HR (95%CI)	logHR (SE)	HR (95%CI)	τ^2 (SE)
VPS:PHT VPS 2 trials, 124 events, 287 patients	-0.0868 (0.1804)	0.92 (0.64,1.31)	-0.08665 (0.18041)	0.92 (0.64,1.31)	0

The overall hazard ratio suggests a slight non-significant trend toward favouring PHT (HR 95%CI: 0.92(0.64, 1.31)) but this result is inconclusive as clinically important results cannot be excluded from the wide confidence interval. There is no evidence of heterogeneity between the two trials reflected by an identical random effects result and an estimate of zero for the between trial variability parameter τ^2.

Summary and comparison of indirect and direct results

Estimates of the hazard ratio and 95% confidence interval for the indirect comparison between VPS and PHT are summarised in Table 6.3. For comparison, the indirect estimates for the other two comparisons are also displayed. Hence, the two trials directly comparing VPS:PHT and three trials directly comparing VPS:CBZ are used to estimate the indirect comparison of PHT:CBZ, whilst two trials comparing VPS:PHT and 3 trials comparing PHT:CBZ provide the indirect evidence for the comparison between VPS:CBZ. Detailed calculations for these two comparisons are summarised in Appendix D (section D.1. and D.2).

The hazard ratios estimated from direct evidence for each pair-wise comparison (Table 6.1 and Table 6.2) agree well with the corresponding indirect comparison (Table 6.3). Good agreement is seen across all three comparisons with overlapping confidence intervals (direct versus indirect) and qualitatively similar results.

For each comparison and assumption of either fixed or random treatment effects, the AD-indirect results are exactly equal to the IPD-indirect results. This is because the two sets of trials included in the IPD-indirect stratified Cox model are independent, hence regression coefficient estimates and standard errors are equivalent to those obtained from fitting two separate stratified models to each set of trials. As the latter estimates (Table 6.1 and Table 6.2) are used in the AD-indirect calculations, the results are not surprisingly equivalent. In practice, the AD-indirect and IPD-indirect results are unlikely to be equivalent due to differences in terms of included patients and methods of analysis that inevitably occur when comparing AD and IPD results [49].

The indirect evidence comparing VPS:PHT (Table 6.3) includes over seven times as many events and five times as many patients compared to the corresponding direct evidence (Table 6.2). The 95% CI for the indirect HR is thus narrower than the direct result. For the PHT:CBZ the indirect evidence consists of just over twice as many events and patients compared to the direct evidence. For the VPS:CBZ comparison, more patients and events are included in the direct evidence analysis. Since four times as many similar sized trials are needed for the indirect approach to have the same power as directly randomised comparisons (Glenny *et al.*, personal communication), the 95% CIs for the direct hazard ratio are each narrower than the indirect evidence for these two comparisons.

Table 6.3. Estimates for each indirect comparison using AD or IPD with fixed or random effects

Comparison	AD-indirect					IPD-indirect				
	Fixed effect		Random effects			Fixed effect		Random effects		
	logHR (SE)	HR (95%CI)	logHR (SE)	HR (95%CI)	τ^2 (SE)	logHR (SE)	HR (95%CI)	logHR (SE)	HR (95%CI)	τ^2 (SE)
VPS:PHT [1]	-0.1829 (0.1449)	0.83 (0.63, 1.11)	-0.1707 (0.2045)	0.84 (0.56, 1.26)	0.0625 (0.083)	-0.1834 (0.1449)	0.83 (0.63, 1.11)	-0.1707 (0.2045)	0.84 (0.56, 1.26)	0.0625 (0.083)
PHT:CBZ [2]	-0.1037 (0.1982)	0.90 (0.61, 1.33)	-0.0916 (0.2454)	0.91 (0.56, 1.48)	0.0625 (0.083)	-0.1038 (0.1982)	0.90 (0.61, 1.33)	-0.0917 (0.2453)	0.91 (0.56, 1.48)	0.0625 (0.083)
VPS:CBZ [3]	-0.0944 (0.2163)	0.91 (0.60, 1.39)	-0.0943 (0.2163)	0.91 (0.60, 1.39)	0	-0.0944 (0.2163)	0.91 (0.60, 1.39)	-0.0942 (0.2163)	0.91 (0.60, 1.39)	0

[1] Based on 6 trials, 887 events, 1551 patients [HR>1 favours VPS]. Calculations displayed in section 6.8.1.

[2] Based on 5 trials, 722 events, 1287 patients [HR>1 favours PHT]. Calculations displayed in Appendix D.1.

[3] Based on 5 trials, 413 events, 838 patients [HR>1 favours VPS]. Calculations displayed in Appendix D.2.

6.8.2. Combining direct and indirect evidence

If both direct and indirect estimates of hazard ratio are available for each comparison, these may be combined to obtain an overall summary of both sources of evidence.

Aggregate data

For each comparison, the indirect log hazard ratio and its variance estimated using an AD-indirect approach (Table 6.3) can be combined with the relevant corresponding direct estimate (Table 6.1 and Table 6.2) using an inverse variance weighted average of both sources of evidence. To illustrate, consider the comparison between VPS and PHT assuming fixed effects within comparison and between sources of evidence. The combined estimate, denoted $\log HR_{VPS:PHTcombined}$ may be calculated as

$$\log HR_{VPS:PHTcombined} = \frac{\frac{\log HR_{VPS:PHTdirect}}{\text{var}(\log HR_{VPS:PHTdirect})} + \frac{\log HR_{VPS:PHTindirect}}{\text{var}(\log HR_{VPS:PHTindirect})}}{\text{var}(\log HR_{VPS:PHTdirect})^{-1} + \text{var}(\log HR_{VPS:PHTindirect})^{-1}}$$

$$= \frac{\frac{-0.0868}{0.0325} + \frac{-0.1829}{0.0210}}{1/0.0325 + 1/0.0210} = -0.1452$$

$$\text{var}(\log HR_{VPS:PHTcombined}) = \frac{1}{\text{var}(\log HR_{VPS:PHTdirect})^{-1} + \text{var}(\log HR_{VPS:PHTindirect})^{-1}}$$

$$= \frac{1}{1/0.0325 + 1/0.0210} = 0.0128$$

In addition to assuming either fixed or random effects within each set of trials that contribute to the indirect estimate, it is also possible to adopt either a fixed or random effects model for pooling both sources of evidence. The latter approach incorporates additional variability between both sources of evidence.

Individual patient data

When IPD are available, a combined analysis of direct and indirect evidence can be achieved by including all data in one stratified Cox regression model with appropriate recognition of covariance as described previously. Alternatively, the indirect estimate (IPD-indirect assuming either fixed or random treatment effects) and direct estimate could be combined using the inverse weighted average approach. Again, a fixed or random effects model may be assumed for pooling both sources of evidence.

Results are given in Table 6.4 using the inverse variance weighted average method for combining aggregate data estimates or IPD estimates assuming either fixed or random effects within each pair wise comparison. The result from fitting a single stratified Cox model with fixed treatment effects for the analysis of direct and indirect evidence is also included for comparison.

The combination of direct and indirect results (Table 6.4) incorporates additional information leading to an improvement in precision compared with the direct evidence considered in isolation (Table 6.1 and Table 6.2). The greatest improvement in precision occurs for the VPS:PHT comparison with minimal improvement for both other comparisons. This is because the indirect evidence for the VPS:PHT comparison is more precise than the direct evidence and therefore contributes more to the combined analysis.

Table 6.4. Combining direct and indirect evidence from 2-arm trials

Comparison*	Inverse variance weighted method				FE stratified Cox model including indirect and direct evidence
	AD-indirect and direct		IPD-indirect and direct		
	Combining FE estimates	Combining RE estimates$^{\phi}$	Combining FE estimates	Combining RE estimates$^{\phi}$	
VPS:PHT					
logHR(SE)	-0.1456 (0.113)	-0.1234 (0.135)	-0.1456 (0.113)	-0.1234 (0.135)	-0.1456 (0.113)
HR(95%CI)	0.86 (0.69,1.08)	0.88 (0.68,1.15)	0.86 (0.69,1.08)	0.88 (0.68,1.15)	0.86 (0.69,1.08)
PHT:CBZ					
logHR(SE)	-0.0331 (0.102)	-0.0237 (0.107)	-0.0331 (0.102)	-0.0237 (0.107)	-0.0332 (0.102)
HR(95%CI)	0.97 (0.79,1.18)	0.98 (0.79,1.21)	0.97 (0.79,1.18)	0.98 (0.79,1.21)	0.97 (0.79,1.18)
VPS:CBZ					
logHR(SE)	-0.1783 (0.077)	-0.1471 (0.132)	-0.1783 (0.077)	-0.1471 (0.132)	-0.1787 (0.077)
HR(95%CI)	0.84 (0.72,0.97)	0.86 (0.67,1.12)	0.84 (0.72,0.97)	0.86 (0.67,1.12)	0.84 (0.72,0.97)

* Each based on 8 trials, 1011 events, 1838 patients. HR>1 favours first drug in comparison

FE: Fixed treatment effect across trials, RE: Random treatment effects across trials

$^{\phi}$ Random treatment effects assumed across trials that provide direct evidence and each set of trials contributing to the indirect comparison

As noted previously, there is good agreement between direct (Table 6.1 and Table 6.2) and indirect results (Table 6.3) across all three comparisons in this example. Adopting either a fixed or random effects model for pooling direct and indirect results using an inverse variance weighted average gave identical estimates of hazard ratio and 95% confidence interval. Although there are only two data points to consider, estimates of I^2 (section 1.3.1) statistic equal to zero for all comparisons suggest that any variability between sources of evidence can be explained by chance. These results add further support to the argument of compatibility and suggest that combination of sources of

evidence is not unreasonable for this example although clinical factors should also be considered.

Although the combined estimates based on AD and IPD methods are identical in this example, estimates generated from IPD have been used in the AD calculations therefore such good agreement is unlikely to reflect reality. Also, in this example, the stratified Cox model with direct and indirect evidence gives identical results to the IV weighted average. As noted in Chapter 3, these two methods can differ under some circumstances although not in the above example.

6.8.3. Including trials with three treatments

The example used in section 6.8.1 and 6.8.2 was somewhat artificial as two trials, De Silva 1996 [45] and Heller 1995 [48], randomised patients to one of three drugs CBZ, VPS and PHT but data for the VPS arm was excluded from both trials to ease interpretation and illustration of methods. Multiple drugs are sometimes compared within the same clinical trial and such information should be fully utilised when considering all the evidence. In the present section, data from all drug groups (CBZ, VPS or PHT) in all eight trials (1095 events, 1948 patients) are included in a single stratified Cox regression model with two treatment indicator variables x_{1ij}, x_{2ij} representing three treatments as described in section 6.8.1 (assuming fixed treatment effects). As data from within the same trial contribute to the estimation of both regression parameters, the covariance between these parameters should be estimated and used for calculating the standard error for the VPS:PHT comparison. Due to the dependence between estimates from a 3-arm trial, a combined analysis based on aggregate data is not considered for reasons described previously.

Due to the inclusion of additional data, the results of fitting a stratified Cox regression model including all 2-arm and 3-arm trials comparing CBZ, VPS and PHT (Table 6.5) indicate further improvement in precision compared with previous combined results (Table 6.4).

Table 6.5. Results from a single fixed effect stratified Cox regression model including all data from 3-arm trials and 2-arm trials

Comparison	HR>1 favours	Regression parameter	logHR (SE)	HR (95%CI)
VPS:PHT	VPS	$\beta_1 - \beta_2$	-0.08377[1] (0.0990)	0.92 (0.76,1.12)
PHT:CBZ	PHT	β_2	-0.0544 (0.09873)	0.95 (0.78,1.15)
VPS:CBZ	VPS	β_1	-0.1382 (0.07011)	0.87 (0.76,1.00)

[1] log hazard ratio is calculated from difference between log hazard ratios of PHT:CBZ and VPS:CBZ comparisons $\hat{\beta}_1 - \hat{\beta}_2$. The standard error is given by $\sqrt{(0.09873^2+0.07011^2-2*0.00243)}$ where $\text{cov}(\hat{\beta}_1, \hat{\beta}_2) = 0.00243$.

6.8.4. Summary of illustration

In this illustration, the AD-indirect results are equal to the IPD-indirect results since the AD-indirect methods use aggregate data generated from IPD. An empirical example using real aggregate data extracted from trial reports should be examined to gain further insight into a comparison of results from both approaches.

There is good agreement between indirect and direct estimates in this example adding support to the clinical justifications for including results in a combined analysis.

Four times as many similar sized trials are needed for the indirect approach to have the same power as a directly randomised comparison and precision of the indirect result will therefore vary by example. For the three comparisons examined here, precision was greater for the indirect log hazard ratio compared to the direct for the VPS:PHT comparison leading to the greatest improvement in precision for the combined estimate.

Realistically, inclusion of a 3-arm trial can only be undertaken with IPD but this improved precision further for all three comparisons.

6.9. Totality of evidence in epilepsy trials

As described in Chapter 4, IPD are available for 4496 patients randomised within 18 RCTs that examine the effect of 6 different AEDs summarised in Table 6.6. These trials provide direct evidence for eight comparisons with respect to the outcomes time to first seizure and time to withdrawal (CBZ-VPS, CBZ-PHT, CBZ-PHB, VPS-PHT, VPS-PHB, PHT-PHB, CBZ-LTG, PHT-OXC) whilst hazard ratios and confidence intervals can be estimated using direct evidence for seven comparisons with respect to the outcome time to 12 month remission (no data for the direct comparison CBZ-LTG). The pooled results, using stratified Cox proportional hazards models, for each comparison and each outcome were estimated in Chapter 4 (Table 4.5). This is the traditional approach to meta-analysis, summarising the available evidence for each pair-wise comparison in turn.

Further evidence is integral to this data set and should be utilised if possible. The relationship between the six AEDs in terms of number of trials and patients (maximum possible overall) contributing to each direct pair-wise comparison is summarised in Figure 6.3.

Figure 6.3. Relationship between AEDs from each direct comparison (maximum number of trials and patients available for analysis for each comparison)

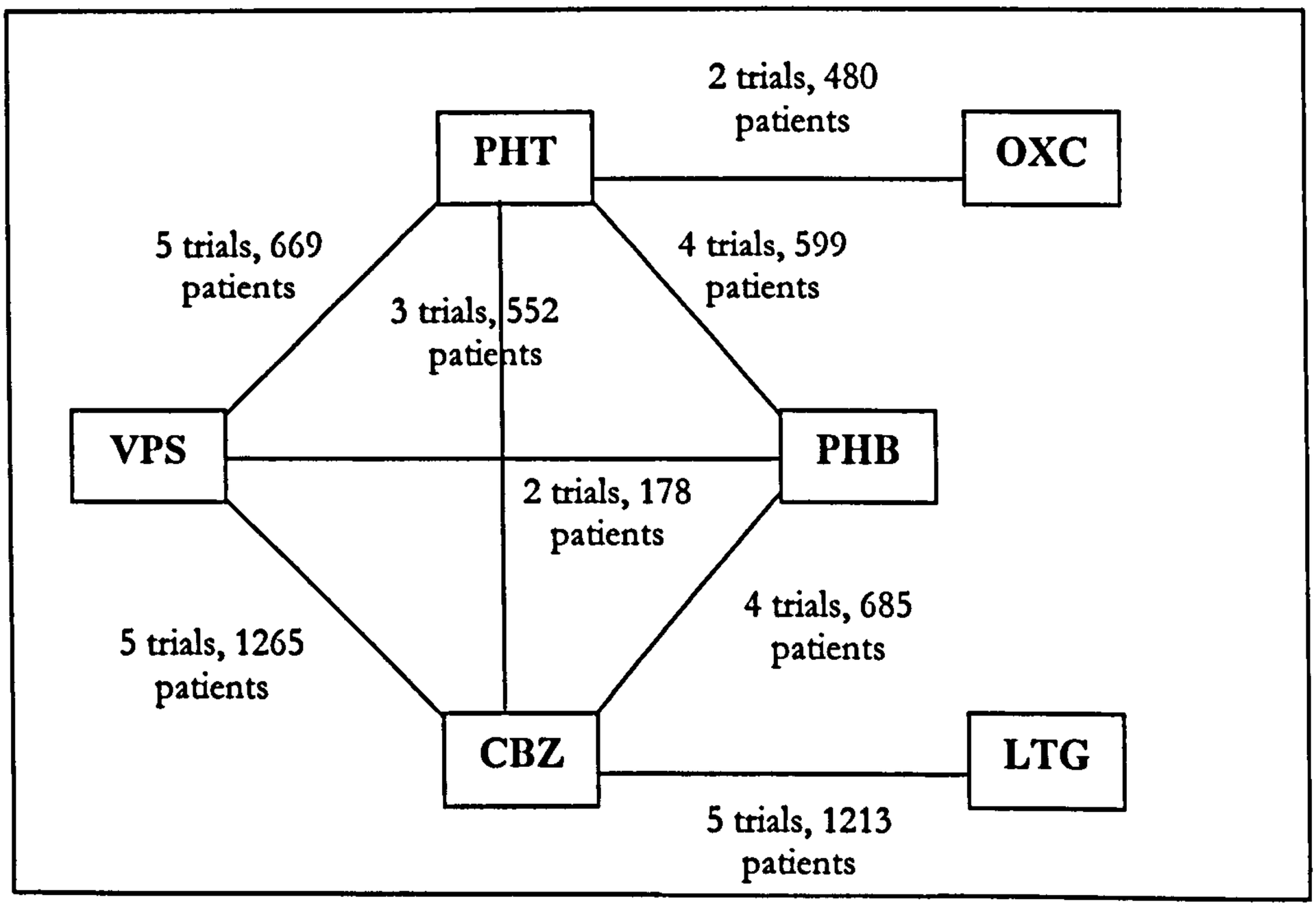


Table 6.6 Number of patients randomised to each drug, pair-wise comparisons examined and IPD availability for each outcome across each trial

Trial	Number randomised to each AED						Pair-wise comparisons examined	IPD available for each time-to-event outcome		
	CBZ	PHB	PHT	VPS	LTG	OXC		Withdrawal	First seizure	12month remission
1.Heller 1995	61	58	63	61	0	0	CBZ-VPS CBZ-PHT CBZ-PHB VPS-PHT VPS-PHB PHT-PHB	Yes	Yes	Yes
2.De Silva 1996	54	10	54	49	0	0	CBZ-VPS CBZ-PHT CBZ-PHB VPS-PHT VPS-PHB PHT-PHB	Yes	Yes	Yes
3.Mattson 1985	155	155	165	0	0	0	CBZ-PHT CBZ-PHB PHT-PHB	Yes	Yes	Yes
4.Mattson 1992	236	0	0	244	0	0	CBZ-VPS	Yes	Yes	Yes
5.Richens 1994	151	0	0	149	0	0	CBZ-VPS	Yes	Yes	Yes

Trial	Number randomised to each AED						Pair-wise comparisons examined	IPD available for each time-to-event outcome		
	CBZ	PHB	PHT	VPS	LTG	OXC		Withdrawal	First seizure	12month remission
6.Verity 1995	**130**	0	0	**130**	0	0	CBZ-VPS	Yes	Yes	Yes
7.Brodie 1995a	**59**	0	0	0	**64**	0	CBZ-LTG	Yes	Yes	No
8.Brodie 1995b	**60**	0	0	0	**59**	0	CBZ-LTG	Yes	Yes	No
9.Reunanen 1996	**120**	0	0	0	**229**	0	CBZ-LTG	No	Yes	No
10.Ramsay 1992	0	0	**50**	**86**	0	0	VPS-PHT	Yes	Yes	No
11.Craig 1994	0	0	**81**	**85**	0	0	VPS-PHT	No	Yes	Yes
12.Turnbull 1985	0	0	**70**	**70**	0	0	VPS-PHT	Yes	Yes	Yes
13.Placencia 1993	**95**	**97**	0	0	0	0	CBZ-PHB	Yes	Yes	Yes
14.Brodie 1999	**48**	0	0	0	**102**	0	CBZ-LTG	Yes	Yes	No
15.Bill 1997	0	0	**144**	0	0	**143**	PHT-OXC	Yes	Yes	Yes
16.Guerreiro 1997	0	0	**96**	0	0	**97**	PHT-OXC	Yes	Yes	Yes
17.Pal 1998	0	**47**	**47**	0	0	0	PHT-PHB	No	Yes	Yes
18.Barrera 2001	**202**	0	0	0	**420**	0	CBZ-LTG	Yes	No	No

CBZ: Carbamazeine, **VPS**: Sodium Valproate, **PHT**: Phenytoin, **PHB**: Phenobarbitone, **LTG**: Lamotrigine, **OXC**: Oxcarbazepine

The trials under consideration provide the capacity to calculate indirect pair-wise comparisons (refer to Figure 6.3). Different levels of evidence are therefore available for the fifteen possible pair-wise comparisons as summarised below.

Evidence available	Pair-wise comparison
Direct and indirect evidence	CBZ-VPS CBZ-PHT CBZ-PHB VPS-PHT VPS-PHB PHT-PHB
Direct evidence alone	CBZ-LTG PHT-OXC
Indirect evidence alone	CBZ-OXC PHB-OXC VPS-OXC LTG-OXC PHB-LTG PHT-LTG VPS-LTG

The indirect evidence may be obtained from multiple sources rather than from only two sets of pair-wise comparisons with a common treatment (refer to Figure 6.3) as considered previously. For example, the indirect evidence for CBZ-VPS may be obtained from the following comparisons: CBZ-PHT and VPS-PHT; CBZ-PHB and VPS-PHB, each of which may in turn be estimated from more than one source of evidence. This complex structure of direct and indirect evidence is referred to from here onwards as the totality of evidence.

6.9.1. Motivation for exploring totality of evidence

The joint analysis of all six drugs from 18 trials to summarise the current total evidence for monotherapy in epilepsy provides the primary motivation for considering the totality of evidence analysis. This analysis requires the assumption that patients are reasonably similar across trials in that individuals entered into each trial would be expected to be eligible for entry into remaining trial settings. If patients were hypothetically switched from one trial to another, the relative treatment effects in each trial should be assumed to remain approximately equal. Patients enrolled into the monotherapy trials are considered to be similar in terms of their epilepsy (Table 4.2, Chapter 4). Most have a

newly diagnosed epilepsy and treatment with either of the AEDs taken as monotherapy would be considered appropriate for such patients. Some trials did in fact include more than two AEDs (Table 6.6) adding further support to the assumption that these comparisons are reasonable in this population of patients. The availability of IPD provides additional justification for exploring the totality of evidence as uniform outcome and event definitions are used across all trials. Further justification of the appropriateness of such an analysis comes from the fact that the protocols used in each original systematic review of pair-wise drug comparisons were identical in terms of clinical questions addressed, outcomes considered, assessment of trial eligibility and review methodology.

Secondly, the unavailability of IPD, or indeed any data, directly comparing some AEDs would lead us to consider the next best level of available evidence. As an example, although the comparison between VPS and LTG would be clinically informative, individual patient data from RCTs that directly compare these AEDs are not currently available. The totality of evidence summary could provide the best evidence for these comparisons, particularly since multiple sources may contribute to the indirect comparison.

The third motivation for exploring indirect comparisons in the epilepsy monotherapy trials is to strengthen conclusions from direct comparisons. This is particularly important for comparisons where the direct evidence available is scarce e.g. PHB-VPS or where there is limited potential for undertaking further studies to provide additional direct evidence. CBZ and VPS are considered first line treatments for epilepsy in Europe and the USA, with VPS the treatment of choice for generalized seizures and CBZ the treatment of choice for partial seizures [69], [124]. Although there is insufficient evidence from RCTs to support these clinical beliefs, this apparent consensus in treatment policy greatly restricts the prospect of conducting further RCTs directly comparing these AEDs. It is important therefore to examine other sources of evidence.

To enable the above issues to be addressed, the totality of evidence from all patients and trials are analysed in the following sections. Due to the complexity and interdependence of pair-wise comparisons within the epilepsy monotherapy trials, these analyses are only

undertaken using individual patient data with an assumption of fixed treatment effects. Only the evidence from randomised controlled trials are included as this would represent a summary of the best available evidence. Indirect comparisons are based on relationships between relative treatment effects. Although indirect comparisons are not themselves based on randomised evidence, if the relative treatment effects are estimated from randomised controlled trials, they would provide better evidence for the indirect comparison than would be estimated from observational studies. For this reason, observational studies have not been considered although could easily be included in a similar way to RCTs.

6.9.2. Exploring treatment main effects using totality of evidence

The Cox proportional hazards model stratified by trial with fixed treatment effect is adopted for the analysis of totality of evidence including all patients from all treatment groups in all trials. The model is a simple extension of the Cox model described for combining direct and indirect evidence or trials with more than 2 arms. The effects of six AEDs are of interest and may be represented in the Cox proportional hazards model by five dummy variables. Choosing OXC (arbitrarily) as the baseline drug, the following totality of evidence model assuming fixed treatment effects, is fitted

$$\lambda_{ij} = \lambda_{0j}(t)\exp(\beta_1 x_{1ij} + \beta_2 x_{2ij} + \beta_3 x_{3ij} + \beta_4 x_{4ij} + \beta_5 x_{5ij}) \qquad (6.9)$$

with the following dummy variable coding structure

Treatment	Dummy variable				
	x_{1ij}	x_{2ij}	x_{3ij}	x_{4ij}	x_{5ij}
CBZ	**1**	0	0	0	0
PHB	0	**1**	0	0	0
PHT	0	0	**1**	0	0
VPS	0	0	0	**1**	0
LTG	0	0	0	0	**1**
OXC	0	0	0	0	0

Estimates of the hazard ratio and its standard error for each pair-wise comparison based on the totality of evidence may be obtained as described in section 6.6.2 and 6.7.2. For example, the hazard ratio and 95% confidence interval for the comparison CBZ to OXC based on the totality of evidence is given by $\exp(\hat{\beta}_1 \pm 1.96 * SE(\hat{\beta}_1))$. For the comparison between CBZ and VPS, the hazard ratio is given by

$$HR_{CBZ:VPS} = \exp[(\hat{\beta}_1 - \hat{\beta}_4)]$$

with standard error

$$SE(\hat{\beta}_1 - \hat{\beta}_4) = \sqrt{SE(\hat{\beta}_1)^2 + SE(\hat{\beta}_4)^2 - 2 * \mathrm{cov}(\hat{\beta}_1, \hat{\beta}_4)}$$

The parameter estimates and standard errors for model (6.9) with respect to each outcome are recorded in Appendix E (Table E.1.1, Table E.1.2, Table E.1.3). As a worked example, consider the comparison between CBZ and VPS for the outcome time to first seizure (Table E.1.1).

$$\hat{\beta}_1 = 0.088 \quad \hat{\beta}_4 = 0.157$$

$$\mathrm{var}(\hat{\beta}_1) = 0.024788 \quad \mathrm{var}(\hat{\beta}_4) = 0.02479 \quad \mathrm{cov}(\hat{\beta}_1, \hat{\beta}_4) = 0.022716$$

$$\mathrm{HR}_{\mathrm{CBZ:VPS}} = \exp[(\hat{\beta}_1 - \hat{\beta}_4)] = \exp[-0.0690] = 0.93$$

$$\begin{aligned}\mathrm{SE}(\mathrm{HR}_{\mathrm{CBZ:VPS}}) = SE(\hat{\beta}_1 - \hat{\beta}_4) &= \sqrt{SE(\hat{\beta}_1)^2 + SE(\hat{\beta}_4)^2 - 2 * \mathrm{cov}(\hat{\beta}_1, \hat{\beta}_4)} \\ &= \sqrt{0.024788 + 0.02479 - 2 * 0.022716} \\ &= 0.064389\end{aligned}$$

$$\mathrm{HR}_{\mathrm{CBZ:VPS}}\ (95\%\mathrm{CI}):\ \exp[(-0.0690) \pm 1.96 * 0.064389] = \mathbf{0.93(0.82, 1.06)}$$

Estimates of the hazard ratio and 95% confidence interval based on totality of evidence are summarised in Table 6.7, Table 6.8, and Table 6.9 for all comparisons and each outcome. In Chapter 4, individual stratified Cox proportional hazards models with a single treatment indicator variable (fixed effect) were fitted to each direct comparison

for each outcome (Table 4.5). These results are reproduced in Tables 6.7, 6.8, 6.9 to allow comparison with totality of evidence results.

Table 6.7. Time to first seizure: Analysis of totality of evidence and direct evidence for comparison

Comparison	HR>1 favours	DIRECT EVIDENCE			TOTALITY OF EVIDENCE
		Number of trials	Events /Total	HR (95% CI)	HR (95% CI)
cbz-oxc	OXC	0		-	1.09 (0.80, 1.49)
phb-oxc	OXC	0		-	0.93 (0.66, 1.29)
pht-oxc	OXC	2	229/472	**1.08 (0.83, 1.40)**	**1.08 (0.83, 1.40)**
vps-oxc	OXC	0		-	1.17 (0.86, 1.59)
ltg-oxc	OXC	0		-	1.27 (0.87, 1.86)
cbz-phb	PHB	4	365/677	**1.17 (0.94, 1.45)**	**1.18 (0.97, 1.44)**
cbz-pht	PHT	3	362/545	**1.10 (0.89, 1.35)**	**1.01 (0.86, 1.19)**
cbz-vps	VPS	5	864/1225	**0.92 (0.81, 1.06)**	**0.93 (0.82, 1.06)**
cbz-ltg	LTG	4	344/741	**0.86 (0.69, 1.08)**	**0.86 (0.69, 1.08)**
phb-pht	PHT	4	351/592	**0.84 (0.68, 1.05)**	**0.86 (0.70, 1.05)**
phb-vps	VPS	2	134/178	**1.05 (0.72, 1.52)**	**0.79 (0.64, 0.98)**
phb-ltg	LTG	0		-	0.73 (0.54, 0.99)
pht-vps	VPS	5	371/639	**0.96 (0.78, 1.18)**	**0.92 (0.78, 1.09)**
pht-ltg	LTG	0		-	0.85 (0.65, 1.13)
vps-ltg	LTG	0		-	0.92 (0.71, 1.19)

Table 6.8. Time to 12 month remission: Analysis of totality of evidence and direct evidence for comparison

Comparison	HR>1 favours	DIRECT EVIDENCE			TOTALITY OF EVIDENCE
		Number of trials	Events /Total	HR (95% CI)	HR (95% CI)
cbz-oxc	CBZ	0	-	-	0.96 (0.67, 1.37)
phb-oxc	PHB	0	-	-	0.90 (0.61, 1.32)
pht-oxc	PHT	2	170/308	**0.92 (0.68, 1.24)**	**0.92 (0.68, 1.24)**
vps-oxc	VPS	0	-	-	0.84 (0.59, 1.20)
cbz-phb	CBZ	4	280/684	**1.08 (0.84, 1.38)**	**1.06 (0.85, 1.33)**
cbz-pht	CBZ	3	289/551	**1.01 (0.80, 1.27)**	**1.04 (0.86, 1.26)**
cbz-vps	CBZ	5	767/1225	**1.14 (0.99, 1.32)**	**1.14 (0.99, 1.31)**
phb-pht	PHB	4	260/562	**0.94 (0.73, 1.22)**	**0.98 (0.78, 1.24)**
phb-vps	PHB	2	130/178	**0.91 (0.62, 1.33)**	**1.07 (0.84, 1.36)**
pht-vps	PHT	4	303/514	**1.04 (0.83, 1.31)**	**1.09 (0.90, 1.33)**

Data for direct comparison CBZ-LTG unavailable for this outcome

Table 6.9. Time to withdrawal: Analysis of totality of evidence and direct evidence for comparison

Comparison	HR>1 favours	DIRECT EVIDENCE			TOTALITY OF EVIDENCE
		Number of trials	Events /Total	HR (95% CI)	HR (95% CI)
cbz-oxc	OXC	0		-	1.61 (0.99, 2.62)
phb-oxc	OXC	0		-	2.35 (1.43, 3.86)
pht-oxc	OXC	2	91/480	**1.65 (1.08, 2.52)**	**1.65 (1.08, 2.52)**
vps-oxc	**OXC**	0		-	1.51 (0.92, 2.47)
ltg-oxc	OXC	0		-	0.96 (0.55, 1.68)
cbz-phb	PHB	4	235/676	**0.68 (0.52, 0.89)**	**0.69 (0.53, 0.88)**
cbz-pht	**PHT**	3	196/546	**0.99 (0.75, 1.31)**	**0.98 (0.77, 1.24)**
cbz-vps	VPS	5	399/1200	**1.03 (0.84, 1.25)**	**1.07 (0.89, 1.28)**
cbz-ltg	LTG	4	209/1032	**1.68 (1.27, 2.21)**	**1.68 (1.27, 2.21)**
phb-pht	PHT	3	211/499	**1.61 (1.21, 2.12)**	**1.43 (1.10, 1.85)**
phb-vps	VPS	2	66/170	**1.75 (1.03, 2.95)**	**1.56 (1.17, 2.07)**
phb-ltg	LTG	0		-	2.45 (1.68, 3.55)
pht-vps	VPS	4	137/495	**1.05 (0.75, 1.47)**	**1.09 (0.85, 1.40)**
pht-ltg	LTG	0		-	1.72 (1.19, 2.46)
vps-ltg	LTG	0		-	1.57 (1.13, 2.19)

Totality of evidence: Statistical interpretation

For each outcome and pair-wise comparison there is good agreement between totality of evidence and direct evidence based results (Tables 6.7, 6.8, 6.9). Due to the inclusion of additional data, the confidence intervals obtained from the totality of evidence model are each narrower, or the same width, compared with the corresponding confidence interval obtained from a model that includes only direct evidence for that particular comparison. For the comparisons between CBZ-LTG and PHT-OXC, the estimates and confidence intervals are unaffected by including totality of evidence since these pair-wise comparisons cannot be obtained indirectly from other trials (refer to Figure 6.3).

The totality of evidence analysis has most noticeable effect in terms of improving precision for the CBZ-PHT and PHB-VPS comparisons for time to first seizure (Table 6.7), PHB-VPS comparison for time to 12-month remission (Table 6.8) and PHB-PHT, PHB-VPS, PHT-VPS for time to withdrawal (Table 6.9). This may be explained by recognising that for these pair-wise comparisons the number of individuals and events within the trials providing direct evidence is smaller relative to those providing indirect evidence.

By using the totality of evidence, hazard ratios and 95% confidence intervals are estimated for pair-wise comparisons of AEDs that have no direct evidence from randomised controlled trials (Table 6.7, 6.8, 6.9). Although these findings are clinically helpful as a summary of the current available evidence, they require cautious interpretation as the results are based solely on indirect evidence. Additional reassurance is gained from the fact that the totality of evidence results agree well with direct estimates where data are available for comparison.

Direct evidence for the PHB-VPS comparison suggests a statistically non-significant difference between PHB and VPS for the outcome time to first seizure. The totality of evidence for this comparison suggests a statistically significant clinical advantage in favour of PHB. Since confidence intervals are estimated for multiple comparisons using the same data there is increased possibility of obtaining a spurious result which should be considered in an overall clinical interpretation. The presentation of 99% confidence intervals for the totality of evidence results may be more appropriate.

Totality of evidence: Clinical interpretation

Time to first seizure

The totality of evidence analysis suggests that time to first seizure is not statistically significantly different for OXC when compared to CBZ, PHB, PHT, VPS and LTG. Since confidence intervals for each hazard ratio are wide, clinically important differences in favour of either drugs cannot be excluded and equivalence cannot be concluded.

Although results are not statistically significant for pair-wise comparisons with CBZ, the evidence suggests that PHB may have an increased length of time before first seizure (clinically better), VPS and LTG have a shorter length of time compared to CBZ and PHT a similar length of time to CBZ.

The totality of evidence suggests that PHB is clinically better for the outcome time to first seizure compared with PHT, VPS and LTG. For the latter two pair wise comparisons the 95% confidence intervals do not include unity but these results are based either on indirect evidence alone (PHB-LTG) or on limited direct evidence (PHB-VPS) and should therefore be viewed with caution.

For remaining pair-wise comparisons with PHT the totality of evidence suggests a trend toward favouring PHT compared to VPS or LTG although neither result is statistically significant.

Finally, although there may be a slight trend in favour of VPS compared to LTG, there is no evidence of a statistically significant difference between the dugs and since this comparison is based entirely on indirect evidence conclusions should be viewed cautiously.

Time to 12 month remission

There is no evidence for a statistically significant difference between OXC and CBZ, PHB, PHT or VPS for this outcome but 95% confidence intervals are too wide to conclude equivalence amongst each pair-wise comparison.

When compared with CBZ, the two drugs PHB and PHT may have similar effectiveness but confidence intervals include clinically important differences and are again too wide and to conclude equivalence. When compared with VPS, the evidence suggests that the time taken to achieve a period of 12 month remission may be shorter with CBZ although the confidence interval includes unity.

PHT and VPS have similar effectiveness when compared with PHB but equivalence cannot be concluded due to the wide confidence interval for the hazard ratio for both pair-wise comparisons.

There is no evidence for a statistically significant difference between PHT and VPS for this outcome.

Time to withdrawal

The totality of evidence analysis suggest that the time before withdrawal due to adverse events or poor seizure control is significantly longer for OXC compared to PHB or PHT, with a suggestion, although not statistically significant, to favour OXC when compared to either CBZ or VPS. OXC and LTG have similar effectiveness but clinically important results cannot be excluded from the wide confidence interval hence equivalence of these two drugs should not be concluded. Direct evidence is only available for the PHT-OXC pair-wise comparison.

Time to withdrawal is significantly longer for CBZ compared to PHB. There is no evidence for a statistically significant difference between CBZ and PHT or VPS for this outcome but equivalence cannot be concluded. Time to withdrawal is significantly longer for LTG when compared to CBZ.

For this outcome, the evidence in favour of PHT, VPS or LTG when each are compared to PHB is statistically significant although the latter pair wise comparison (PHB-LTG) does not include direct randomised evidence.

There is no evidence for a significant difference between PHT and VPS whilst both drugs are significantly worse than LTG although these two results are based entirely on indirect evidence.

Clinical conclusions for totality of evidence analysis

The overall trends from the totality of evidence analysis suggest that PHB may be the most effective drug in terms of the outcome time to first seizure whilst LTG is least effective. For the outcome time to 12 month remission, comparisons with LTG cannot be made. Since none of the results from pair-wise comparisons are statistically significant and confidence intervals are wide for many, conclusions regarding the best and worst drug cannot be drawn. OXC and LTG are most favoured for increasing time to withdrawal and hence indicating better tolerability whilst PHB is least favoured. However, pair wise comparisons with OXC are mostly based on indirect evidence and should be viewed cautiously. These results highlight the need for randomised controlled trials providing direct comparisons with LTG and OXC. The largest randomised trial (SANAD) in epilepsy is currently underway and will provide evidence for a direct comparison between CBZ-OXC, CBZ-LTG and LTG-OXC.

6.9.3. Exploring the interaction between treatment and epilepsy type using totality of evidence for time to first seizure

As described in section 6.5, an interaction between treatment and a covariate can complicate or invalidate the interpretation of an indirect comparison. The unadjusted totality of evidence results presented in section 6.9.2 are based on a Cox model including main effect of treatment terms with an underlying assumption that the relative treatment effect for each comparison (all fifteen comparisons) is similar across all covariate values. For the epilepsy monotherapy comparisons, this assumption is clinically unlikely since there are strong beliefs that some AEDs have different effects for generalised and partial epilepsy, i.e. a treatment by epilepsy type interaction. The clinical anticipation of the underlying interaction motivated further investigation of the effect of epilepsy type upon the totality of evidence results. The following strategy was adopted for the outcome time to first seizure.

Step 1. For each direct evidence pair-wise comparison, separate Cox models (stratified by trial with fixed treatment effect) including treatment indicator variable and the main effect of epilepsy type were examined.

Step 2. For each direct evidence pair-wise comparison, separate Cox models (stratified by trial with fixed treatment effect) including treatment indicator variable, main effect of epilepsy type, and a treatment-type interaction term was examined.

Step 3. The totality of evidence model (model 6.9) was fitted with a common main effect of epilepsy type represented by dummy variable x_{6ij} which takes the value of one for partial epilepsy and zero for generalised epilepsy.

$$\lambda_{ij} = \lambda_{0j}(t)\exp(\beta_1 x_{1ij} + \beta_2 x_{2ij} + \beta_3 x_{3ij} + \beta_4 x_{4ij} + \beta_5 x_{5ij} + \beta_6 x_{6ij}) \quad \textbf{(6.10)}$$

Step 4. The totality of evidence model (model 6.9) was fitted with a common main effect of epilepsy type x_{6ij} plus five terms representing interaction between treatment and epilepsy type,

$$\begin{aligned}\lambda_{ij} = \lambda_{0j}(t)\exp(&\beta_1 x_{1ij} + \beta_2 x_{2ij} + \beta_3 x_{3ij} + \beta_4 x_{4ij} + \beta_5 x_{5ij} + \beta_6 x_{6ij} \\ &+ \beta_7 x_{7ij} + \beta_8 x_{8ij} + \beta_9 x_{9ij} + \beta_{10} x_{10ij} + \beta_{11} x_{11ij})\end{aligned} \quad \textbf{(6.11)}$$

where, for partial epilepsy

Treatment	Dummy variable				
	x_{7ij}	x_{8ij}	x_{9ij}	x_{10ij}	x_{11ij}
CBZ	**1**	0	0	0	0
PHB	0	**1**	0	0	0
PHT	0	0	**1**	0	0
VPS	0	0	0	**1**	0
LTG	0	0	0	0	**1**
OXC	0	0	0	0	0

To investigate the statistical significance of terms added to a Cox regression model, analyses of nested models should be based on the same data. For 60/3785 patients across all trials (31 of whom had an event), data on type of epilepsy were not available.

These patients are excluded in the following models for consistency hence the main effect of treatment models may differ slightly to previously reported results in Table 4.5 and Table 6.7. These exclusions are not considered to introduce much bias as 98% of the data are retained.

In a standard Cox model, the change in the value of –2logL (log-likelihood ratio statistic) due to adding variables $x_{(p+1)ij}, x_{(p+2)ij}, \ldots, x_{(p+q)ij}$, to a model including variables $x_{1ij}, x_{2ij}, \ldots, x_{pij}$ provides a test of the null hypothesis that the q parameters $\beta_{p+1}, \beta_{p+2}, \ldots, \beta_{p+q}$ are all zero. For the totality of evidence analysis, the change in –2logL for a model including all interaction with treatment variables (model 6.11) compared to the model without (model 6.10) provides an overall test for the null hypothesis that all interaction terms are zero. To explore whether the hazard ratio for individual pair-wise comparisons are different for partial and generalised epilepsy in a totality of evidence analysis, the individual interaction term relating to that comparison should be excluded and the change in –2logL for the reduced model should be compared to that of fitting a model with all interaction terms (model 6.11). For example, the difference in –2logL between model (6.11) and a model omitting variable x_{7ij} would provide a test for $H_0 : \beta_7 = 0$, i.e. that the hazard ratio CBZ:OXC is the same for partial and generalised epilepsy. For comparisons that do not involve the baseline drug OXC, the model can be re-parameterised using a different drug as the baseline. Values of –2logL, change in –2logL on omitting each individual interaction with treatment term, and corresponding p-value for each pair-wise comparison are summarised in Appendix E (Table E.2.4).

Results: Exploring interactions between treatment and epilepsy type

Parameter estimates and values of –2logL for models exploring the main effect of epilepsy type and interaction with treatment using only direct evidence (step 1 and step 2) are recorded in Appendix E (Table E.2.1). Corresponding parameter estimates from the totality of evidence analysis (step 3 and step 4) are also summarised in Appendix E (Table E.2.2 and Table E.2.3). The hazard ratio and 95% confidence intervals calculated from these parameter estimates are summarised in Table 6.10. The p-value for an interaction between treatment and epilepsy type is recorded in Table 6.10 for each

comparison from the analysis of direct evidence only and for totality of evidence. The p-value is presented to enable results in terms of statistical significance to be contrasted between direct and totality of evidence analyses. A more detailed summary of the calculations involved with values of –2logL are given in Appendix E (Table E.2.1 for direct evidence, Table E.2.4 for totality of evidence).

Statistical interpretation

For the outcome time to first seizure, there is evidence for a statistically significant interaction between treatment and epilepsy type for individual direct evidence comparisons between CBZ-PHB and CBZ-VPS. Although not statistically significant there are also trends to suggest that the hazard ratio may differ between partial and generalised epilepsies for the direct comparison between PHB-VPS and PHT-VPS. As interactions with treatment are found for individual direct comparisons, the totality of evidence analysis also incorporates terms for interactions between treatment and epilepsy type.

Comparison of the adjusted direct and adjusted totality of evidence results indicate generally good agreement across pair-wise comparisons, with improved precision for the hazard ratio within each epilepsy type subgroup.

In this example, the p-value for a treatment-type interaction in the adjusted totality of evidence analysis is smaller compared to the adjusted direct analysis except for the CBZ-PHB pair-wise comparison. Evidence for a statistically significant treatment-type interaction is suggested for eight comparisons (VPS-OXC, LTG-OXC, CBZ-PHB, CBZ-VPS, CBZ-LTG, PHB-PHT, PHB-VPS, PHB-LTG) in the totality of evidence analysis. For three of these comparisons, the evidence for an interaction comes from indirect comparisons only and should be interpreted with caution.

Table 6.10. Analysis of totality of evidence for time to first seizure adjusted for epilepsy type and type by treatment interaction terms

HR>1 favours	Comparison	Totality unadjusted HR(95% CI)	Type	Direct Adjusted HR (95% CI)	Direct p-value for interaction	Totality Adjusted HR (95% CI)	Totality p-value for interaction
OXC	cbz-oxc	1.09 (0.80,1.49)	P			1.13 (0.80, 1.59)	0.56
			G			0.99 (0.64, 1.53)	
OXC	phb-oxc	0.93 (0.66,1.29)	P			0.85 (0.59, 1.23)	0.43
			G			1.05 (0.64, 1.74)	
OXC	pht-oxc	1.08 (0.83,1.40)	P	1.10 (0.81, 1.49)	0.52	1.14 (0.85, 1.53)	0.14
			G	0.91 (0.54, 1.52)		0.79 (0.52, 1.22)	
OXC	vps-oxc	1.17 (0.86,1.59)	P			1.39 (0.98, 1.97)	**0.03**
			G			0.82 (0.53, 1.27)	
OXC	ltg-oxc	1.27 (0.87,1.86)	P			1.55 (1.02, 2.36)	**0.07**
			G			0.95 (0.56, 1.61)	
PHB	cbz-phb	1.18 (0.97,1.44)	P	1.39 (1.09, 1.78)	**0.01**	1.33 (1.06, 1.67)	**0.07**
			G	0.74 (0.49, 1.12)		0.94 (0.67, 1.31)	
PHT	cbz-pht	1.01 (0.86,1.19)	P	1.13 (0.89, 1.43)	0.84	0.99 (0.82, 1.20)	0.11
			G	1.07 (0.71, 1.62)		1.25 (0.96, 1.61)	
VPS	cbz-vps	0.93 (0.82,1.06)	P	0.82 (0.70, 0.96)	**0.02**	0.81 (0.70, 0.94)	**0.002**
			G	1.15 (0.91, 1.45)		1.21 (0.97, 1.50)	

Table 6.10. Analysis of totality of evidence for time to first seizure adjusted for epilepsy type and type by treatment interaction terms

HR>1 favours	Comparison	Totality unadjusted HR(95% CI)	Type	Direct Adjusted HR (95% CI)	Direct p-value for interaction	Totality Adjusted HR (95% CI)	Totality p-value for interaction
LTG	cbz-ltg	0.86	P	0.75 (0.57, 0.99)	0.30	0.73 (0.56, 0.94)	**0.05**
		(0.69,1.08)	G	0.98 (0.64, 1.49)		1.03 (0.74, 1.45)	
PHT	phb-pht	0.86	P	0.79 (0.62, 1.02)	0.28	0.75 (0.59, 0.94)	**0.005**
		(0.70,1.05)	G	1.05 (0.68, 1.62)		1.33 (0.94, 1.89)	
VPS	phb-vps	0.79	P	0.77 (0.45, 1.32)	0.16	0.61 (0.47, 0.79)	**0.0003**
		(0.64,0.98)	G	1.28 (0.78, 2.09)		1.29 (0.92, 1.80)	
LTG	phb-ltg	0.73	P			0.55 (0.39, 0.77)	**0.003**
		(0.54,0.99)	G			1.10 (0.71, 1.72)	
VPS	pht-vps	0.92	P	0.82 (0.61, 1.09)	0.26	0.82 (0.67, 1.00)	0.26
		(0.78,1.09)	G	1.03 (0.77, 1.38)		0.97 (0.76, 1.24)	
LTG	pht-ltg	0.85	P			0.74 (0.54, 1.01)	0.53
		(0.65,1.13)	G			0.83 (0.56, 1.23)	
LTG	vps-ltg	0.92	P			0.90 (0.67, 1.20)	0.80
		(0.71,1.19)	G			0.86 (0.60, 1.24)	

Clinical interpretation

There is insufficient evidence for an interaction between treatment and epilepsy type for the CBZ-OXC, PHB-OXC and PHT-OXC comparisons. When OXC is compared to VPS or LTG, comparisons which are based entirely on indirect evidence, the totality adjusted result suggests that patients with partial epilepsy would benefit more from OXC whilst for generalised epilepsy the patterns of effect are less clear but suggest that VPS may be better than OXC whilst there is insufficient evidence of a difference between LTG and OXC.

For pair-wise comparisons with CBZ, the adjusted totality results suggest that for the CBZ-PHB comparison, PHB is better for partial epilepsy with similar effects for both drugs with generalised epilepsy; for the CBZ-PHT comparison, the two drugs have similar effects for partial epilepsy whilst PHT is favoured for generalised epilepsies; for the CBZ-VPS comparison, CBZ is better for partial seizures and VPS is better for generalised seizures; for the CBZ-LTG comparison, CBZ is better for partial seizures but both drugs have similar effects for generalised seizures.

For remaining pair-wise comparisons involving PHB, patients with partial epilepsy have a better clinical outcome with PHB compared to PHT, VPS and LTG. For these pair-wise comparisons patients with generalised epilepsy have a better clinical outcome with PHT, VPS or LTG but the confidence intervals for the latter subgroup are wide.

There is no evidence for an interaction between treatment and epilepsy type for the PHT-VPS, PHT-LTG and VPS-LTG comparisons.

Conclusions for totality of evidence analysis adjusted for interactions between treatment and epilepsy type

The totality of evidence analysis utilises evidence from direct and indirect comparisons. Since interactions between treatment and epilepsy type were anticipated clinically and identified statistically using the direct pair-wise comparison evidence, the totality of evidence analysis requires appropriate adjustment to be made for these interactions.

The confidence intervals for hazard ratios estimated within each epilepsy type subgroup are narrower for the totality of evidence adjusted analysis. The adjusted analysis allows the clinical effect of drugs to be compared separately for patients with partial and generalised epilepsy in terms of pair-wise comparisons that have not been undertaken previously. For these comparisons, the totality of evidence adjusted analysis provides the best summary of the current available evidence and highlights the need for direct randomised evidence to confirm patterns identified. In some cases however, a randomised controlled trial providing the best level of direct evidence may never be undertaken and clinicians will need to make informed decisions based on the currently available evidence and their clinical knowledge. The adjusted interpretations are more useful clinically due to strong beliefs that certain drugs may be better for one or other epilepsy type.

The availability of IPD for this example allowed a thorough investigation into whether interactions exist and the opportunity for incorporation into the totality of evidence analysis. Such analyses could not be undertaken with aggregate data for this example. In practice, IPD is likely to be required for similar analyses in other areas.

6.10. Discussion

External evidence from indirect comparisons can play an important role in clinical research and may be particularly valuable if evidence from a direct randomised comparison either does not exist, comprises a limited amount of data, or is unlikely to ever be examined in future trials. Several methods exist to allow such an analysis but only methods which make the most of the power of randomisation within each trial and maintain the within trial structure by stratification or comparing relative effects across trials have been considered in this thesis. Nevertheless, evidence from an indirect comparison will always be limited by the fact that differences across trials in patient mix could bias indirect treatment effects and caution is required when interpreting indirect evidence. Since an indirect comparison is not itself a randomised comparison it suffers from many of the disadvantages associated with observational data. It may be argued that evidence from observational studies should be considered when exploring indirect comparisons. However, since the methods considered within this thesis maintain the within trial structure and power of randomisation, the pair-wise comparisons estimated from randomised evidence are likely to be less biased than those obtained from

observational studies. Therefore, the indirect comparison estimated from randomised evidence is likely to be less biased than an indirect comparison estimated from observational data.

In principle, estimates of indirect treatment effect may be obtained using aggregate data but limitations of reporting adequate data for time-to-event outcomes is likely to impose restrictions for this approach. Although estimates of indirect comparison based on aggregate data agree very well with results using IPD in the examples examined, such agreement is unlikely to occur in practice as the AD-indirect comparison calculations utilised IPD generated estimates.

In addition to overcoming the problem of inadequate aggregate data, the availability of IPD ensures that outcome definitions can be standardised across trials contributing to the indirect comparison and also facilitates the most in-depth investigation of consistency in clinical characteristics across trials which can increase confidence with interpretation and generalisability of results. A further limitation with AD-indirect methods is the necessary assumption that relative effects are consistent across covariate values. This assumption can be explored more thoroughly with IPD and if appropriate, adjustments can be made to reflect the validity of such an assumption.

In some situations, combining direct and indirect estimates may be clinically justifiable and the combined estimate may be achieved using an inverse variance weighted average of both sources of evidence. To reflect increased uncertainty around the indirect comparison relative to the direct comparison, proportionately less weight could be applied to the former value. A sensitivity analysis approach applying different weightings between 0 and 100% could provide a range of scenarios to evaluate. However, as four times as many similar sized trials are needed for the indirect approach to have the same power as directly randomised comparisons (Glenny *et al*, personal communication), the indirect evidence is already implicitly 'down-weighted' to a certain extent.

Several pair-wise comparisons involving 6 anti-epileptic drugs have been explored in separate systematic reviews of randomised controlled trials using IPD. Each systematic review provides the best level of evidence currently available for each direct comparison. This is the traditional approach to meta-analysis. From a clinician or

patient's perspective, the results of pair-wise comparisons between alternative AEDs can be difficult to interpret in isolation if there are multiple treatment options available. In this thesis, patients from each treatment group in each trial from these original systematic reviews were included in an overall analysis of the totality of evidence. This analysis provided evidence for pair-wise comparisons that had not been previously undertaken in a randomised controlled trial setting and also improved precision for comparisons where direct evidence was also available. This totality of evidence analysis was felt to be clinically appropriate as the original systematic reviews used identical protocols, trial and patient inclusion criteria and review methodology. The patients included in the monotherapy trials were clinically similar and treatment with any of the AEDs would be considered appropriate.

The totality of evidence analysis presented here simultaneously incorporates direct and indirect evidence from trials with 2 or more treatment groups. The model assumes each relative treatment effect is a fixed effect but can in principle be extended to incorporate random treatment effects with appropriate recognition of covariance terms if necessary. This totality of evidence analysis is considered reasonable for the epilepsy monotherapy trials due to identical review protocols and trial/patient inclusion criteria used for each pair-wise comparison. The model itself does not allow for the possibility that different sources of indirect evidence may produce inconsistent estimates of treatment effect to each other or indeed compared to the direct evidence. The potential problem of inconsistency is unlikely to be an issue for this example due to trial similarities and the a priori clinical justifications for undertaking the combined analysis. A recent publication by Lumley [125] describes an approach for meta-analysis of a network of treatment comparisons from 2-arm trials with continuous or binary data. For example, an indirect estimate of treatment effect between A and B could be estimated from the sets of trials comparing A with C or B with C but also from two further sets of trials comparing A with D and B with D. The approach allows the degree of agreement between these sources of indirect comparison to be quantified and incorporated into the model. Their approach was not considered in this thesis and the authors note that the model would require further extensions to handle multi-armed trials correctly.

The totality of evidence analysis presented in this chapter is novel in terms of the clinical results and in the application of a stratified Cox model for the combined analysis of

direct and indirect evidence for patient level time-to-event data. Ades [127] describe an approach for the analysis of mixed pair-wise comparisons for categorical data, combining information from different studies on different, but structurally related outcomes within a Bayesian framework. Higgins and Whitehead [112] also describe an approach for combining aggregate level categorical data from indirect and direct comparisons within a Bayesian framework. Although the analyses presented in this thesis are undertaken within a classical framework, there are potential benefits to a fully Bayesian analysis which can incorporate uncertainty surrounding parameter estimates which would seem particularly attractive in this setting.

With an increasing number of meta-analyses of pair-wise comparisons of treatments in different clinical areas, the approach to summarising the totality of evidence is likely to be useful in other fields. However, careful consideration is required to evaluate whether such an analysis would be appropriate. If patients are clinically diverse across trials, the totality of evidence analysis may be difficult to apply to a particular population of patients. Although the robustness of an indirect comparison cannot be assessed without direct evidence, such a summary is important as it represents the current available evidence and may highlight the need for high quality randomised evidence directly comparing the AEDs of interest. The indirect evidence could also be used to provide the parameters required for sample size calculation in future RCTs.

This chapter has addressed some important issues associated with indirect comparisons and in particular how IPD may provide additional benefits. From a clinical perspective, the totality of evidence analysis adjusted for interactions between treatment and epilepsy type requires further application to the outcomes time to 12 month remission and time to withdrawal. From a methodological perspective, further work is required to extend the computation capabilities of the stratified Cox model with random treatment effects. This will enable multiple, non-independent random effects to be fitted within the same model which would be the appropriate random effects model to facilitate inclusion of trials with more than two treatments or for the combined analysis of both direct and indirect evidence. This would also be the appropriate model for a random effects totality of evidence analysis which would estimate and incorporate between trial variability within each comparison.

CHAPTER 7

Concluding remarks and further work

Meta-analysis undertaken as a component of a rigorously conducted systematic review is a valuable tool for clinicians, researchers, consumers, and for policy making. The analysis may be based on either aggregate or individual patient data. Although the latter IPD approach is often more costly, time consuming and elaborate, an increasing number of meta-analyses are adopting this approach as it is well recognised as the gold-standard [25]. A recent systematic review (Mark Simmons, personal communication) revealed that IPD based meta-analyses are dominated by reviews that consider time-to-event data in the primary analysis. Limitations of suitably reliable data often preclude the use of an aggregate data approach when considering time-to-event outcomes. The availability of IPD enables thorough investigation of sources of heterogeneity and can be valuable for undertaking multiple indirect comparisons. The main aim of this thesis was to summarise issues, contribute to the development of methodology, and attempt to establish evidence relating to the comparison between IPD and AD based approaches to meta-analysis of time-to-event outcomes.

7.1. Methodological conclusions

Undertaking an AD meta-analysis of time-to-event outcomes can, in theory, be straightforward provided an estimate of log hazard ratio and its variance can be extracted from publications or obtained from original trialists. A further requirement is

that the calculation of length of time, end-point and censoring status should be clearly defined and consistent across all trials. For example, in epilepsy trials the outcome time to first seizure may be defined as time from randomisation to first seizure or, time from end of 6 week drug titration period to first seizure.

In practice, estimates of log hazard ratio and its variance are not often presented directly. Methods exist to approximate the log hazard ratio and its variance using other summary data related to the log-rank test and survival curves [32] that may be presented more frequently. The survival curve approach has been extended by Williamson, Tudur *et al* [33] to incorporate data related to the number at risk which are often quoted within the trial report. This approach was developed to overcome the assumption of constant censoring across the entire follow-up period, an assumption required for a previously proposed survival curve method [32]. The examples examined suggest that the incorporation of numbers at risk can improve approximation but the usefulness of this approach will clearly rely on availability of the required data. A review of survival analyses reported in cancer journals pre-1995 revealed that whilst 88 per cent of articles presented survival curves, only 8 per cent included the numbers at risk on the figures [31]. A more recent review of clinical trials published in four general medical journals (Lancet, New England Journal of Medicine, British Medical Journal, Journal of the American Medical Association) between July and October 1999 [36] found that almost all trials presenting survival plots where there was variable length of follow-up had presented numbers at risk (Tim Clayton, personal communication). For two aggregate data reviews in colorectal cancer and liver surgery examined in this thesis, 3 out of 4 trials presenting survival curves also quoted numbers at risk. Pocock *et al* [36] make a specific recommendation that numbers at risk should be displayed when presenting survival curves. Both AD approaches based on survival curves [32], [33] require careful extraction of probabilities from published curves that may introduce bias. In the absence of direct estimates of log hazard ratio and its variance, the present author would recommend that survival curves be used as a last resort with preference given to the approach incorporating numbers at risk if data are available. Sensitivity analyses are a useful tool to explore the impact of assumptions made with the methods for approximating the individual trial log hazard ratio and its variance.

The aggregate data methods based on extracting the log-rank test p-value are more reliable than survival curve approaches for the examples examined. However, the reliability of this approach will depend on the level of accuracy in reporting the relevant aggregate data. For example, the p-value should be reported to at least 2 decimal places and a log-rank test p-value quoted as 'NS' is not useful. As a further note of caution with this approach, the direction of effect may not be obvious from a p-value (particularly if the treatments have a similar effect) and other sources of data such as a published survival curve should be examined for clarification.

Underlying the interpretation of a hazard ratio is an assumption that hazards are reasonably proportional across time. The assumption can be assessed using three alternative aggregate data based methods that were proposed in this thesis (published by *Statistics in Medicine* [33]). The method of exploring whether individual trial estimates of log hazard ratio vary according to trial summary measure of follow-up is attractive due to its simplicity but would only indicate evidence against the assumption if there were evidence for statistical heterogeneity and sufficient variability in the summary measure of follow-up across trials. Constructing log cumulative hazard plots are feasible but their interpretation is subjective and made particularly difficult when the curves are close together, indicative of small treatment effects. The final method of exploring interval based estimates of log hazard ratio is reasonably straightforward but the test proposed is likely to have low power as the alternative hypothesis is general. The methods were illustrated using a meta-analysis of 5 randomised controlled trials. Ideally, further examples are required to explore the reliability of these methods and the author would recommend that interpretation of results should be undertaken cautiously.

Meta-analysis with individual patient data will involve several important stages. As a minimum, the process would require data acquisition, possibly some data entry, data cleaning, manipulation and validation prior to undertaking the analyses of interest. For the IPD reviews of epilepsy monotherapy trials examined throughout this thesis, the complexity of the outcome data meant that a substantial amount of time was required for programming of the outcomes. Examples of these complexities include different approaches for recording seizure dates across trials, and definitions of endpoint for the outcome time to withdrawal. In other clinical areas, the programming of outcomes may not be as time consuming and detailed.

A fixed effect meta-analysis of individual patient time-to-event data may be undertaken using a number of alternative methods. The stratified log-rank analysis, stratified Cox model, and inverse variance weighted average of Cox model estimates were reviewed and compared in this thesis. The simulation study revealed that similar results were obtained using the three methods for meta-analysis in the absence of heterogeneity and for moderate treatment effects. For larger effects, the stratified Cox model and IV weighted average gave similar estimates and are to be preferred to the stratified log-rank analysis. In the presence of heterogeneity, the performance of all three methods was poor. Most published reviews of individual patient failure time data, including reviews of epilepsy monotherapy trials, adopt the stratified log-rank approach to analysis. Re-analysis of the epilepsy meta-analyses using stratified Cox regression models indicate patterns of differences consistent with those suggested by the simulation study. The empirical comparison of methods further suggested that the occurrence of tied event times could be an important factor in terms of choice of method for analysis. Further simulations with a larger number of patients and repetitions and further empirical comparisons are required to establish more specific guidelines on choice of method. Other methods that may be considered by reviewers would be an analysis of all data ignoring the effect of trial (using an overall log-rank analysis or Cox model) or a Cox regression model adjusted for trial effects using indicator variables as described in Chapter 5. However, since the benefit of within-trial randomisation is effectively ignored with both approaches, they are not considered appropriate for meta-analysis in the opinion of the present author. It should be noted that the Cox regression model may not be an appropriate choice if the assumption of proportional hazards is violated. Furthermore, if the time-to-event data follow a particular parametric distribution, a more powerful analysis may be achieved by assuming a particular form of probability distribution for the data. However, since fewer assumptions are made with a semi-parametric Cox model and estimates are likely to be conservative, this approach is expected to be sufficient for the majority of situations.

Exploring clinical, methodological and statistical heterogeneity is an important aspect of the meta-analysis process and possible factors for heterogeneity should be given careful consideration during protocol development. Factors which could modify the treatment effect may be explored using regression models with either aggregate or individual

patient data. For time-to-event outcomes, the aggregate based meta-regression models are reasonably straightforward provided that sufficient data may be extracted for trial level treatment effects and trial level covariates. The limitations of AD meta-regression, apart from the potential lack of suitable data and restrictions related to the number of trials, are well documented [91], [126], [103]. Issues to consider are that associations derived from meta-regressions are observational, and have a weaker interpretation than the causal relationships derived from randomized comparisons [126], and the approach is subject to *ecological bias* which arises when results based on aggregate data are incorrectly assumed to apply at the individual level. Thompson *et al* [126] further note that data dredging is the main pitfall in reaching reliable conclusions from meta-regression which can only be avoided by the pre-specification of covariates.

The availability of individual patient covariate data allows heterogeneity to be explored more thoroughly and avoids potential ecological bias that may exist with aggregate level analyses. These analyses may be undertaken by fitting a Cox regression model assuming either fixed or random treatment effects. Trial effects may be recognised by incorporating indicator variables, stratification, or adopting a random trial effects approach. Including trial indicator variables (fixed effect) assumes that each within-trial hazard rate for each treatment group is proportional to a common baseline hazard function. However, since trials included in a meta-analysis are likely to vary in terms of clinical characteristics this assumption is rather restrictive. The Cox model stratified by trial requires that only the hazard rates for treatment groups within each trial are proportional over time. Furthermore, since the stratified model only uses individuals within a trial to define each risk set in the likelihood construction, the stratified approach is more in keeping with the principle of maintaining randomisation within meta-analysis. The main disadvantage with the stratified model is that no direct estimate of the importance of the strata effect is produced and that the precision of estimated coefficients may be diminished if there are a large number of strata [96]. Since most IPD meta-analyses are based on fewer than 10 trials (Mark Simmonds, personal communication) and the importance of trial effects per se are not generally of interest in meta-analysis, the author would recommend using the stratified Cox regression model. The appropriateness of the final approach allowing for random trial effects is somewhat less clear as the underlying assumption, that individuals recruited into different trials are a random sample from a wider collection of populations, is difficult to relate to the

concept of meta-analysis. Higgins *et al* [60] believe the assumption of random trial effects is a degree less plausible than that of random treatment effects since it is likely that treatment effects will be more similar than trial populations across trials in a meta-analysis. In the context of proportional hazards models, O'Quigley and Stare [104] conclude that for group sizes (trial sample size) of five or more, the model corresponding to a random trial effects model examined in this thesis provides no more than modest efficiency gains when compared to a stratified model. On the other hand, for moderate to large numbers of very small groups, of sizes two or three, the efficiency gains of the random effects model can be far from negligible [104]. However, since most meta-analyses do not involve a large number of small trials with two or three patients, it would seem that the stratified model would be most appropriate for the majority of situations.

Allowance for residual heterogeneity, heterogeneity of treatment effects left unexplained by any included covariates, may be achieved by fitting random treatment effects. The development of methodology for incorporating random treatment effects within a Cox model required extending previously reported [62] concepts and a computer program for parameter estimation. The resulting collection of programs for fitting Cox models with random treatment effects were also adapted to incorporate the Efron method for handling ties which is considered a better approximation [96] particularly for the outcome time to 12 month remission examined in epilepsy reviews of monotherapy trials.

Indirect comparisons are useful where direct comparisons either do not exist, comprise a limited amount of data, or are unlikely to be examined in future trials. Investigating methodology for indirect comparisons was very much motivated by the reviews of epilepsy monotherapy trials each of which considered a pair-wise comparison of anti-epileptic drugs. The totality of evidence analysis was felt to be clinically appropriate as the original systematic reviews used identical protocols, trial and patient inclusion criteria and review methodology. Careful consideration should be given to these facts before undertaking similar analyses in other areas. The trials and comparisons explored within these epilepsy reviews form a network of direct and indirect evidence which represents the totality of evidence currently available for monotherapy comparisons. An

improvement in precision for the relative treatment effects is gained from the totality of evidence analysis with good agreement between direct evidence and totality of evidence results in the epilepsy example. Moreover, as the analysis is based on a wider selection of patient populations, the results may be seen to apply to a wider population. The model also provided estimates of hazard ratio and 95% confidence intervals for pair-wise comparisons for which direct randomised evidence are not available. Although this represents the best current evidence for these particular comparisons, the results require careful interpretation due to the nature of an indirect comparison that should be regarded with similar concerns as evidence from observational studies. The fixed effect stratified Cox model is possibly over-simplistic for the totality of evidence analysis, as it makes no allowance for residual heterogeneity of the treatment effects. However the approach to parameter estimation based on the penalized partial likelihood adopted in this thesis would require further modification to allow for the covariance structure between correlated random effects. Since confidence intervals for hazard ratios related to each pair-wise comparison would be as wide, or wider, from a random treatment effects approach, the fixed effect analysis results including the totality of evidence presented in this thesis should be viewed only as a preliminary guide. A further note of caution should be given to the potential for spurious results due to multiple comparisons. This issue was not addressed in detail but could be partially overcome by presenting 99% confidence intervals for the totality of evidence analysis.

7.2. Aggregate data compared to individual patient data

There is no doubt that an IPD approach to meta-analysis is the gold-standard method that will provide the best summary of evidence and the most flexibility, particularly if considering time-to-event outcomes. However, if the extra effort and strain on resources and time were to only provide minor advantages over and above what may be achieved with aggregate data, the IPD approach may not be worthwhile in some settings. Empirical evidence relating to different aspects of the comparison between aggregate and individual patient data approaches is required to enable reviewers to make decisions about which approach may be appropriate if the choice were available. However, when trying to compare methods, there are a number of issues to address which can complicate the general assessment and will likely vary according to example. This thesis has attempted to explore some of these issues to contribute some knowledge to the growing body of evidence concerned with this question. The final decision of

which approach to use will very much depend on the example under consideration, availability of data and resources.

An aggregate data approach with time-to-event data can usually be undertaken quite quickly which can be useful to provide estimates for sample size calculations or as part of the data monitoring process for a randomised trial, or for a preliminary assessment of whether collecting IPD would be worthwhile. Provided the reported aggregate data were based on all randomised patients and definitions of the end-point and censoring procedures were consistent across trials, the AD and IPD (assuming same follow-up) approaches should provide similar estimates of overall treatment effect. The degree of discrepancy would depend on method of approximation or analysis with most uncertainty surrounding the reliability of aggregate data approaches based on published survival curves. If the majority of the evidence from an aggregate data approach were only from survival curves, the author is of the opinion that IPD would be worthwhile.

Most AD meta-analyses of time-to-event outcomes would differ to corresponding IPD based analyses in terms of included patients, amount of follow-up and included trials and reviewers should assess the potential impact of these factors. For example, if AD were only available from a small portion of eligible trials, or if the AD were based on data with a high proportion of patient exclusions, an IPD approach would be worth considering particularly for the former case in which meta-analysis may not be possible. The empirical examples examined in this thesis indicate that results from each approach will be varied and identifying a specific pattern of bias that can be attributed to the aggregate data approach is complex, if not impossible for the comparison of treatment effect.

Individual patient data can bring further advantages apart from the capacity to re-instate previously excluded patients, overcome the potential for within study selective reporting, incorporate additional follow-up data or the ability to identify or include unpublished studies. One particular advantage is the potential for exploring prognostic factors and sources of heterogeneity. With AD, exploring potential source of heterogeneity is problematic due to data limitations, potential for ecological bias, and problems of interpretation. Recent publications [103], [89] have recommended that IPD should be used whenever possible to reliably study patient characteristics and investigate

heterogeneity. The empirical comparison examined in this thesis supports these recommendations.

The individual patient data available for the epilepsy reviews of monotherapy trials were particularly valuable for exploring indirect comparisons and the totality of evidence analyses. IPD provides much more scope for these analyses with the potential to include covariate data, explore consistency of evidence and investigate the potential for interactions between treatment and covariates which could invalidate the interpretation of evidence from indirect comparisons. Such in-depth analyses would not have been possible without IPD.

Further empirical evidence for the comparison between AD and IPD approaches in terms of main treatment effects and meta-regression analyses will be collected in due course with two separate research projects currently underway [57], (Jesse Berlin, personal communication, ESTEEM project).

7.3. Areas for further research

Several areas for further research that are of particular interest to the author have been identified in the process of producing this thesis. Additional empirical evidence comparing the survival curve approach with and without incorporating numbers at risk for estimating log hazard ratio and its variance would be useful. Further simulated data based on a larger numbers of patients and repetitions are required to reliably explore and compare alternative methods of meta-analysis with particular emphasis on Cox regression models for estimating the overall treatment effect and heterogeneity parameter. The suite of programs for fitting random effects Cox models presented in this thesis are extremely valuable but their usefulness may be limited in practice as a large amount of computing time may be required to obtain parameter estimates. Further improvements to the estimation process may be achievable and should be explored in more detail. The potential impact of tied data in terms of discrepancy between estimates from stratified log-rank analyses and stratified Cox models was highlighted for the outcome time to 12 month remission in the epilepsy example. Standard theoretical results for the connection between the Log-rank and Cox regression model described in section 3.5.1 are valid when there are no tied event times. Further modifications to the likelihood function are required to facilitate the occurrence of ties and explore the

connection between these methods under these conditions. The occurrence of ties may be a problem in other reviews with time-to-event outcomes, and this should be explored further. Expanding the current simulation study to explore varying degrees of ties will allow the impact on estimation to be evaluated. The resulting information could be linked with data from current IPD reviews to evaluate how much of a problem ties may be in practice. Either way, it is clear that to adequately explore this issue and potentially adjust analyses accordingly, IPD is essential. The theory and facility to fit a Cox regression model with multiple random treatment effects requires investigation. This model would allow the totality of evidence analysis to be explored with the assumption of random treatment effects which would incorporate any between trial variability in relative effects and is perhaps a more realistic analysis. Further advantages may also be gained from adopting a Bayesian approach to the analysis. From a clinical point of view in relation to the epilepsy reviews included in this thesis, the totality of evidence analyses require further work in terms of exploring treatment-covariate interactions for the outcomes time to withdrawal and 12 month remission as well as exploratory analyses of other factors. Finally, the models considered throughout, for the analysis of individual patient failure time data are each based around the semi-parametric Cox regression model. Further research is required to investigate the potential benefits of other models such as an accelerated life model, or models that assume some parametric distribution for the data.

APPENDIX A

SAS programs for fitting random effects Cox models using the Breslow method for handling ties

A.1. Cox model with fixed trial indicator variables and random treatment effects (FE/RE) using Breslow method for handling ties

```
/* data contains variable id (observation number), sastime (time-to-event), censor
(censoring variable for time-to-event), treat (treatment indicator variable), no1001-
no1005 (indicator variables representing 5 included trials) */

proc sort data=data;
 by sastime descending censor;
run;
data data;
 set data;
 dif=sastime-lag(sastime);
 atrisk=(dif^=0);
 run;
data data1;
 set data;
 where censor=1;
 keep id sastime censor;
run;
data data2;
 set data;
 where censor=0;
 keep id sastime censor;
run;
```

```
data data3;
 set data1;
 keep sastime censor;

proc transpose data=data3 out=data4;
 by sastime;
 var censor;

data censor;
 set data4;
 censor=sum(of col1-col357);
 keep sastime censor;
run;

data data6;
 set censor data2;
proc sort data=data6;
 by sastime;
data data7;
 set data6;
 keep sastime censor;
proc transpose data=data7 out=data8;
 by sastime;
 var censor;
data data9;
 set data8;
 /*CHANGE SO THAT RANGE IS THE NUBER OF COLUMNS PRINTED*/
 /*OFF IN DATA8 */
 censor=sum(of col1-col13);
 keep sastime censor;
run;
data data10;
 set data9;
 keep sastime;
data data;
 merge data data10;
 by sastime;
run;
```

POINT X

```
/* Program for analysis of epilepsy data */

proc iml worksize=100000;

cc1=5;        /* number of trials */
cc2=1225;   /* total sample size */

use data;
read all into P;
```

```
d=P[,3];          /* vector indicating censoring (censor) */
set=P[,10];       /* vector indicating number at risk (atrisk) */

 use data6;       /* vector indicating ties */
 read all into H;
 t=H[,3];
  nn=nrow(t);
 *-----------------------*;
 j=t[1];
  if j > 1 then
   do;
    tt0=repeat(0,j-1,1);
    tt1=j//tt0;
   end;
  else tt1=j;
 *-----------------------*;
  do q=2 to nn;
   j=t[q];
    if j > 1 then
     do;
      tt0=repeat(0,j-1,1);
      tt2=j//tt0;
     end;
    else tt2=j;
   tt1=tt1//tt2;
  end;
 TIES=diag(tt1);
 X=P[,1]||P[,6:9];                /* design matrix (fixed effects) */
Z1=P[,5:9]#P[,1];                 /* design matrix (random effects interaction) */

/* vector of fixed effects */
variable={'treat', 'no1002', 'no1003', 'no1004', 'no1005'};
fixnum=5;                         /* number of fixed effects */
initbeta=j(fixnum,1,0);           /* vector of initial values for beta */

/* random effects */
initu1=j(cc1,1,0);                /* vector of initial values for variance for interaction */

compo={'theta1'};                 /* variance component of frailty */
inithe1=1;                        /* interaction */

prelike=0;
iterate=0;
dif=10000;
L=0;

do k=1 to 5000 until(dif<1e-6);  /* convergence criterion */
iterate=iterate+1;
prelike=L;
newbeta=initbeta;                 /* change estimate */
newu1=initu1;
```

```
newthe1=inithe1;

/* parameter estimation */
 eta=X*newbeta+Z1*newu1;  /* systematic component */
 w1=exp(eta);
 W=diag(w1);
 WW=diag(set);
  do i=1 to cc2;
   if WW[i,i]=1 then WW[i,i:cc2]=1;
  end;
 w2=WW*w1;
  do kk=1 to cc2;
   if w2[kk]=0 then w2[kk]=10000;
  end;
a1=tt1/w2;
A=diag(a1);
M=WW`;
BB=M*A*j(cc2,1,1);
B=diag(BB);

L11=d-W*M*A*j(cc2,1,1);      /* derivative of partial likelihood */
 AA=diag(set/w2);
L12=W*B-W*M*TIES*AA*AA*M`*W; /* second derivative of likelihood */
XZ1=X`//Z1`;
XZ2=X||Z1;
 V=XZ1*L12*XZ2;   /* covariance matrix */

 L1=(d#(eta-tt1#log(w2)))`*j(cc2,1,1);   /* log-likelihood */

/* update the estimate */
 para=newbeta//newu1;
 qq=(newu1/newthe1);
 q=j(fixnum,1,0)//qq;

 RR1=I(cc1)/newthe1;
 RR2=j(fixnum,fixnum,0);
 R=block(RR2,RR1);
V=V+R;    /* covariance matrix */

L=L1
   -0.5#(cc1#log(newthe1)+(newu1`*newu1/newthe1));

if prelike>L then newpara=para+0.5#(-ginv(V)*q+ginv(V)*(XZ1*L11));
 else newpara=para-ginv(V)*q+ginv(V)*(XZ1*L11);    /* updated parameter */

 initbeta=newpara[1:fixnum,];
 initu1=newpara[fixnum+1:fixnum+cc1,];

 invV=ginv(V);
   A11=invV[1:fixnum,1:fixnum];   /* covariance matrix of beta */
   A22=invV[fixnum+1:fixnum+cc1,fixnum+1:fixnum+cc1];
```

```
   inithe1=initu1`*initu1/(cc1-trace(A22)/newthe1);

dif=abs(prelike-L);
difx=prelike-L;
paradif=newpara-para;

end;


/* S.E. of beta */
 varbeta=vecdiag(A11);
 sebeta=sqrt(varbeta);

/* S.E. of theta (REML estimate) */
 inithev1=2#(inithe1##2)/
                (cc1-2#trace(A22)/newthe1+trace(A22**2)/(inithe1##2));
 sethe1=sqrt(inithev1);

/* print */
vname={'estimate', 'se'};
estse1=initbeta||sebeta;
estse2=inithe1||sethe1;
reset noname;
print,'fixed effects',,estse1[rowname=variable colname=vname];
print,'variance of random effects (interaction)',,estse2[rowname=compo
colname=vname];

interterms={'no1001', 'no1002', 'no1003', 'no1004', 'no1005'};
print,'frailty (interaction)',,initu1[rowname=interterms];

quit;
```

A.2. Cox model stratified by trial with random treatment effects (SFE/RE) using Breslow method for handling ties

```
/* data contains variable id (observation number), sastime (time-to-event), censor
(censoring variable for time-to-event), treat (treatment indicator variable), no1001-
no1005 (indicator variables representing 5 included trials), hospno (variable representing
trial number with values 1-5) */

proc sort data=data;
 by hospno sastime descending censor;
run;

data data;
 set data;
 dif=sastime-lag(sastime);
 atrisk=(dif^=0);
run;
```

```
data data;
set data;
by hospno sastime descending censor;
if first.hospno then if atrisk=0 then atrisk=1;
run;

data data1;
 set data;
 where censor=1;
 keep id hospno sastime censor;
run;
data data2;
 set data;
 where censor=0;
 keep id hospno sastime censor;
run;

data data3;
 set data1;
 keep hospno sastime censor;
 proc transpose data=data3 out=data4;
  by hospno sastime descending censor;
  var censor;
run;
proc print;
run;
 data censor;
  set data4;
  censor=sum(of col1-col115);
  keep hospno sastime censor;
 run;

data data6;
 set censor data2;
proc sort data=data6;
 by hospno sastime;
 run;

data data7;
 set data6;
 keep hospno sastime censor;

proc transpose data=data7 out=data8;
 by hospno sastime;
 var censor;
 run;

data data9;
 set data8;
 censor=sum(of col1-col8);
 keep hospno sastime censor;
```

```
data data10;
 set data9;
 keep hospno sastime same;

data data;
 merge data data10;
 by hospno sastime;
run;

data datahosp;
set data;
run;

/* analysis of epilepsy data */
proc iml;

cc1=5;          /* number of trials */
cc2=1225;  /* total sample size */
nn1=122;        /* number in each trial */
nn2=103;
nn3=288;
nn4=246;
nn5=466;

nn1a=122;  /* cumulative number in each trial */
nn2a=225;
nn3a=513;
nn4a=759;
nn5a=1225;

use datahosp;
 read all into P;
 hosp=P[,4];    /* vector indicating trial */

 d=P[,3];                    /* vector indicating censoring (censor) */
 d1=P[1:nn1a,3];
 d2=P[nn1a+1:nn2a,3];
 d3=P[nn2a+1:nn3a,3];
 d4=P[nn3a+1:nn4a,3];
 d5=P[nn4a+1:nn5a,3];

set=P[,13];
set1=P[1:nn1a,13];  /* vector indicating number at risk (atrisk) */
set2=P[nn1a+1:nn2a,13];
set3=P[nn2a+1:nn3a,13];
set4=P[nn3a+1:nn4a,13];
set5=P[nn4a+1:nn5a,13];

/* sorting out ties in each trial */
use data6;       /* vector indicating ties */
```

```
read all into H;
t=H[,3];
nn=nrow(t);   /* number of rows of t */

*-----------------------*;
j=t[1];   /* dealing with the first observation */
 if j > 1 then
  do;
   tt0=repeat(0,j-1,1);
   tt1=j//tt0;
  end;
 else tt1=j;
*-----------------------*;
 do q=2 to nn;
  j=t[q];
   if j > 1 then
    do;
     tt0=repeat(0,j-1,1);
     tt2=j//tt0;
    end;
   else tt2=j;
  tt1=tt1//tt2;   /* vector with cc2 entries */
 end;

t1=tt1[1:nn1a];           /* trial 1 */
t2=tt1[nn1a+1:nn2a]; /* trial 2 */
 t3=tt1[nn2a+1:nn3a];
 t4=tt1[nn3a+1:nn4a];
 t5=tt1[nn4a+1:nn5a];

TIES1=diag(t1);
TIES2=diag(t2);
TIES3=diag(t3);
TIES4=diag(t4);
TIES5=diag(t5);

X=P[,1]; /* design matrix trial (fixed effects) */
X1=P[1:nn1a,1];                  /*trial 1*/
X2=P[nn1a+1:nn2a,1];          /* trial 2 */
X3=P[nn2a+1:nn3a,1];
X4=P[nn3a+1:nn4a,1];
X5=P[nn4a+1:nn5a,1];

Z1=P[,8:12]#P[,1];  /* design matrix (random effects interaction) */
Z11=P[1:nn1a,8:12]#P[1:nn1a,1];
Z12=P[nn1a+1:nn2a,8:12]#P[nn1a+1:nn2a,1];
Z13=P[nn2a+1:nn3a,8:12]#P[nn2a+1:nn3a,1];
Z14=P[nn3a+1:nn4a,8:12]#P[nn3a+1:nn4a,1];
Z15=P[nn4a+1:nn5a,8:12]#P[nn4a+1:nn5a,1];

/* fixed effects */
```

```
variable={'treat'};       /* vector of fixed effects */
fixnum=1;                 /* number of fixed effects */
initbeta=j(fixnum,1,0);  /* vector with initial values of beta */

/* random effects */
frailty={'no1001', 'no1002', 'no1003', 'no1004', 'no1005'};
initu1=j(cc1,1,0);  /* initial value of deviations */

compo={'theta1'};
inithe1=1;      /* initial value for random effects variance */

prelike=0;
iterate=0;
dif=10000;
L=0;

do k=1 to 500 until(dif<1e-6);  /* convergence criterion */
iterate=iterate+1;
prelike=L;

newbeta=initbeta;   /* change estimate */

newu1=initu1;
newthe1=inithe1;

/* parameter estimation */
eta1=X1*newbeta+Z11*newu1;  /* systematic component */
eta2=X2*newbeta+Z12*newu1;   /* systematic component */
eta3=X3*newbeta+Z13*newu1;   /* systematic component */
eta4=X4*newbeta+Z14*newu1;  /* systematic component */
eta5=X5*newbeta+Z15*newu1;   /* systematic component */

w1a=exp(eta1);
w2a=exp(eta2);
w3a=exp(eta3);
w4a=exp(eta4);
w5a=exp(eta5);

W1b=diag(w1a);
W2b=diag(w2a);
W3b=diag(w3a);
W4b=diag(w4a);
W5b=diag(w5a);

WW=diag(set);
WW1=diag(set1);
WW2=diag(set2);
WW3=diag(set3);
WW4=diag(set4);
WW5=diag(set5);
```

```
 do i=1 to nn1;
  if WW1[i,i]=1 then WW1[i,i:nn1]=1;
 end;
 w21=WW1*w1a;
  do kk=1 to nn1;
   if w21[kk]=0 then w21[kk]=10000;
  end;
a11=t1/w21;
 AA1=diag(a11);
M1=WW1`;
BB1=M1*AA1*j(nn1,1,1);
 B1=diag(BB1);
L111=d1-W1b*M1*AA1*j(nn1,1,1); /* derivative of partial likelihood */
AAA1=diag(set1/w21);
 L121=W1b*B1-W1b*M1*TIES1*AAA1*AAA1*M1`*W1b; /* second derivative */
/*****************************/
 do i=1 to nn2;
  if WW2[i,i]=1 then WW2[i,i:nn2]=1;
 end;
 w22=WW2*w2a;
  do kk=1 to nn2;
   if w22[kk]=0 then w22[kk]=10000;
  end;
a12=t2/w22;
 AA2=diag(a12);
M2=WW2`;
BB2=M2*AA2*j(nn2,1,1);
 B2=diag(BB2);
L112=d2-W2b*M2*AA2*j(nn2,1,1); /* derivative of partial likelihood */
  AAA2=diag(set2/w22);
 L122=W2b*B2-W2b*M2*TIES2*AAA2*AAA2*M2`*W2b; /* second derivative */
/*****************************/
 do i=1 to nn3;
  if WW3[i,i]=1 then WW3[i,i:nn3]=1;
 end;
 w23=WW3*w3a;
  do kk=1 to nn3;
   if w23[kk]=0 then w23[kk]=10000;
  end;
a13=t3/w23;
 AA3=diag(a13);
M3=WW3`;
BB3=M3*AA3*j(nn3,1,1);
 B3=diag(BB3);
L113=d3-W3b*M3*AA3*j(nn3,1,1); /* derivative of partial likelihood */
AAA3=diag(set3/w23);
 L123=W3b*B3-W3b*M3*TIES3*AAA3*AAA3*M3`*W3b; /* second derivative */
/*****************************/
 do i=1 to nn4;
  if WW4[i,i]=1 then WW4[i,i:nn4]=1;
 end;
```

```
 w24=WW4*w4a;
  do kk=1 to nn4;
   if w24[kk]=0 then w24[kk]=10000;
  end;
a14=t4/w24;
 AA4=diag(a14);
M4=WW4`;
BB4=M4*AA4*j(nn4,1,1);
 B4=diag(BB4);
L114=d4-W4b*M4*AA4*j(nn4,1,1); /* derivative of partial likelihood */
AAA4=diag(set4/w24);
L124=W4b*B4-W4b*M4*TIES4*AAA4*AAA4*M4`*W4b; /* second derivative */
/*****************************/
 do i=1 to nn5;
  if WW5[i,i]=1 then WW5[i,i:nn5]=1;
 end;
 w25=WW5*w5a;
  do kk=1 to nn5;
   if w25[kk]=0 then w25[kk]=10000;
  end;
a15=t5/w25;
 AA5=diag(a15);
M5=WW5`;
BB5=M5*AA5*j(nn5,1,1);
 B5=diag(BB5);
 L115=d5-W5b*M5*AA5*j(nn5,1,1);/* derivative of partial likelihood */
AAA5=diag(set5/w25);
 L125=W5b*B5-W5b*M5*TIES5*AAA5*AAA5*M5`*W5b; /* second derivative */
/*****************************/

L11=L111//L112//L113//L114//L115;
L12=block(L121,L122,L123,L124,L125);

XZ1=X`//Z1`;
XZ2=X||Z1;
V=XZ1*L12*XZ2;   /* covariance matrix */

L1=(d1#(eta1-t1#log(w21)))`*j(nn1,1,1)+(d2#(eta2-
t2#log(w22)))`*j(nn2,1,1)+(d3#(eta3-t3#log(w23)))`*j(nn3,1,1)+(d4#(eta4-
t4#log(w24)))`*j(nn4,1,1)
+(d5#(eta5-t5#log(w25)))`*j(nn5,1,1); /* log-likelihood */

/* update the estimate */
para=newbeta//newu1;
qq=(newu1/newthe1);
q=j(fixnum,1,0)//qq;

RR1=I(cc1)/newthe1;
RR2=j(fixnum,fixnum,0);
R=block(RR2,RR1);
```

```
V=V+R;   /* covariance matrix */

L=L1-0.5#(cc1#log(newthe1)+(newu1`*newu1/newthe1));

if prelike>L then newpara=para+0.5#(-ginv(V)*q+ginv(V)*(XZ1*L11));
 else newpara=para-ginv(V)*q+ginv(V)*(XZ1*L11);

initbeta=newpara[1:fixnum,];
initu1=newpara[fixnum+1:fixnum+cc1,];

invV=ginv(V);
   A11=invV[1:fixnum,1:fixnum];   /* covariance matrix of beta */
   A22=invV[fixnum+1:fixnum+cc1,fixnum+1:fixnum+cc1];
inithe1=initu1`*initu1/(cc1-(trace(A22)/newthe1));

dif=abs(prelike-L);
difx=prelike-L;
paradif=newpara-para;

end;

/* S.E. of beta */
 varbeta=vecdiag(A11);
 sebeta=sqrt(varbeta);

/* S.E. of theta (REML estimate) */
inithev1=2#(inithe1##2)/
              (cc1-2#trace(A22)/newthe1+trace(A22**2)/(inithe1##2));
 sethe1=sqrt(inithev1);

/* print */
vname={'estimate', 'se'};
estse1=initbeta||sebeta;
estse2=inithe1||sethe1;
reset noname;
print,'fixed effects',,estse1[rowname=variable colname=vname];
print,'variance of random effects (interaction)',,estse2[rowname=compo
colname=vname];

interterms={'no1001', 'no1002', 'no1003', 'no1004', 'no1005'};
print,'frailty (interaction)',,initu1[rowname=interterms];

quit;
```

A.3. Cox model with random trial effects and random treatment effects (RE/RE) using Breslow method for handling ties

This program is similar to the original program supplied by Takuhiro Yamaguchi which has been adapted and extended in this thesis to fit other models of interest in meta-analysis of individual patient time-to-event data (Appendix A1, A2, B1, B2, B3).

```
/* data sorting steps are same as Appendix A..1 up to POINT X */

/* Program for analysis of epilepsy data */

proc iml;

cc1=5;          /* number of trials */
cc2=1225;       /* total sample size */

use data;
 read all into P;

 d=P[,3];       /* vector indicating censoring (censor) */

set=P[,12];     /* vector indicating number at risk (atrisk) */

use data6;      /* vector indicating ties */
read all into H;
t=H[,3];
 nn=nrow(t);
*----------------------*;
j=t[1];
 if j > 1 then
  do;
   tt0=repeat(0,j-1,1);
   tt1=j//tt0;
  end;
 else tt1=j;
*----------------------*;
 do q=2 to nn;
  j=t[q];
   if j > 1 then
    do;
     tt0=repeat(0,j-1,1);
     tt2=j//tt0;
    end;
   else tt2=j;
  tt1=tt1//tt2;
 end;
TIES=diag(tt1);
```

```
X=P[,1]||P[,4:5];  /* design matrix (fixed effects) */

Z0=P[,7:11];          /* design matrix (random effects baseline) */
Z1=Z0#P[,1];          /* design matrix (random effects interaction) */


/* fixed effects */
variable={'treat'};              /* vector of fixed effects */
fixnum=1;                        /* number of fixed effects */
initbeta=j(fixnum,1,0);          /* vector of initial values */

/* random effects */
 frailty={'no1001', 'no1002', 'no1003', 'no1004', 'no1005'};  /* vector of frailty */
 initu0=j(cc1,1,0);  /* initial value of variance for baseline */
 initu1=j(cc1,1,0);  /* initial value of variance for interaction */

 compo={'theta0','theta1'};  /* variance component of frailty */
 inithe0=1;  /* baseline */
 inithe1=1;  /* interaction */

prelike=0;
iterate=0;
dif=10000;
L=0;

do k=1 to 500 until(dif<1e-6);  /* convergence criterion */
iterate=iterate+1;
prelike=L;

newbeta=initbeta;
newu0=initu0;
newu1=initu1;
newthe0=inithe0;
newthe1=inithe1;

/* parameter estimation */
 eta=X*newbeta+Z0*newu0+Z1*newu1;  /* systematic component */

 w1=exp(eta);
W=diag(w1);

WW=diag(set);
 do i=1 to cc2;
  if WW[i,i]=1 then WW[i,i:cc2]=1;
 end;
w2=WW*w1;
 do kk=1 to cc2;
  if w2[kk]=0 then w2[kk]=10000;
 end;
a1=tt1/w2;
```

```
A=diag(a1);

M=WW`;
BB=M*A*j(cc2,1,1);
B=diag(BB);

L11=d-W*M*A*j(cc2,1,1);    /* derivative of partial likelihood */
  AA=diag(set/w2);
L12=W*B-W*M*TIES*AA*AA*M`*W;  /* second derivative */

XZ1=X`//Z0`//Z1`;
 XZ2=X||Z0||Z1;
 V=XZ1*L12*XZ2;              /* covariance matrix */

 L1=(d#(eta-tt1#log(w2)))`*j(cc2,1,1);   /* log-likelihood */

para=newbeta//newu0//newu1;
 qq=(newu0/newthe0)//(newu1/newthe1);
 q=j(fixnum,1,0)//qq;

 RR0=I(cc1)/newthe0;
 RR1=I(cc1)/newthe1;
 RR2=j(fixnum,fixnum,0);
 R=block(RR2,RR0,RR1);
V=V+R;    /* covariance matrix */

L=L1 - 0.5
#(cc1#log(newthe0#newthe1)+(newu0`*newu0/newthe0+newu1`*newu1/newthe1));

if prelike>L then newpara=para+0.5#(-ginv(V)*q+ginv(V)*(XZ1*L11));
 else newpara=para-ginv(V)*q+ginv(V)*(XZ1*L11);    /* updated parameter */

 initbeta=newpara[1:fixnum,];
 initu0=newpara[fixnum+1:fixnum+cc1,];
 initu1=newpara[fixnum+1+cc1:fixnum+cc1#2,];

  invV=ginv(V);
    A11=invV[1:fixnum,1:fixnum];   /* covariance matrix of beta */
    A22=invV[fixnum+1:fixnum+cc1,fixnum+1:fixnum+cc1];
    A33=invV[fixnum+1+cc1:fixnum+cc1#2,fixnum+1+cc1:fixnum+cc1#2];
 inithe0=initu0`*initu0/(cc1-trace(A22)/newthe0);
  inithe1=initu1`*initu1/(cc1-trace(A33)/newthe1);

dif=abs(prelike-L);
difx=prelike-L;
paradif=newpara-para;

end;

/* S.E. of beta */
 varbeta=vecdiag(A11);
```

```
sebeta=sqrt(varbeta);

/* S.E. of theta (REML estimate) */
inithev0=2#(inithe0##2)/
            (cc1-2#trace(A22)/newthe0+trace(A22**2)/(inithe0##2));
sethe0=sqrt(inithev0);
inithev1=2#(inithe1##2)/
            (cc1-2#trace(A33)/newthe1+trace(A33**2)/(inithe1##2));
sethe1=sqrt(inithev1);

/* print */
vname={'estimate', 'se'};
estse1=initbeta||sebeta;
estse2=(inithe0||sethe0)//(inithe1||sethe1);
reset noname;
print,'fixed effects',,estse1[rowname=variable colname=vname];
print,'variance of random effects (baseline, interaction)',,estse2[rowname=compo
colname=vname];
print,'frailty (baseline)',,initu0[rowname=frailty];
print,'frailty (interaction)',,initu1[rowname=frailty];

quit;
```

APPENDIX B

SAS programs for fitting random effect Cox models using the Efron method for handling ties

B.1. Cox model with fixed trial indicator variables and random treatment effects using Efron method for handling ties

```
/* data contains variable id (observation number), sastime (time-to-event), censor
(censoring variable for time-to-event), treat (treatment indicator variable), no1001-
no1005 (indicator variables representing 5 included trials) */

proc sort data=data;
 by sastime descending censor;
run;

data data;
 set data;
 dif=sastime-lag(sastime);
 atrisk=(dif^=0);
run;

data data1;
 set data;
 where censor=1;
 keep id sastime censor;
run;
data data2;
 set data;
 where censor=0;
```

```
 keep id sastime censor;
run;

data data3;
 set data1;
 keep sastime censor;
proc transpose data=data3 out=data4;
 by sastime;
 var censor;

 data censor;
  set data4;
  censor=sum(of col1-col357);
  keep sastime censor;
 run;

data data6;
 set censor data2;
proc sort data=data6;
 by sastime;
data data7;
 set data6;
 keep sastime censor;
proc transpose data=data7 out=data8;
 by sastime;
 var censor;
data data9;
 set data8;
 censor=sum(of col1-col13);
 keep sastime censor;

data data10;
 set data9;
 keep sastime same;
data data;
 merge data data10;
 by sastime;
run;
```

POINT X

```
/* Program for analysis of epilepsy data */

proc iml;

cc1=5;          /* number of trials */
cc2=1225;   /* total sample size */

use data;
 read all into P;
 d=P[,3];                    /* vector indicating censoring (censor) */
```

```
set=P[,12];      /* vector indicating number at risk (atrisk) */

use data6;       /* vector indicating ties */
read all into H;
t=H[,3];
 nn=nrow(t);
*----------------------*;
 j=t[1];
 if j > 1 then
  do;
   tt0=repeat(0,j-1,1);
    tt1=j//tt0;
  end;
 else tt1=j;
*----------------------*;
 do q=2 to nn;
  j=t[q];
   if j > 1 then
    do;
     tt0=repeat(0,j-1,1);
     tt2=j//tt0;
    end;
   else tt2=j;
  tt1=tt1//tt2;
 end;

/* creating vector for efron */
*----------------------*;
 j=t[1];
 if j > 1 then
  do;
     efron0=1;
        pos0=0;
                    pos0a=0;
      do k=1 to j-1;
     efron2=(j-k)/j;
        pos2=k;
                    pos2a=j-(k+1);
     efron0=efron0//efron2;
        pos0=pos0//pos2;
                    pos0a=pos0a//pos2a;
      end;

  end;
 else do;
   efron0=j;
      pos0=0;
                  pos0a=0;
 end;
*----------------------*;
 do q=2 to nn;
```

```
j=t[q];
 if j > 1 then
  do;
   efron2=1;
   pos2=0;
         pos2a=0;
   do k=1 to j-1;
     efron1=(j-k)/j;
           pos1=k;
           pos1a=j-(k+1);
     efron2=efron2//efron1;
            pos2=pos2//pos1;
           pos2a=pos2a//pos1a;
    end;
   end;
  else do;
   efron2=j;
      pos2=0;
      pos2a=0;
     end;
 efron0=efron0//efron2;
 pos0=pos0//pos2;
 pos0a=pos0a//pos2a;
end;

X=P[,1]||P[,8:11];   /* design matrix (fixed effects) */

Z1=P[,7:11]#P[,1];


/* fixed effects */
 variable={'treat','no1002','no1003','no1004','no1005'};
 fixnum=5;   /* number of fixed effects */
 initbeta=j(fixnum,1,0);  /* vector of initial values */

 initu1=j(cc1,1,0);  /* initial value of variance for interaction */

 compo={'theta1'};   /* variance component of frailty */
 inithe1=1;                /* interaction */

prelike=0;
iterate=0;
dif=10000;
L=0;

do k=1 to 500 until(dif<1e-6);  /* convergence criterion */
iterate=iterate+1;
prelike=L;

newbeta=initbeta;   /* change estimate */
newu1=initu1;
```

```
newthe1=inithe1;

/* parameter estimation */
 eta=X*newbeta+Z1*newu1;  /* systematic component */

 w1=exp(eta);
 W=diag(w1);
 WW1=diag(efron0);

  do i=1 to cc2;
   if WW1[i,i]=1 then WW1[i,i:cc2]=1;
  if WW1[i,i]<1 then if WW1[i,i]>0 then do;
     val=WW1[i,i];
        s=pos0[i];
                    sa=pos0a[i];
                    if i ^= cc2 then do;
                   WW1[i,i-s:i+sa]=val;
                            if i+sa ^= cc2 then do;
                                    WW1[i,i+sa+1:cc2]=1;
                            end;
                    end;
                    if i=cc2 then do;
                     WW1[i,i-s:i+sa]=val;
            end;
 end;
end;

setefs=j(cc2,1,.);
setefs1=j(cc2,1,.);
 do i=1 to cc2;
   setef=set[i];
   setef1=set[i];
   if efron0[i]<1 then if efron0[i]>0 then do;
     setef=efron0[i];
        setef1=1;
   end;
   setefs[i]=setef;
   setefs1[i]=setef1;
 end;

WW2s=diag(setefs);
 do i=1 to cc2;
 if WW2s[i,i]=1 then WW2s[i,i:cc2]=1;
 if WW2s[i,i]<1 then if WW2s[i,i]>0 then do;
    val=WW2s[i,i];
       s=pos0[i];
                    sa=pos0a[i];
                    if i ^= cc2 then do;
                   WW2s[i,i-s:i+sa]=val;
                            if i+sa ^= cc2 then do;
                                    WW2s[i,i+sa+1:cc2]=1;
```

```
                    end;

                end;
                if i=cc2 then do;
                  WW2s[i,i-s:i+sa]=val;
                end;
  end;
 end;

M2s=WW2s`;

 w2a=WW2s*w1;
  do kk=1 to cc2;
   if w2a[kk]=0 then w2a[kk]=10000;
  end;

a1a=setefs1/w2a;
As=diag(a1a);
AAs=diag(d/w2a);
M1=WW1`;

 BB1=M1*AAs*j(cc2,1,1);
 B1=diag(BB1);

 L11=d-W*M1*As*j(cc2,1,1);    /* derivative of partial likelihood */

 TIES2=diag(d);

 L12=W*B1-W*M1*TIES2*AAs*AAs*M1`*W;   /* second derivative */

XZ1=X`//Z1`;
XZ2=X||Z1;
V=XZ1*L12*XZ2;

L1=(d#(eta-log(w2a)))`*j(cc2,1,1);   /* log-likelihood */

/* update the estimate */
 para=newbeta//newu1;
 qq=(newu1/newthe1);
 q=j(fixnum,1,0)//qq;

 RR1=I(cc1)/newthe1;
 RR2=j(fixnum,fixnum,0);
 R=block(RR2,RR1);
V=V+R;    /* covariance matrix */

L=L1 - 0.5 #(cc1#log(newthe1)+(newu1`*newu1/newthe1));

if prelike>L then newpara=para+0.5#(-ginv(V)*q+ginv(V)*(XZ1*L11));
 else newpara=para-ginv(V)*q+ginv(V)*(XZ1*L11);    /* updated parameter */
```

```
initbeta=newpara[1:fixnum,];
initu1=newpara[fixnum+1:fixnum+cc1,];

invV=ginv(V);
   A11=invV[1:fixnum,1:fixnum];   /* covariance matrix of beta */
   A22=invV[fixnum+1:fixnum+cc1,fixnum+1:fixnum+cc1];
   inithe1=initu1`*initu1/(cc1-trace(A22)/newthe1);


dif=abs(prelike-L);
difx=prelike-L;
paradif=newpara-para;

end;


/* S.E. of beta */
 varbeta=vecdiag(A11);
 sebeta=sqrt(varbeta);

/* S.E. of theta (REML estimate) */
inithev1=2#(inithe1##2)/
                (cc1-2#trace(A22)/newthe1+trace(A22**2)/(inithe1##2));
 sethe1=sqrt(inithev1);

vname={'estimate', 'se'};
estse1=initbeta||sebeta;
estse2=inithe1||sethe1;
reset noname;
print,'fixed effects',,estse1[rowname=variable colname=vname];
print,'variance of random effects (interaction)',,estse2[rowname=compo
colname=vname];

interterms={'no1001', 'no1002', 'no1003', 'no1004', 'no1005'};
print,'frailty (interaction)',,initu1[rowname=interterms];

quit;
```

B.2. Cox model stratified by trial with random treatment effects using Efron method for handling ties

```
/* data contains variable id (observation number), sastime (time-to-event), censor
(censoring variable for time-to-event), treat (treatment indicator variable), no1001-
no1005 (indicator variables representing 5 included trials), hospno (variable indicating
trial number 1-5) */

proc sort data=data;
 by hospno sastime descending censor;
run;
```

```
data data;
 set data;
 dif=sastime-lag(sastime);
 atrisk=(dif^=0);
keep id sastime censor treat hospno
    eptype sex
    no1001 no1002 no1003 no1004 no1005
    atrisk;
run;
data data;
set data;
by hospno sastime descending censor;
if first.hospno then if atrisk=0 then atrisk=1;
run;

data data1;
 set data;
 where censor=1;
 keep id hospno sastime censor;
run;

data data2;
 set data;
 where censor=0;
 keep id hospno sastime censor;
run;

 data data3;
  set data1;
  keep hospno sastime censor;
 proc transpose data=data3 out=data4;
  by hospno sastime descending censor;
  var censor;
run;
 data censor;
  set data4;
  censor=sum(of col1-col115);
  keep hospno sastime censor;
 run;
proc print data=data4;
run;

data data6;
 set censor data2;
proc sort data=data6;
 by hospno sastime;
 run;

data data7;
```

```
 set data6;
 keep hospno sastime censor;
proc transpose data=data7 out=data8;
 by hospno sastime;
 var censor;
 run;

data data9;
 set data8;
 censor=sum(of col1-col8);
 keep hospno sastime censor;
data data10;
 set data9;
 keep hospno sastime same;

data data;
 merge data data10;
 by hospno sastime;
run;
data datahosp;
set data;
run;

/* analysis of epilepsy data */
proc iml;

cc1=5;          /* number of trials */
cc2=1225;       /* total sample size */
nn1=122;        /* number in each trial */
nn2=103;
nn3=288;
nn4=246;
nn5=466;

nn1a=122;       /* cumulative number in each trial */
nn2a=225;
nn3a=513;
nn4a=759;
nn5a=1225;

use datahosp;
read all into P;
 hosp=P[,4];    /* vector indicating trial */

d=P[,3];          /* vector indicating censoring (censor) */
d1=P[1:nn1a,3];
d2=P[nn1a+1:nn2a,3];
d3=P[nn2a+1:nn3a,3];
d4=P[nn3a+1:nn4a,3];
d5=P[nn4a+1:nn5a,3];
```

```
set=P[,13];      /* vector indicating number at risk (atrisk) */
set1=P[1:nn1a,13];
set2=P[nn1a+1:nn2a,13];
set3=P[nn2a+1:nn3a,13];
set4=P[nn3a+1:nn4a,13];
set5=P[nn4a+1:nn5a,13];

/* sorting out ties in each trial */
use data6;       /* vector indicating ties */
read all into H;
t=H[,3];
nn=nrow(t);   /* number of rows of t */

*-----------------------*;
j=t[1];   /* first observation */
 if j > 1 then
  do;
   tt0=repeat(0,j-1,1);
   tt1=j//tt0;
  end;
 else tt1=j;
*-----------------------*;
 do q=2 to nn;
  j=t[q];
   if j > 1 then
    do;
     tt0=repeat(0,j-1,1);
     tt2=j//tt0;
    end;
   else tt2=j;
  tt1=tt1//tt2;   /* vector with cc2 entries */
 end;

/* creating vector for efron */
*-----------------------*;
 j=t[1];
 if j > 1 then
  do;
    efron0=1;
       pos0=0;
            pos0a=0;
   do k=1 to j-1;
  efron2=(j-k)/j;
     pos2=k;
           pos2a=j-(k+1);
  efron0=efron0//efron2;
     pos0=pos0//pos2;
           pos0a=pos0a//pos2a;
   end;

end;
```

```
 else do;
  efron0=j;
     pos0=0;
           pos0a=0;
 end;
*-----------------------*;
 do q=2 to nn;
  j=t[q];
   if j > 1 then
    do;
     efron2=1;
  pos2=0;
     pos2a=0;
 do k=1 to j-1;
   efron1=(j-k)/j;
        pos1=k;
                  pos1a=j-(k+1);
   efron2=efron2//efron1;
        pos2=pos2//pos1;
                  pos2a=pos2a//pos1a;
  end;
 end;
else do;
 efron2=j;
    pos2=0;
          pos2a=0;
    end;
 efron0=efron0//efron2;
 pos0=pos0//pos2;
 pos0a=pos0a//pos2a;
end;

efron01=efron0[1:nn1a];              /* trial 1 */
efron02=efron0[nn1a+1:nn2a];         /* trial 2 */
efron03=efron0[nn2a+1:nn3a];
efron04=efron0[nn3a+1:nn4a];
efron05=efron0[nn4a+1:nn5a];

pos01=pos0[1:nn1a];                  /* trial 1 */
pos02=pos0[nn1a+1:nn2a];             /* trial 2 */
pos03=pos0[nn2a+1:nn3a];
pos04=pos0[nn3a+1:nn4a];
pos05=pos0[nn4a+1:nn5a];

pos0a1=pos0a[1:nn1a];                /* trial 1 */
pos0a2=pos0a[nn1a+1:nn2a];           /* trial 2 */
pos0a3=pos0a[nn2a+1:nn3a];
pos0a4=pos0a[nn3a+1:nn4a];
pos0a5=pos0a[nn4a+1:nn5a];

X=P[,1];        /* design matrix (fixed effects) */
```

```
X1=P[1:nn1a,1];                    /* trial 1*/
X2=P[nn1a+1:nn2a,1];               /* trial 2 */
X3=P[nn2a+1:nn3a,1];
X4=P[nn3a+1:nn4a,1];
X5=P[nn4a+1:nn5a,1];

 Z1=P[,8:12]#P[,1];     /* design matrix (random effects interaction) */
Z11=P[1:nn1a,8:12]#P[1:nn1a,1];
Z12=P[nn1a+1:nn2a,8:12]#P[nn1a+1:nn2a,1];
Z13=P[nn2a+1:nn3a,8:12]#P[nn2a+1:nn3a,1];
Z14=P[nn3a+1:nn4a,8:12]#P[nn3a+1:nn4a,1];
Z15=P[nn4a+1:nn5a,8:12]#P[nn4a+1:nn5a,1];

/* fixed effects */
 variable={'treat'};               /* vector of fixed effects */
 fixnum=1;                         /* number of fixed effects */
 initbeta=j(fixnum,1,0);           /* vector of initial values */

/* random effects */
 frailty={'no1001', 'no1002', 'no1003', 'no1004', 'no1005'};
 initu1=j(cc1,1,0);                /* initial value of variance for interaction */

 compo={'theta1'};
 inithe1=1;                        /* interaction */

prelike=0;
iterate=0;
dif=10000;
L=0;

do k=1 to 500 until(dif<1e-6);  /* convergence criterion */
iterate=iterate+1;
prelike=L;

newbeta=initbeta;
newu1=initu1;
newthe1=inithe1;

/* parameter estimation */
 eta1=X1*newbeta+Z11*newu1;
eta2=X2*newbeta+Z12*newu1;
eta3=X3*newbeta+Z13*newu1;
eta4=X4*newbeta+Z14*newu1;
eta5=X5*newbeta+Z15*newu1;

w1a=exp(eta1);
w2a=exp(eta2);
w3a=exp(eta3);
w4a=exp(eta4);
w5a=exp(eta5);
```

```
W1b=diag(w1a);
W2b=diag(w2a);
W3b=diag(w3a);
W4b=diag(w4a);
W5b=diag(w5a);

WW=diag(set);
WW1=diag(set1);
WW2=diag(set2);
WW3=diag(set3);
WW4=diag(set4);
WW5=diag(set5);

/* trial 1 */
WW11=diag(efron01);

 do i=1 to nn1;
  if WW11[i,i]=1 then WW11[i,i:nn1]=1;
  if WW11[i,i]<1 then if WW11[i,i]>0 then do;
     val=WW11[i,i];
        s=pos01[i];
              sa=pos0a1[i];
              if i ^= nn1 then do;
            WW11[i,i-s:i+sa]=val;
                  if i+sa ^= nn1 then do;
                      WW11[i,i+sa+1:nn1]=1;
                  end;
              end;
              if i=nn1 then do;
               WW11[i,i-s:i+sa]=val;
         end;
   end;
  end;

setefs=j(nn1,1,.);
setefs1=j(nn1,1,.);
 do i=1 to nn1;
   setef=set1[i];
   setef1=set1[i];
   if efron01[i]<1 then if efron01[i]>0 then do;
     setef=efron01[i];
        setef1=1;
  end;
  setefs[i]=setef;
  setefs1[i]=setef1;
end;

 WW2s1=diag(setefs);
 do i=1 to nn1;
 if WW2s1[i,i]=1 then WW2s1[i,i:nn1]=1;
 if WW2s1[i,i]<1 then if WW2s1[i,i]>0 then do;
```

```
val=WW2s1[i,i];
   s=pos01[i];
        sa=pos0a1[i];
        if i ^= nn1 then do;
       WW2s1[i,i-s:i+sa]=val;
            if i+sa ^= nn1 then do;
                WW2s1[i,i+sa+1:nn1]=1;
            end;

        end;
        if i=nn1 then do;
             WW2s1[i,i-s:i+sa]=val;
           end;
  end;
 end;

M2s1=WW2s1`;

 w2a1=WW2s1*w1a;
  do kk=1 to nn1;
   if w2a1[kk]=0 then w2a1[kk]=10000;
  end;

a1a=setefs1/w2a1;
As1=diag(a1a);
AAs1=diag(d1/w2a1);
M11=WW11`;

 BB11=M11*AAs1*j(nn1,1,1);
 B11=diag(BB11);

  L111=d1-W1b*M11*As1*j(nn1,1,1);    /* derivative of partial likelihood */

  TIES21=diag(d1);

 L121=W1b*B11-W1b*M11*TIES21*AAs1*AAs1*M11`*W1b;  /* second derivative
*/

/* trial 2 */
 WW12=diag(efron02);

do i=1 to nn2;
 if WW12[i,i]=1 then WW12[i,i:nn2]=1;
 if WW12[i,i]<1 then if WW12[i,i]>0 then do;
    val=WW12[i,i];
       s=pos02[i];
            sa=pos0a2[i];
            if i ^= nn2 then do;
           WW12[i,i-s:i+sa]=val;
                if i+sa ^= nn2 then do;
                    WW12[i,i+sa+1:nn2]=1;
```

```
                end;
            end;
            if i=nn2 then do;
             WW12[i,i-s:i+sa]=val;
        end;
  end;
 end;

setefs=j(nn2,1,.);
setefs1=j(nn2,1,.);
 do i=1 to nn2;
  setef=set2[i];
  setef1=set2[i];
  if efron02[i]<1 then if efron02[i]>0 then do;
    setef=efron02[i];
       setef1=1;
  end;
  setefs[i]=setef;
  setefs1[i]=setef1;
end;

 WW2s2=diag(setefs);
  do i=1 to nn2;
  if WW2s2[i,i]=1 then WW2s2[i,i:nn2]=1;
  if WW2s2[i,i]<1 then if WW2s2[i,i]>0 then do;
     val=WW2s2[i,i];
        s=pos02[i];
             sa=pos0a2[i];
             if i ^= nn2 then do;
            WW2s2[i,i-s:i+sa]=val;
                  if i+sa ^= nn2 then do;
                       WW2s2[i,i+sa+1:nn2]=1;
                  end;

             end;
             if i=nn2 then do;
              WW2s2[i,i-s:i+sa]=val;
             end;
 end;
end;

M2s2=WW2s2`;

 w2a2=WW2s2*w2a;
  do kk=1 to nn2;
   if w2a2[kk]=0 then w2a2[kk]=10000;
  end;

a1a=setefs1/w2a2;
As2=diag(a1a);
AAs2=diag(d2/w2a2);
```

```
M12=WW12`;

 BB12=M12*AAs2*j(nn2,1,1);
 B12=diag(BB12);

L112=d2-W2b*M12*As2*j(nn2,1,1);    /* derivative of partial likelihood */

  TIES22=diag(d2);

L122=W2b*B12-W2b*M12*TIES22*AAs2*AAs2*M12`*W2b;   /* second derivative
*/

/* trial 3 */
WW13=diag(efron03);

 do i=1 to nn3;
  if WW13[i,i]=1 then WW13[i,i:nn3]=1;
  if WW13[i,i]<1 then if WW13[i,i]>0 then do;
     val=WW13[i,i];
        s=pos03[i];
              sa=pos0a3[i];
              if i ^= nn3 then do;
             WW13[i,i-s:i+sa]=val;
                    if i+sa ^= nn3 then do;
                         WW13[i,i+sa+1:nn3]=1;
                    end;
                    end;
                    if i=nn3 then do;
                     WW13[i,i-s:i+sa]=val;
           end;
    end;
   end;

setefs=j(nn3,1,.);
setefs1=j(nn3,1,.);
 do i=1 to nn3;
   setef=set3[i];
   setef1=set3[i];
   if efron03[i]<1 then if efron03[i]>0 then do;
     setef=efron03[i];
        setef1=1;
  end;
  setefs[i]=setef;
  setefs1[i]=setef1;
end;

 WW2s3=diag(setefs);
  do i=1 to nn3;
  if WW2s3[i,i]=1 then WW2s3[i,i:nn3]=1;
  if WW2s3[i,i]<1 then if WW2s3[i,i]>0 then do;
     val=WW2s3[i,i];
```

```
s=pos03[i];
      sa=pos0a3[i];
      if i ^= nn3 then do;
    WW2s3[i,i-s:i+sa]=val;
          if i+sa ^= nn3 then do;
              WW2s3[i,i+sa+1:nn3]=1;
          end;

      end;
      if i=nn3 then do;
       WW2s3[i,i-s:i+sa]=val;
                end;
   end;
  end;

M2s3=WW2s3`;

 w2a3=WW2s3*w3a;
  do kk=1 to nn3;
   if w2a3[kk]=0 then w2a3[kk]=10000;
  end;

a1a=setefs1/w2a3;
As3=diag(a1a);
AAs3=diag(d3/w2a3);
M13=WW13`;

 BB13=M13*AAs3*j(nn3,1,1);
 B13=diag(BB13);

L113=d3-W3b*M13*As3*j(nn3,1,1);    /* derivative of partial likelihood */

TIES23=diag(d3);

L123=W3b*B13-W3b*M13*TIES23*AAs3*AAs3*M13`*W3b;   /* second derivative
*/

/* trial 4 */
 WW14=diag(efron04);

do i=1 to nn4;
 if WW14[i,i]=1 then WW14[i,i:nn4]=1;
 if WW14[i,i]<1 then if WW14[i,i]>0 then do;
    val=WW14[i,i];
       s=pos04[i];
            sa=pos0a4[i];
            if i ^= nn4 then do;
          WW14[i,i-s:i+sa]=val;
                if i+sa ^= nn4 then do;
                    WW14[i,i+sa+1:nn4]=1;
                end;
```

```
                end;
                if i=nn4 then do;
                  WW14[i,i-s:i+sa]=val;
          end;
   end;
  end;

setefs=j(nn4,1,.);
setefs1=j(nn4,1,.);
 do i=1 to nn4;
  setef=set4[i];
  setef1=set4[i];
  if efron04[i]<1 then if efron04[i]>0 then do;
    setef=efron04[i];
       setef1=1;
  end;
  setefs[i]=setef;
  setefs1[i]=setef1;
end;

 WW2s4=diag(setefs);
  do i=1 to nn4;
  if WW2s4[i,i]=1 then WW2s4[i,i:nn4]=1;
  if WW2s4[i,i]<1 then if WW2s4[i,i]>0 then do;
     val=WW2s4[i,i];
        s=pos04[i];
        sa=pos0a4[i];
        if i ^= nn4 then do;
      WW2s4[i,i-s:i+sa]=val;
            if i+sa ^= nn4 then do;
                WW2s4[i,i+sa+1:nn4]=1;
            end;

        end;
        if i=nn4 then do;
          WW2s4[i,i-s:i+sa]=val;
        end;
 end;
end;

M2s4=WW2s4`;

 w2a4=WW2s4*w4a;
  do kk=1 to nn4;
   if w2a4[kk]=0 then w2a4[kk]=10000;
  end;

a1a=setefs1/w2a4;
As4=diag(a1a);
AAs4=diag(d4/w2a4);
M14=WW14`;
```

```
BB14=M14*AAs4*j(nn4,1,1);
B14=diag(BB14);

L114=d4-W4b*M14*As4*j(nn4,1,1);    /* derivative of partial likelihood */

TIES24=diag(d4);

L124=W4b*B14-W4b*M14*TIES24*AAs4*AAs4*M14`*W4b;   /* second derivative
*/

/* trial 5 */
 WW15=diag(efron05);

 do i=1 to nn5;
  if WW15[i,i]=1 then WW15[i,i:nn5]=1;
  if WW15[i,i]<1 then if WW15[i,i]>0 then do;
     val=WW15[i,i];
        s=pos05[i];
              sa=pos0a5[i];
              if i ^= nn5 then do;
             WW15[i,i-s:i+sa]=val;
                    if i+sa ^= nn5 then do;
                         WW15[i,i+sa+1:nn5]=1;
                    end;
               end;
               if i=nn5 then do;
                WW15[i,i-s:i+sa]=val;
          end;
   end;
  end;

setefs=j(nn5,1,.);
setefs1=j(nn5,1,.);
 do i=1 to nn5;
   setef=set5[i];
   setef1=set5[i];
   if efron05[i]<1 then if efron05[i]>0 then do;
     setef=efron05[i];
        setef1=1;
   end;
   setefs[i]=setef;
   setefs1[i]=setef1;
end;

 WW2s5=diag(setefs);
  do i=1 to nn5;
  if WW2s5[i,i]=1 then WW2s5[i,i:nn5]=1;
  if WW2s5[i,i]<1 then if WW2s5[i,i]>0 then do;
     val=WW2s5[i,i];
        s=pos05[i];
```

```
  sa=pos0a5[i];
  if i ^= nn5 then do;
WW2s5[i,i-s:i+sa]=val;
      if i+sa ^= nn5 then do;
          WW2s5[i,i+sa+1:nn5]=1;
      end;

  end;
  if i=nn5 then do;
   WW2s5[i,i-s:i+sa]=val;
  end;
   end;
  end;

M2s5=WW2s5`;

 w2a5=WW2s5*w5a;
  do kk=1 to nn5;
   if w2a5[kk]=0 then w2a5[kk]=10000;
  end;

a1a=setefs1/w2a5;
As5=diag(a1a);
AAs5=diag(d5/w2a5);
M15=WW15`;

 BB15=M15*AAs5*j(nn5,1,1);
 B15=diag(BB15);

L115=d5-W5b*M15*As5*j(nn5,1,1);    /* derivative of partial likelihood */

TIES25=diag(d5);

L125=W5b*B15-W5b*M15*TIES25*AAs5*AAs5*M15`*W5b;   /* second derivative
*/

L11=L111//L112//L113//L114//L115;
L12=block(L121,L122,L123,L124,L125);

XZ1=X`//Z1`;
XZ2=X||Z1;
V=XZ1*L12*XZ2;   /* covariance matrix */

/* log-likelihood */
 L1=(d1#(eta1-log(w2a1)))`*j(nn1,1,1)+(d2#(eta2-log(w2a2)))`*j(nn2,1,1)
+(d3#(eta3-log(w2a3)))`*j(nn3,1,1)+(d4#(eta4-log(w2a4)))`*j(nn4,1,1)
+(d5#(eta5-log(w2a5)))`*j(nn5,1,1);

 para=newbeta//newu1;
 qq=(newu1/newthe1);
 q=j(fixnum,1,0)//qq;
```

```
 RR1=I(cc1)/newthe1;
 RR2=j(fixnum,fixnum,0);
 R=block(RR2,RR1);
V=V+R;    /* covariance matrix */

L=L1-0.5#(cc1#log(newthe1)+(newu1`*newu1/newthe1));

if prelike>L then newpara=para+0.5#(-ginv(V)*q+ginv(V)*(XZ1*L11));
 else newpara=para-ginv(V)*q+ginv(V)*(XZ1*L11);    /* updated parameter */
 initbeta=newpara[1:fixnum,];
 initu1=newpara[fixnum+1:fixnum+cc1,];

invV=ginv(V);
   A11=invV[1:fixnum,1:fixnum];   /* covariance matrix of beta */
   A22=invV[fixnum+1:fixnum+cc1,fixnum+1:fixnum+cc1];
   inithe1=initu1`*initu1/(cc1-(trace(A22)/newthe1));

dif=abs(prelike-L);
difx=prelike-L;
paradif=newpara-para;

end;

/* S.E. of beta */
varbeta=vecdiag(A11);
sebeta=sqrt(varbeta);

inithev1=2#(inithe1##2)/
              (cc1-2#trace(A22)/newthe1+trace(A22**2)/(inithe1##2));
sethe1=sqrt(inithev1);

vname={'estimate', 'se'};
estse1=initbeta||sebeta;
estse2=inithe1||sethe1;
reset noname;
print,'fixed effects',,estse1[rowname=variable colname=vname];
print,'variance of random effects (interaction)',,estse2[rowname=compo
colname=vname];
interterms={'no1001', 'no1002', 'no1003', 'no1004','no1005'};
print,'frailty (interaction)',,initu1[rowname=interterms];

quit;
```

B.3. Cox model with random trial effects and random treatment effects using Efron method for handling ties

/* data sorting steps are same as Appendix B..1 up to **POINT X** */

```
/* Program for analysis of epilepsy data */

proc iml;

cc1=5;                  /* number of trials */
cc2=1225;               /* total sample size */

 use data;
 read all into P;
 d=P[,3];               /* vector indicating censoring (censor) */
 set=P[,12];            /* vector indicating number at risk (atrisk) */

use data6;              /* vector indicating ties */
read all into H;
t=H[,3];
 nn=nrow(t);
*-----------------------*;
j=t[1];
 if j > 1 then
  do;
   tt0=repeat(0,j-1,1);
   tt1=j//tt0;
  end;
 else tt1=j;
*-----------------------*;
 do q=2 to nn;
  j=t[q];
   if j > 1 then
    do;
     tt0=repeat(0,j-1,1);
     tt2=j//tt0;
    end;
   else tt2=j;
  tt1=tt1//tt2;
 end;

/* creating vector for efron */
*-----------------------*;
 j=t[1];
if j > 1 then
 do;
   efron0=1;
      pos0=0;
                    pos0a=0;
    do k=1 to j-1;
   efron2=(j-k)/j;
      pos2=k;
                    pos2a=j-(k+1);
   efron0=efron0//efron2;
      pos0=pos0//pos2;
```

```
                    pos0a=pos0a//pos2a;
      end;

 end;
 else do;
  efron0=j;
     pos0=0;
                 pos0a=0;
 end;
*-----------------------*;
 do q=2 to nn;
 j=t[q];
  if j > 1 then
   do;
    efron2=1;
    pos2=0;
          pos2a=0;
   do k=1 to j-1;
     efron1=(j-k)/j;
           pos1=k;
           pos1a=j-(k+1);
     efron2=efron2//efron1;
           pos2=pos2//pos1;
           pos2a=pos2a//pos1a;
    end;
   end;
   else do;
    efron2=j;
       pos2=0;
                   pos2a=0;
      end;
 efron0=efron0//efron2;
 pos0=pos0//pos2;
 pos0a=pos0a//pos2a;
end;

X=P[,1];                /* design matrix (fixed effects) */
Z0=P[,7:11];            /* design matrix (random effects baseline) */
Z1=Z0#P[,1];            /* design matrix (random effects interaction) */

/* fixed effects */
 variable={'treat'};
 fixnum=1;                          /* number of fixed effects */
 initbeta=j(fixnum,1,0);            /* vector of initial values */

/* random effects */
 frailty={'no1001', 'no1002', 'no1003', 'no1004', 'no1005'};
 initu0=j(cc1,1,0);                 /* initial value of variance for baseline */
 initu1=j(cc1,1,0);                 /* initial value of variance for interaction */

 compo={'theta0','theta1'};   /* variance component of frailty */
```

```
inithe0=1;                    /* baseline */
inithe1=1;                    /* interaction */

prelike=0;
iterate=0;
dif=10000;
L=0;

do k=1 to 500 until(dif<1e-6);  /* convergence criterion */
iterate=iterate+1;
prelike=L;

newbeta=initbeta;
newu0=initu0;
newu1=initu1;
newthe0=inithe0;
newthe1=inithe1;

/* parameter estimation */
 eta=X*newbeta+Z0*newu0+Z1*newu1;
 w1=exp(eta);
W=diag(w1);

 WW1=diag(efron0);

do i=1 to cc2;
 if WW1[i,i]=1 then WW1[i,i:cc2]=1;
 if WW1[i,i]<1 then if WW1[i,i]>0 then do;
    val=WW1[i,i];
       s=pos0[i];
                  sa=pos0a[i];
                  if i ^= cc2 then do;
                WW1[i,i-s:i+sa]=val;
                        if i+sa ^= cc2 then do;
                                WW1[i,i+sa+1:cc2]=1;
                        end;
                  end;
                  if i=cc2 then do;
                   WW1[i,i-s:i+sa]=val;
          end;
   end;
  end;

setefs=j(cc2,1,.);
setefs1=j(cc2,1,.);
 do i=1 to cc2;
   setef=set[i];
   setef1=set[i];
   if efron0[i]<1 then if efron0[i]>0 then do;
     setef=efron0[i];
        setef1=1;
```

```
  end;
  setefs[i]=setef;
  setefs1[i]=setef1;
end;

 WW2s=diag(setefs);
  do i=1 to cc2;
  if WW2s[i,i]=1 then WW2s[i,i:cc2]=1;
  if WW2s[i,i]<1 then if WW2s[i,i]>0 then do;
     val=WW2s[i,i];
        s=pos0[i];
        sa=pos0a[i];
        if i ^= cc2 then do;
      WW2s[i,i-s:i+sa]=val;
              if i+sa ^= cc2 then do;
                    WW2s[i,i+sa+1:cc2]=1;
              end;

        end;
        if i=cc2 then do;
         WW2s[i,i-s:i+sa]=val;
        end;
 end;
end;

M2s=WW2s`;

 w2a=WW2s*w1;
  do kk=1 to cc2;
   if w2a[kk]=0 then w2a[kk]=10000;
  end;

a1a=setefs1/w2a;
As=diag(a1a);
AAs=diag(d/w2a);
M1=WW1`;

 BB1=M1*AAs*j(cc2,1,1);
 B1=diag(BB1);

 L11=d-W*M1*As*j(cc2,1,1);    /* derivative of partial likelihood */

 TIES2=diag(d);

 L12=W*B1-W*M1*TIES2*AAs*AAs*M1`*W;   /* second derivative */

XZ1=X`//Z0`//Z1`;
XZ2=X||Z0||Z1;
V=XZ1*L12*XZ2;   /* covariance matrix */

 L1=(d#(eta-log(w2a)))`*j(cc2,1,1);   /* log-likelihood */
```

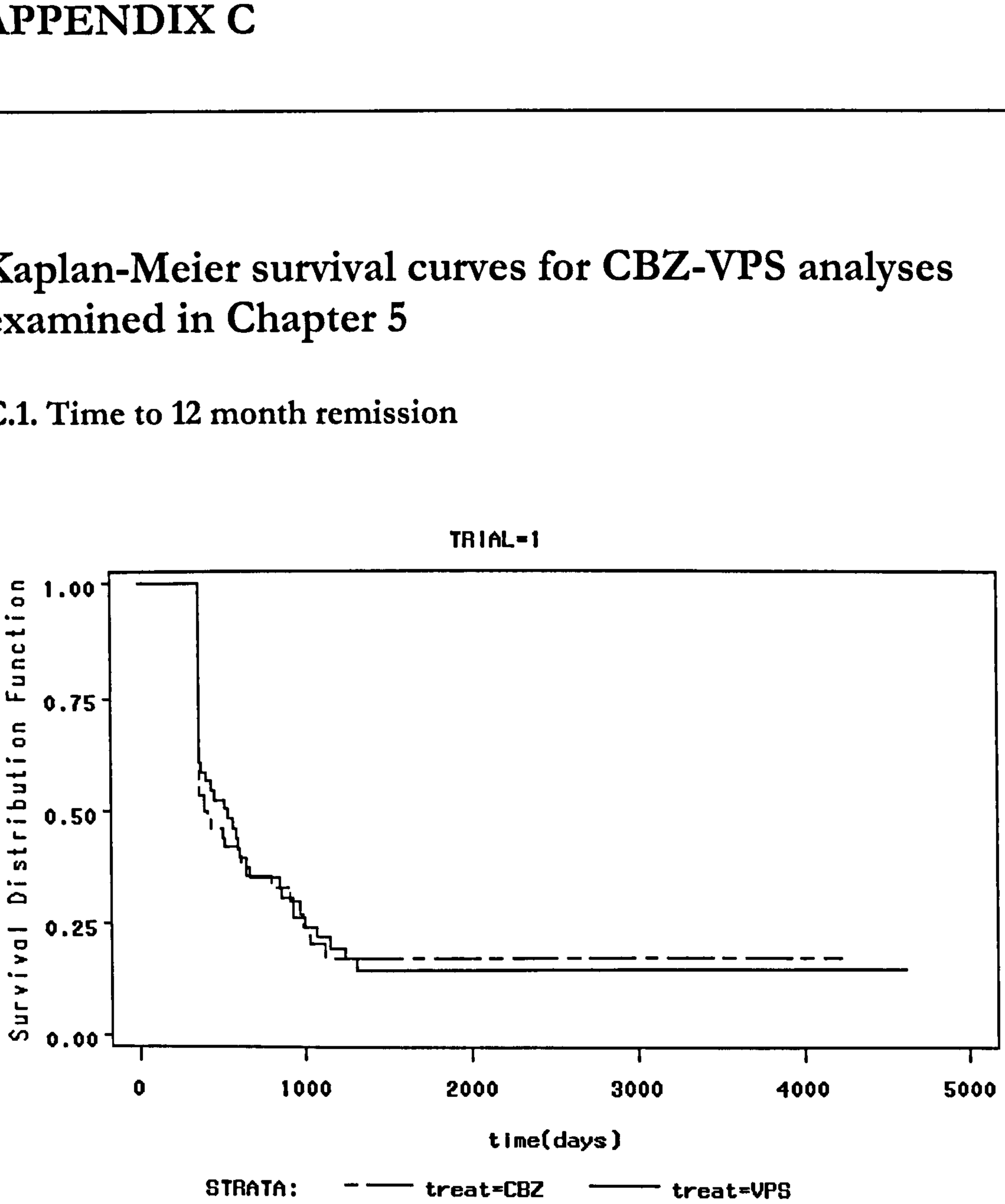
```
para=newbeta//newu0//newu1;
qq=(newu0/newthe0)//(newu1/newthe1);
q=j(fixnum,1,0)//qq;

RR0=I(cc1)/newthe0;
RR1=I(cc1)/newthe1;
RR2=j(fixnum,fixnum,0);
R=block(RR2,RR0,RR1);
V=V+R;    /* covariance matrix */

L=L1 - 0.5
#(cc1#log(newthe0#newthe1)+(newu0`*newu0/newthe0+newu1`*newu1/newthe1));

if prelike>L then newpara=para+0.5#(-ginv(V)*q+ginv(V)*(XZ1*L11));
 else newpara=para-ginv(V)*q+ginv(V)*(XZ1*L11);    /* updated parameter */

 initbeta=newpara[1:fixnum,];
 initu0=newpara[fixnum+1:fixnum+cc1,];
 initu1=newpara[fixnum+1+cc1:fixnum+cc1#2,];

  invV=ginv(V);
    A11=invV[1:fixnum,1:fixnum];   /* covariance matrix of beta */
    A22=invV[fixnum+1:fixnum+cc1,fixnum+1:fixnum+cc1];
    A33=invV[fixnum+1+cc1:fixnum+cc1#2,fixnum+1+cc1:fixnum+cc1#2];
 inithe0=initu0`*initu0/(cc1-trace(A22)/newthe0);
 inithe1=initu1`*initu1/(cc1-trace(A33)/newthe1);

dif=abs(prelike-L);
difx=prelike-L;
paradif=newpara-para;

end;

/* S.E. of beta */
 varbeta=vecdiag(A11);
 sebeta=sqrt(varbeta);

/* S.E. of theta (REML estimate) */
 inithev0=2#(inithe0##2)/
              (cc1-2#trace(A22)/newthe0+trace(A22**2)/(inithe0##2));
 sethe0=sqrt(inithev0);
 inithev1=2#(inithe1##2)/
              (cc1-2#trace(A33)/newthe1+trace(A33**2)/(inithe1##2));
 sethe1=sqrt(inithev1);

/* print */
vname={'estimate', 'se'};
estse1=initbeta||sebeta;
estse2=(inithe0||sethe0)//(inithe1||sethe1);
reset noname;
```

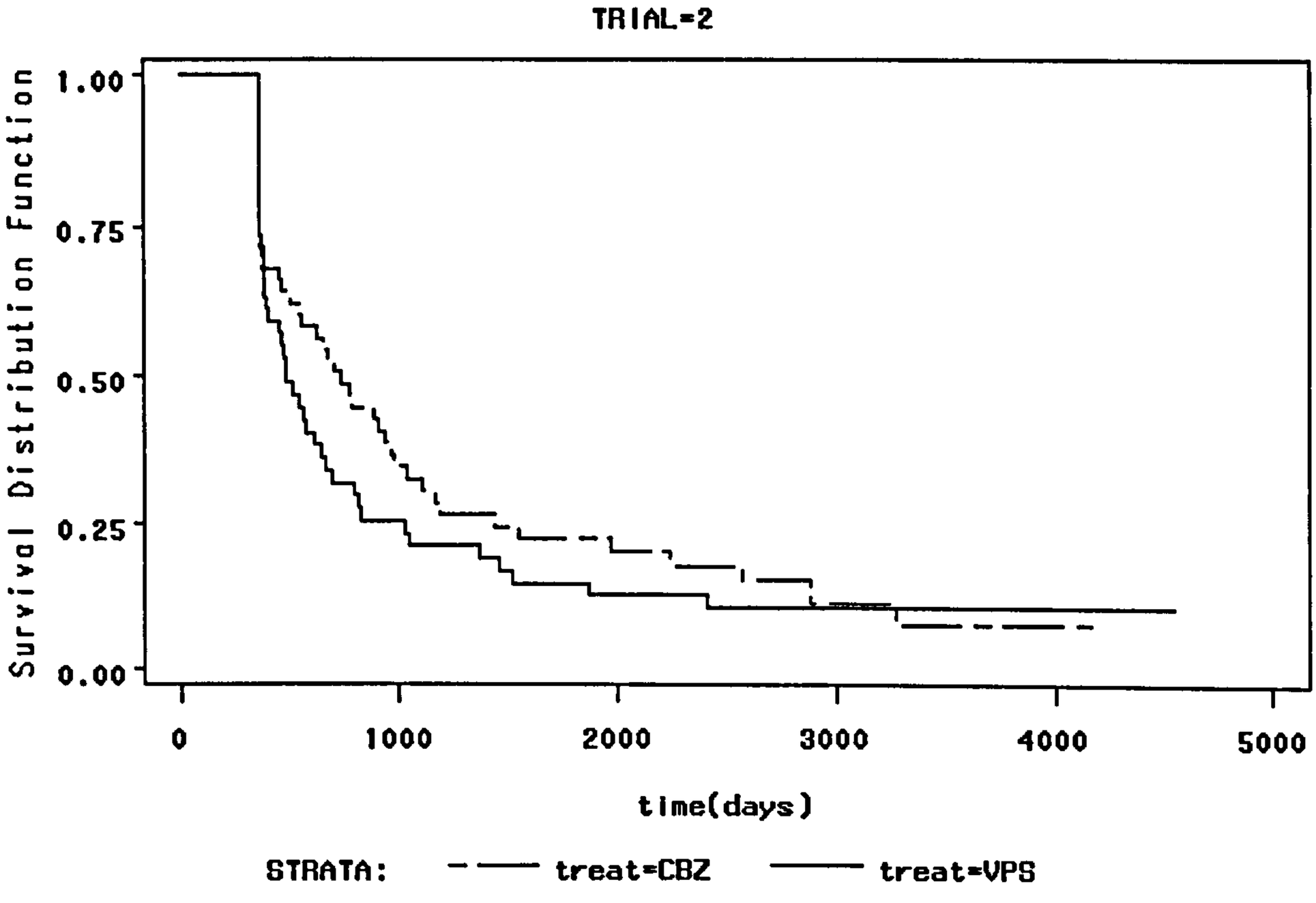
```
print,'fixed effects',,estse1[rowname=variable colname=vname];
print,'variance of random effects (baseline, interaction)',,estse2[rowname=compo
colname=vname];
print,'frailty (baseline)',,initu0[rowname=frailty];
print,'frailty (interaction)',,initu1[rowname=frailty];

quit;
```

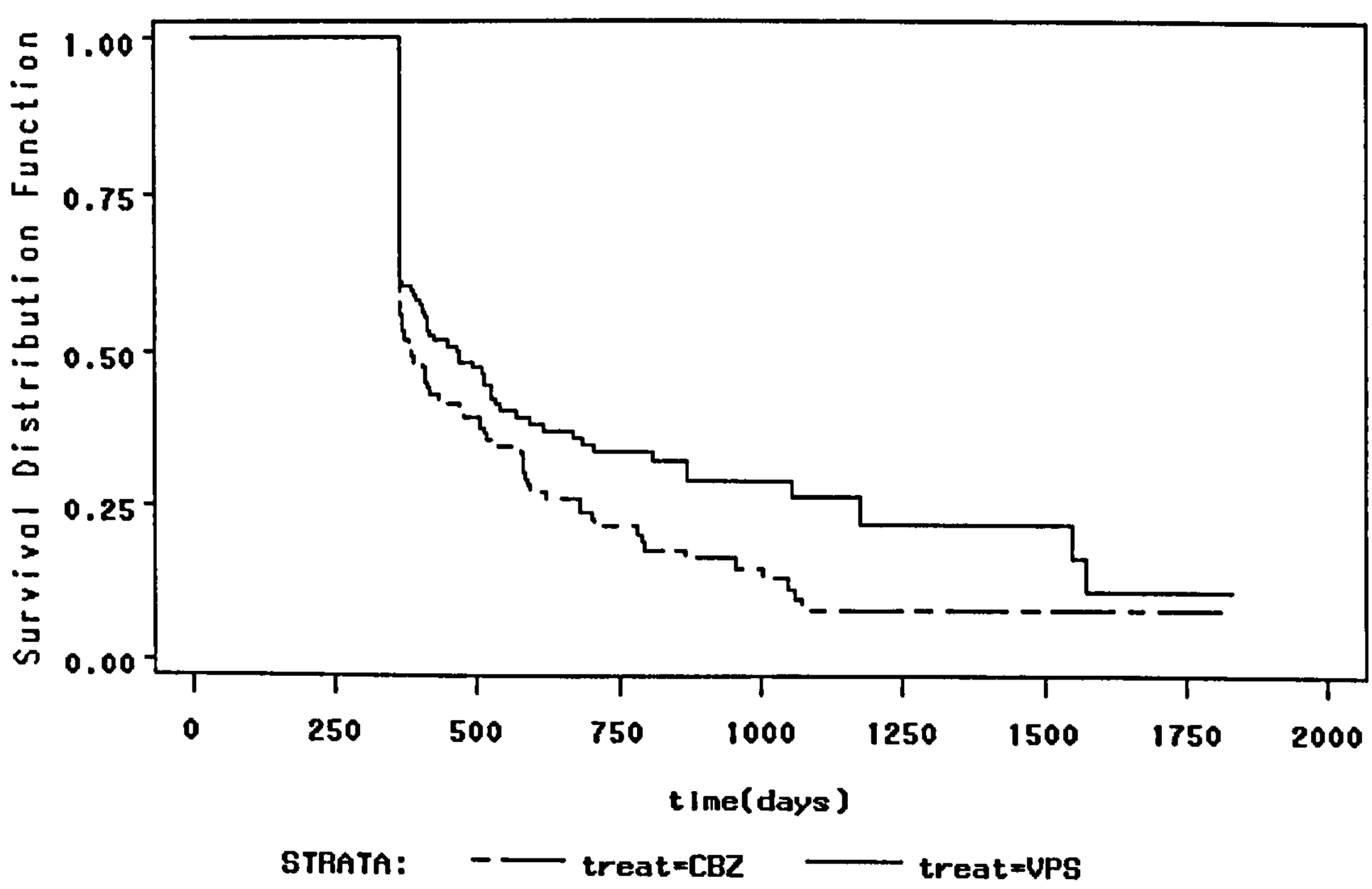
APPENDIX C

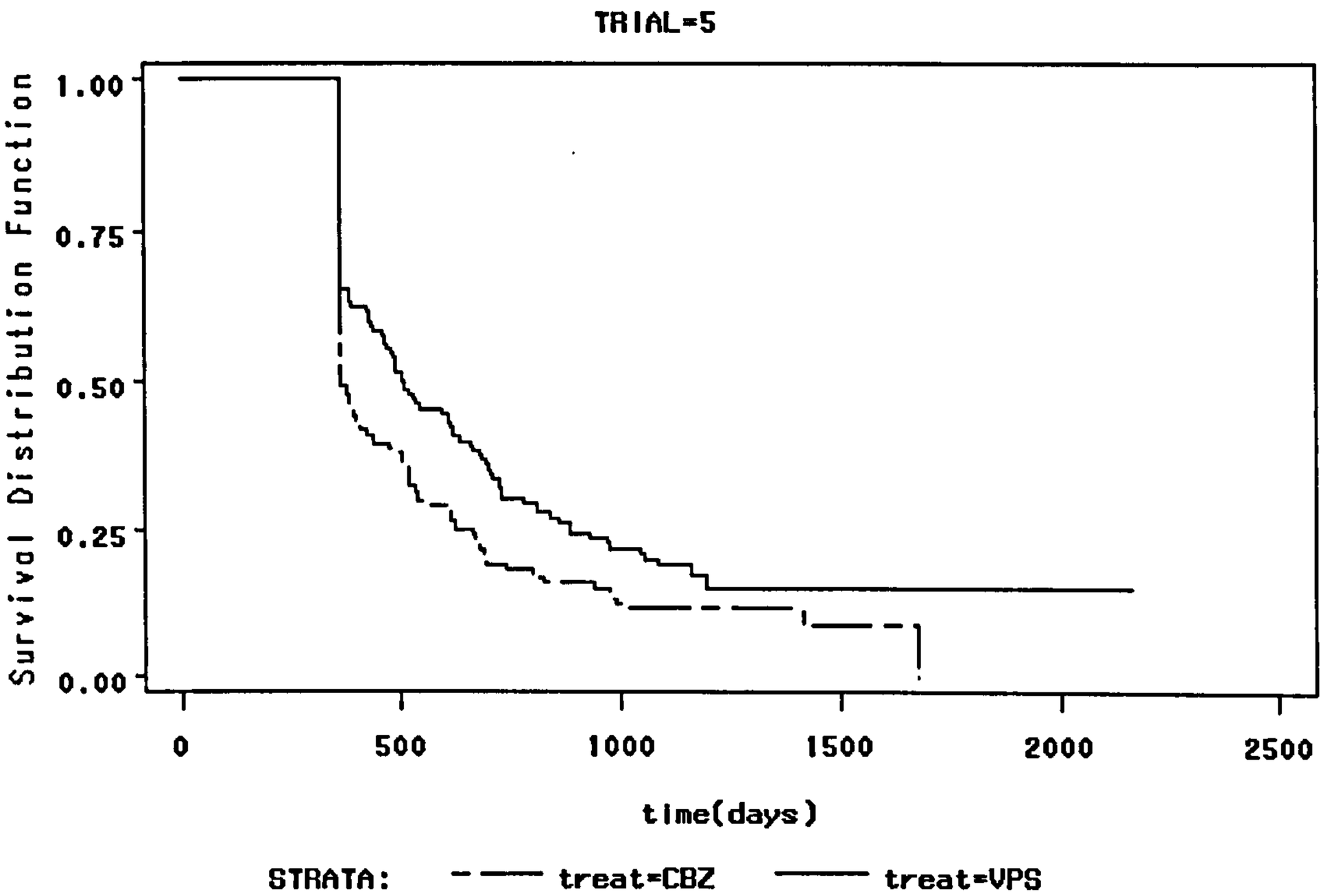
Kaplan-Meier survival curves for CBZ-VPS analyses examined in Chapter 5

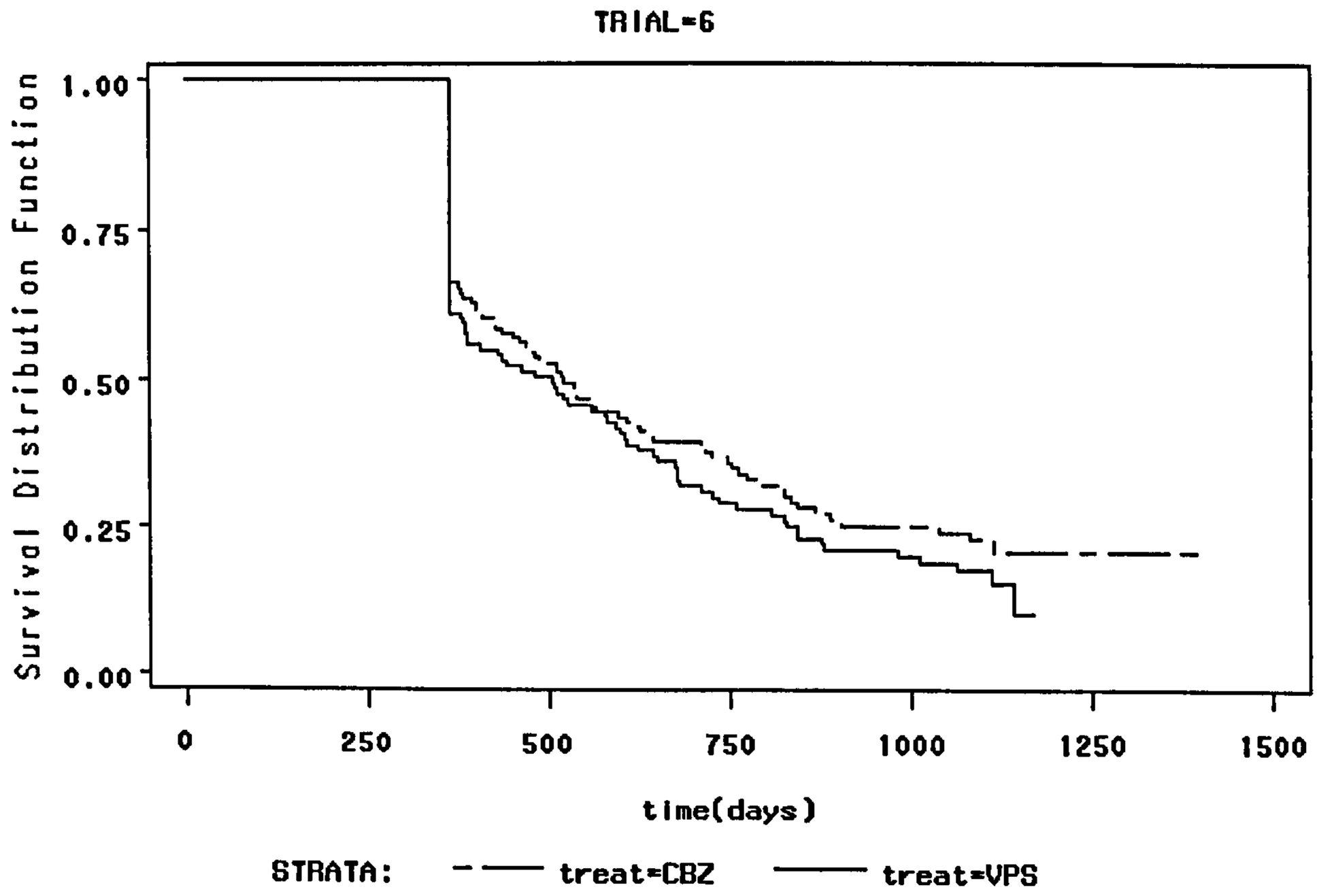
C.1. Time to 12 month remission



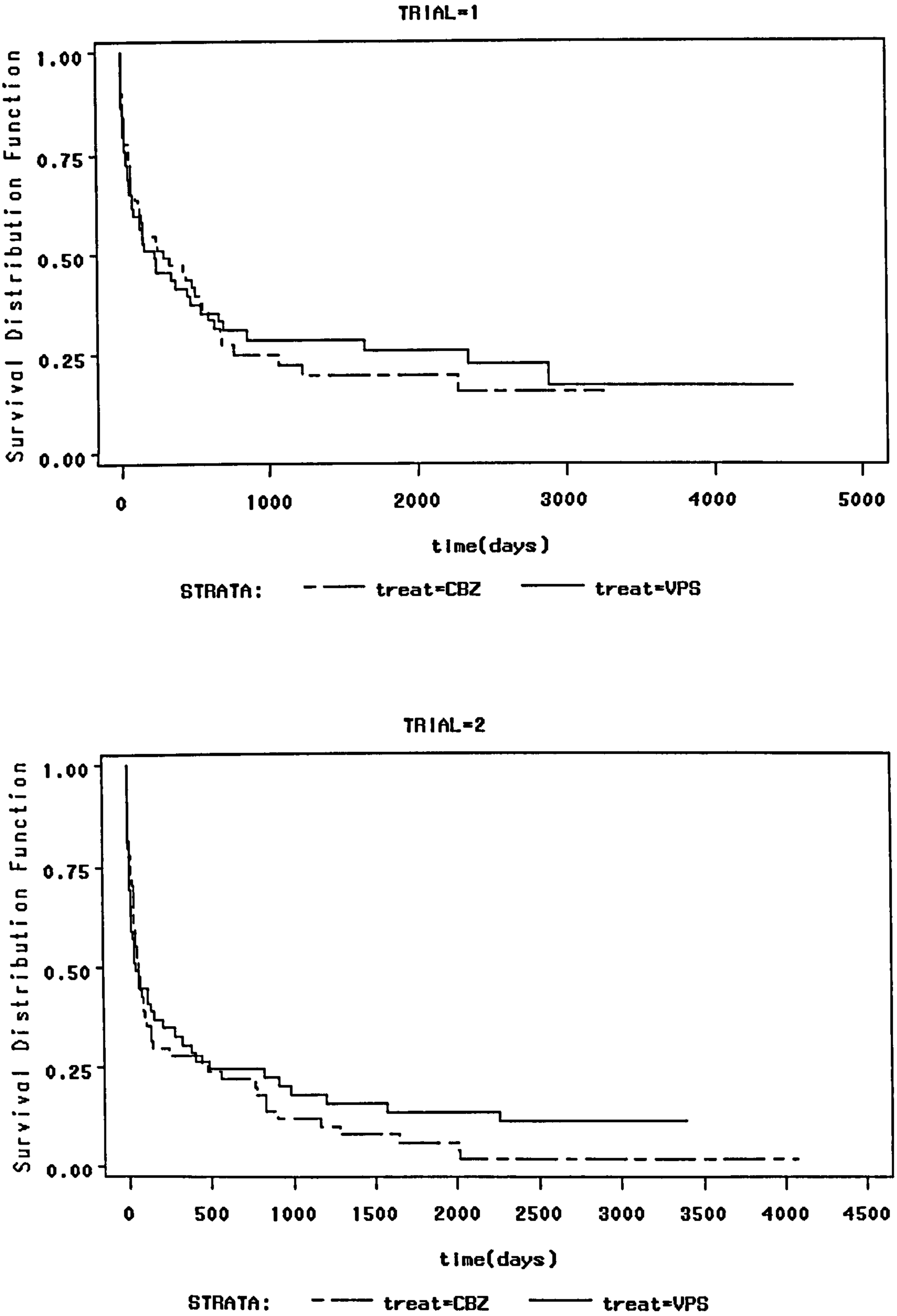

TRIAL=2
Survival Distribution Function
1.00
0.75
0.50
0.25
0.00
0
1000
2000
3000
4000
5000
time(days)
STRATA: treat=CBZ treat=VPS

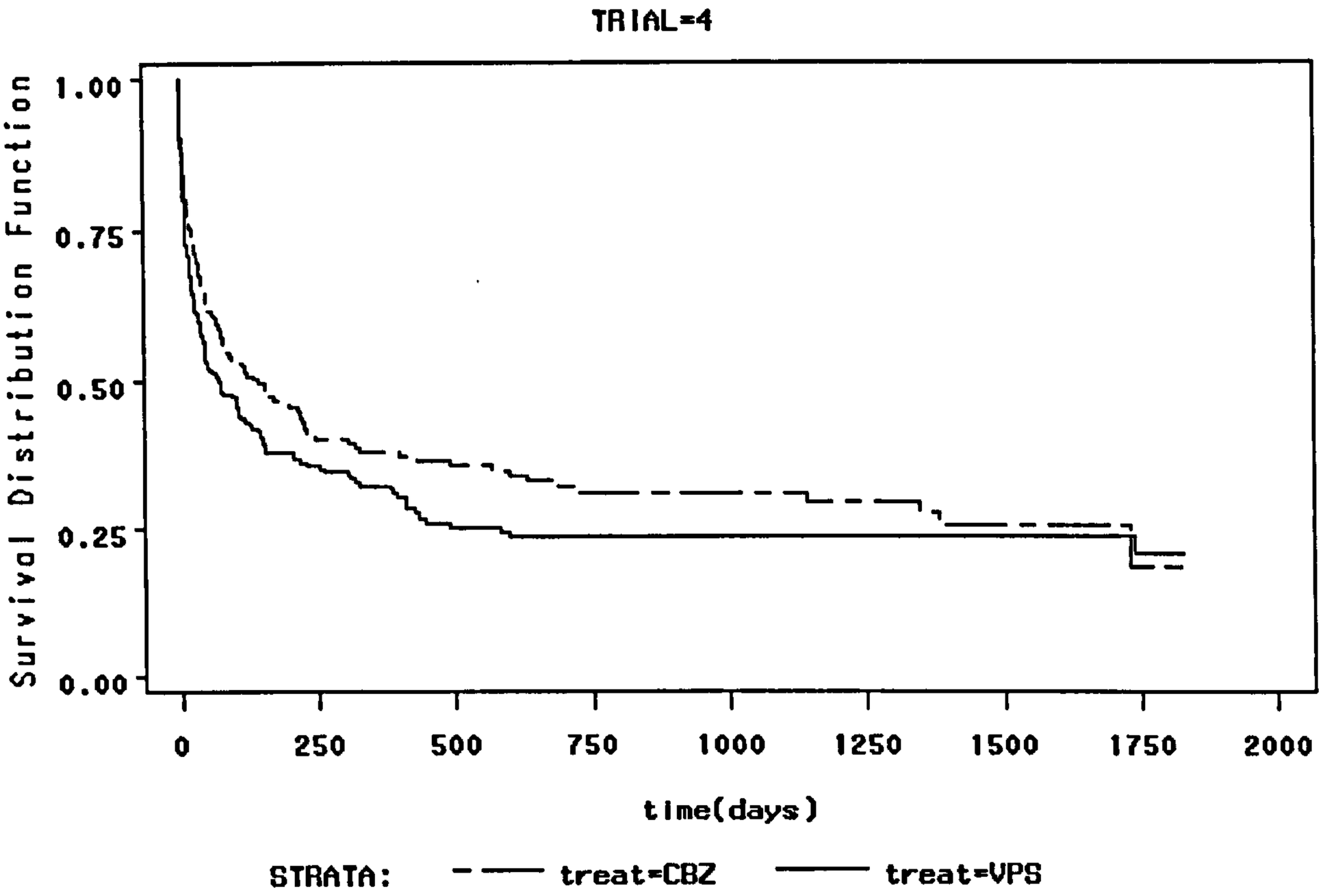

TRIAL=4
Survival Distribution Function
1.00
0.75
0.50
0.25
0.00
0
250
500
750
1000
1250
1500
1750
2000
time(days)
STRATA: treat=CBZ treat=VPS

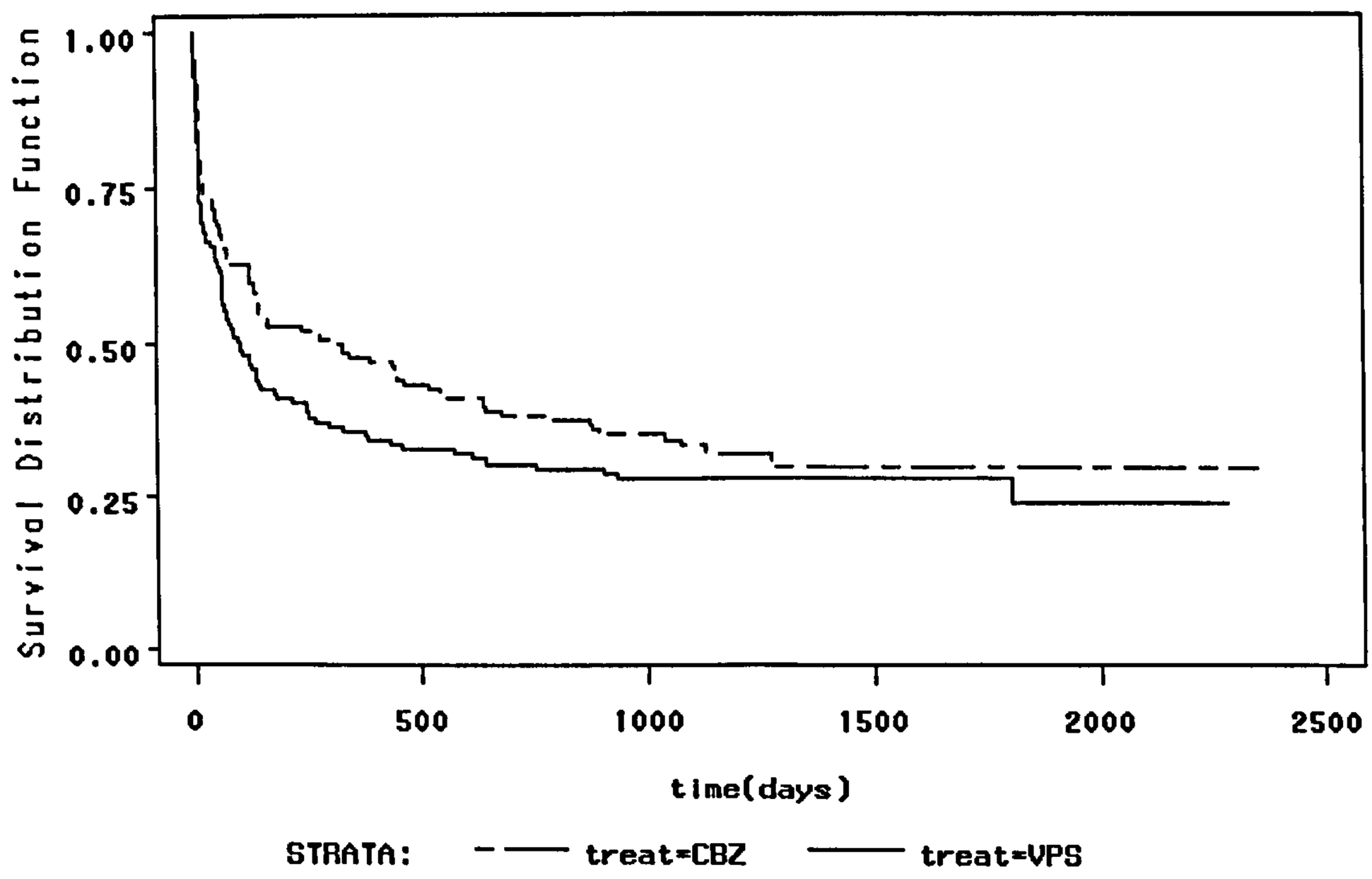

TRIAL=5
Survival Distribution Function
1.00
0.75
0.50
0.25
0.00
0
500
1000
1500
2000
2500
time(days)
STRATA: treat=CBZ treat=VPS

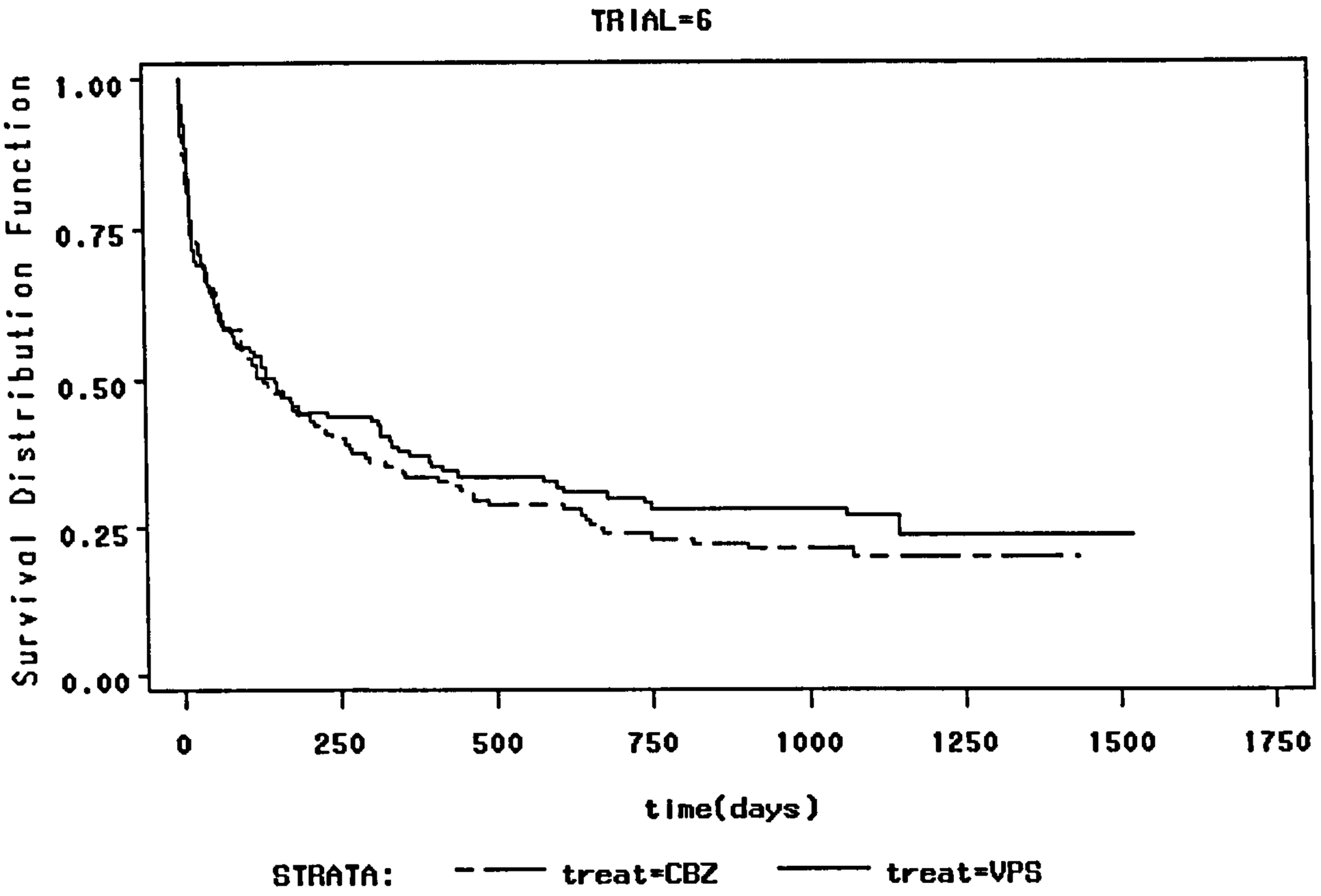

TRIAL=6
Survival Distribution Function
1.00
0.75
0.50
0.25
0.00
0
250
500
750
1000
1250
1500
time(days)
STRATA: treat=CBZ treat=VPS

C.2. Time to first seizure




TRIAL=4
Survival Distribution Function
1.00
0.75
0.50
0.25
0.00
0
250
500
750
1000
1250
1500
1750
2000
time(days)
STRATA: treat=CBZ treat=VPS


TRIAL=5
Survival Distribution Function
1.00
0.75
0.50
0.25
0.00
0
500
1000
1500
2000
2500
time(days)
STRATA: treat=CBZ treat=VPS


TRIAL=6
Survival Distribution Function
1.00
0.75
0.50
0.25
0.00
0
250
500
750
1000
1250
1500
1750
time(days)
STRATA: treat=CBZ treat=VPS

APPENDIX D

Calculations for indirect comparisons relating to Chapter 6

D.1. Calculating indirect comparison for PHT:CBZ

Aggregate data

$$\log \hat{HR}_{PHT:CBZindirect} = \log \hat{HR}_{PHT:VPS} - \log \hat{HR}_{CBZ:VPS}$$

$$\text{var}(\log \hat{HR}_{PHT:CBZindirect}) = \text{var}(\log \hat{HR}_{PHT:VPS}) + \text{var}(\log \hat{HR}_{CBZ:VPS})$$

Under the assumption of a fixed effect of treatment across trials within each pair-wise comparison,

AD-indirect fixed

$$\log \hat{HR}_{PHT:CBZindirect} = 0.0868 - (0.1905) = -0.1037$$

$$\text{var}(\log \hat{HR}_{PHT:CBZindirect}) = 0.1804^2 + 0.0824^2 = 0.0393$$

whilst assuming random treatment effects across trials within each pair-wise comparison,

AD-indirect random

$$\log \hat{HR}_{PHT:CBZindirect} = 0.08665 - (0.17827) = -0.09162$$
$$\text{var}(\log \hat{HR}_{PHT:CBZindirect}) = 0.18041^2 + 0.16617^2 = 0.0602$$

Individual patient data

In model (6.5) and (6.8), the treatment coding structure adopted is such that $x_{1ij} = 1$ for treatment group PHT, $x_{2ij} = 1$ for treatment group CBZ, with VPS taking value zero for both dummy variables.

Under the assumption of a fixed effect of treatment across trials within each pair-wise comparison,

$$\log \hat{HR}_{PHT:VPS} = \hat{\beta}_1 = 0.08665, \quad SE(\hat{\beta}_1) = 0.18041$$
$$\log \hat{HR}_{CBZ:VPS} = \hat{\beta}_2 = 0.19049, \quad SE(\hat{\beta}_2) = 0.08243$$

IPD-indirect fixed

$$\log \hat{HR}_{PHT:CBZindirect} = \hat{\beta}_1 - \hat{\beta}_2$$
$$= 0.08665 - (0.19049) = -0.10384$$
$$\text{var}(\log \hat{HR}_{PHT:CBZindirect}) = \text{var}(\hat{\beta}_1) + \text{var}(\hat{\beta}_2)$$
$$= 0.18041^2 + 0.08243^2 = 0.0393$$

Under the assumption of random treatment effects across trials within each pair-wise comparison, the following results are obtained

$$\log \hat{HR}_{PHT:VPS} = \hat{\beta}_1 = 0.0866, \quad SE(\hat{\beta}_1) = 0.18041,$$
$$\hat{\tau}_1^2 = 0$$

$$\log \hat{HR}_{CBZ:VPS} = \hat{\beta}_2 = 0.1783, \quad SE(\hat{\beta}_2) = 0.1662,$$
$$\hat{\tau}_2^2 = 0.0625, \quad SE(\hat{\tau}_2^2) = 0.0828$$

IPD-indirect random

$$\log \hat{HR}_{PHT:CBZindirect} = \hat{\beta}_1 - \hat{\beta}_2 = 0.0866 - (0.1783) = -0.0917$$

$$\text{var}(\log \hat{HR}_{PHT:CBZindirect}) = \text{var}(\hat{\beta}_1) + \text{var}(\hat{\beta}_2) = 0.18041^2 + 0.1662^2 = 0.0602$$

$$\hat{\tau}^2_{PHT:CBZindirect} = \hat{\tau}_1^2 + \hat{\tau}_2^2 = 0 + 0.0625 = 0.0625$$

D.2. Calculating indirect comparison for VPS:CBZ

Aggregate data

$$\log \hat{HR}_{VPS:CBZindirect} = \log \hat{HR}_{VPS:PHT} - \log \hat{HR}_{CBZ:PHT}$$

$$\text{var}(\log \hat{HR}_{VPS:CBZindirect}) = \text{var}(\log \hat{HR}_{VPS:PHT}) + \text{var}(\log \hat{HR}_{CBZ:PHT})$$

Under the assumption of a fixed effect of treatment across trials within each pair-wise comparison,

AD-indirect fixed

$$\log \hat{HR}_{VPS:CBZindirect} = -0.0868 - (0.0076) = -0.0944$$

$$\text{var}(\log \hat{HR}_{VPS:CBZindirect}) = 0.1804^2 + 0.1193^2 = 0.0468$$

whilst assuming random treatment effects across trials within each pair-wise comparison,

AD-indirect random

$$\log \hat{HR}_{VPS:CBZindirect} = -0.08665 - (0.0076) = -0.0943$$

$$\text{var}(\log \hat{HR}_{VPS:CBZindirect}) = 0.18041^2 + 0.1193^2 = 0.0468$$

Individual patient data

In model (6.5) and (6.8), the treatment coding structure adopted is such that $x_{1ij} = 1$ for treatment group VPS, $x_{2ij} = 1$ for treatment group CBZ, with PHT taking value zero for both dummy variables.

Under the assumption of a fixed effect of treatment across trials within each pair-wise comparison,

$$\log \hat{HR}_{VPS:PHT} = \hat{\beta}_1 = -0.08682, \quad SE(\hat{\beta}_1) = 0.18041$$
$$\log \hat{HR}_{CBZ:PHT} = \hat{\beta}_2 = 0.00755, \quad SE(\hat{\beta}_2) = 0.11931$$

IPD-indirect fixed

$$\log \hat{HR}_{VPS:CBZindirect} = \hat{\beta}_1 - \hat{\beta}_2$$
$$= -0.08682 - (0.00755) = -0.09437$$
$$\text{var}(\log \hat{HR}_{VPS:CBZindirect}) = \text{var}(\hat{\beta}_1) + \text{var}(\hat{\beta}_2)$$
$$= 0.18041^2 + 0.1193^2 = 0.0468$$

Under the assumption of random treatment effects across trials within each pair-wise comparison, the following results are obtained

$$\log \hat{HR}_{VPS:PHT} = \hat{\beta}_1 = -0.0866, \quad SE(\hat{\beta}_1) = 0.18041,$$
$$\hat{\tau}_1^2 = 0$$

$$\log \hat{HR}_{CBZ:PHT} = \hat{\beta}_2 = 0.0076, \quad SE(\hat{\beta}_2) = 0.1193,$$
$$\hat{\tau}_2^2 = 0$$

IPD-indirect random

$$\log \hat{HR}_{VPS:CBZindirect} = \hat{\beta}_1 - \hat{\beta}_2$$
$$= -0.0866 - (0.0076) = -0.0942$$

$$\text{var}(\log \hat{HR}_{VPS:PHTindirect}) = \text{var}(\hat{\beta}_1) + \text{var}(\hat{\beta}_2)$$
$$= 0.18041^2 + 0.1193^2 = 0.0468$$

$$\hat{\tau}^2{}_{VPS:PHTindirect} = \hat{\tau}_1^2 + \hat{\tau}_2^2$$
$$= 0$$

APPENDIX E

Parameter estimates for totality of evidence relating to Chapter 6

E.1. Totality of evidence model exploring treatment main effects (model 6.10)

Table E.1.1 Time to first seizure: Parameter estimates, standard errors and variance-covariance matrix from Cox model stratified by trial including totality of evidence (17 trials, 2130 events, 3785 total)

Parameter	Estimate (standard error)	Variance-covariance matrix				
		β_1 (cbz)	β_2 (phb)	β_3 (pht)	β_4 (vps)	β_5 (ltg)
β_1 (cbz)	0.088 (0.157)	0.024788				
β_2 (phb)	-0.078 (0.169)	0.021372	0.028434			
β_3 (pht)	0.078 (0.132)	0.017529	0.017529	0.017529		
β_4 (vps)	0.157 (0.157)	0.022716	0.020646	0.017529	0.02479	
β_5 (ltg)	0.237 (0.194)	0.024788	0.021372	0.017529	0.02272	0.03776
-2logL	21215.580					
change$^{\phi}$, df, p-value	6.85, 5 p=0.23					

$^{\phi}$ Change compared to null model without treatment indicator variables

Table E.1.2. Time to 12 month remission: Parameter estimates, standard errors and variance-covariance matrix from Cox model stratified by trial including totality of evidence (12 trials, 1430 events, 2728 total)

Parameter	Estimate (standard error)	Variance-covariance matrix			
		β_1 (cbz)	β_2 (phb)	β_3 (pht)	β_4 (vps)
β_1 (cbz)	-0.045 (0.182)	0.033259			
β_2 (phb)	-0.106 (0.195)	0.029015	0.03803		
β_3 (pht)	-0.086 (0.154)	0.023794	0.02379	0.023794	
β_4 (vps)	-0.176 (0.183)	0.030900	0.02829	0.023794	0.03342
-2logL change$^{\phi}$, df, p-value	13079.385 3.95,4 p=0.41				

$^{\phi}$ Change compared to null model without treatment indicator variables

Table E.1.3. Time to withdrawal: Parameter estimates, standard errors and variance-covariance matrix from Cox model stratified by trial including totality of evidence (15 trials, 1070 events, 3830 total)

Parameter	Estimate (standard error)	Variance-covariance matrix				
		β_1 (cbz)	β_2 (phb)	β_3 (pht)	β_4 (vps)	β_5 (ltg)
β_1 (cbz)	0.478 (0.247)	0.06091				
β_2 (phb)	0.855 (0.253)	0.05432	0.06418			
β_3 (pht)	0.501 (0.216)	0.04648	0.04648	0.04648		
β_4 (vps)	0.414 (0.251)	0.05745	0.05303	0.04648	0.06301	
β_5 (ltg)	-0.039 (0.284)	0.06091	0.05432	0.04648	0.05745	0.080632
-2logL change$^{\phi}$, df, p-value	11279.256 29.51, 5, p<0.0001					

$^{\phi}$ Change compared to null model without treatment indicator variables

E.2. Totality of evidence model exploring effect of epilepsy type

Table E.2.1. Direct evidence: Exploring main effect of epilepsy type and interaction with treatment for the outcome time to first seizure (results exclude patients with missing epilepsy type)

Comparison	Parameter	Terms included in model		
		Treatment	Treatment + type	Treatment + type + treatment*type
pht-oxc 2 trials 227 events 468 total	Trt (pht) Type (partial) Type*Trt -2logL change$^{\phi}$, df, p-value	0.070(0.133) 2314.268 0.28, 1, p=0.60	0.046(0.133) 0.363(0.155) 2308.497 **5.77, 1,** **p=0.02**	-0.098(0.264) 0.271(0.209) 0.195(0.306) 2308.088 0.41, 1, p=0.52
cbz-phb 4 trials 365 events 676 total	Trt (cbz) Type (partial) Type*Trt -2logL change$^{\phi}$, df, p-value	0.157(0.110) 3238.227 2.05, 1, p=0.15	0.173(0.110) 0.396(0.146) 3230.852 **7.38, 1,** **p=0.01**	-0.301(0.211) 0.005(0.204) 0.632(0.243) 3224.182 **6.67, 1,** **p=0.01**
cbz-pht 3 trials 362 events 545 total	Trt (cbz) Type (partial) Type*Trt -2logL change$^{\phi}$, df, p-value	0.093(0.106) 3231.833 0.77, 1, p=0.38	0.106(0.106) 0.690(0.149) 3210.833 **21.37, 1,** **p<0.001**	0.069(0.212) 0.666(0.191) 0.050(0.244) 3210.421 0.04, 1, p=0.84
cbz-vps 5 trials 864 events 1225 total	Trt (cbz) Type (partial) Type*Trt -2logL change$^{\phi}$, df, p-value	-0.08(0.068) 8593.508 1.35, 1, p=0.24	-0.09(0.068) 0.503(0.085) 8558.785 **34.72, 1,** **p<0.001**	0.140(0.120) 0.676(0.114) -0.339(0.146) 8553.387 **5.40, 1,** **p=0.02**
cbz-ltg **4 trials** **315 events** **686 total**	Trt (cbz) Type (partial) Type*Trt	-0.163(0.119)	-0.211(0.119) 0.663(0.125)	-0.023(0.214) 0.758(0.156) -0.267(0.256)

Comparison	Parameter	Terms included in model		
		Treatment	Treatment + type	Treatment + type + treatment*type
	-2logL	3014.089	2983.900	2982.818
	change$^{\phi}$, df, p-value	1.90, 1, p=0.17	**30.19, 1, p<0.001**	1.08, 1, p=0.30
phb-pht				
4 trials	Trt (phb)	-0.170(0.113)	-0.164(0.113)	0.044(0.222)
351 events	Type (partial)		0.474(0.150)	0.579(0.181)
592 total	Type*Trt			-0.274(0.254)
	-2logL	3089.253	3079.157	3077.999
	change$^{\phi}$, df, p-value	2.27, 1, p=0.13	**10.09, 1, p=0.001**	1.16, 1, p=0.28
phb-vps				
2 trials	Trt (phb)	0.045(0.189)	0.007(0.189)	0.244(0.251)
134 events	Type (partial)		0.531(0.177)	0.716(0.221)
178 total	Type*Trt			-0.510(0.366)
	-2logL	1018.065	1009.238	1007.280
	change$^{\phi}$, df, p-value	0.06, 1, p=0.81	**8.83, 1, p=0.003**	1.96, 1, p=0.16
pht-vps				
5 trials	Trt (pht)	-0.043(0.105)	-0.088(0.105)	0.032(0.149)
371 events	Type (partial)		0.872(0.114)	0.991(0.156)
639 total	Type*Trt			-0.235(0.209)
	-2logL	3192.808	3133.542	3132.282
	change$^{\phi}$, df, p-value	0.16, 1, p=0.68	**59.27, 1, p<0.001**	1.26, 1, p=0.26

Table E.2.2. Totality of evidence: Exploring main effect of epilepsy type and interaction with treatment for the outcome time to first seizure (17 trials, 2099 events, 3725 total excluding all patients with missing epilepsy type)

Parameter	Terms included in model		
	Treatment	Treatment + type	Treatment + type + treatment*type
β_1 (cbz)	0.080 (0.158)	0.062 (0.158)	-0.013 (0.224)
β_2 (phb)	-0.086 (0.169)	-0.116 (0.169)	0.053 (0.254)
β_3 (pht)	0.070 (0.133)	0.034 (0.133)	-0.232 (0.218)
β_4 (vps)	0.149 (0.158)	0.141 (0.158)	-0.200 (0.223)
β_5 (ltg)	0.243 (0.198)	0.267 (0.198)	-0.046 (0.268)
β_6 (partial)		0.562 (0.055)	0.274 (0.209)
β_7 (partial*cbz)			0.133 (0.228)
β_8 (partial*phb)			-0.216 (0.272)
β_9 (partial*pht)			0.361 0.239)
β_{10} (partial*vps)			0.530 (0.232)
β_{11} (partial*ltg)			0.483 (0.261)
-2logL	20854.289	20746.423	20725.387
change$^{\phi}$, df, p-value	6.954, 1, p=0.22	107.866, 1, p<0.001	21.036, 1, p<0.001

ϕ Change in –2logL compared to the previous model without the additional parameter(s) under consideration

Table E.2.3. Totality of evidence: Variance-covariance matrix for parameters included in the model with treatment + epilepsy type + interactions with treatment for time to first seizure (model 6.11) (17 trials, 2099 events, 3725 total excluding all patients with missing epilepsy type)

	β_1 (cbz)	β_2 (phb)	β_3 (pht)	β_4 (vps)	β_5 (ltg)	β_6 (partial)	β_7 (partial*cbz)	β_8 (partial*phb)	β_9 (partial*pht)	β_{10} (partial*vps)	β_{10} (partial*ltg)
β_1 (cbz)	0.050	0.043	0.040	0.044	0.046	0.030	-0.036	-0.031	-0.032	-0.030	-0.030
β_2 (phb)	0.043	0.065	0.040	0.042	0.043	0.030	-0.031	-0.052	-0.032	-0.030	-0.030
β_3 (pht)	0.040	0.040	0.047	0.041	0.040	0.030	-0.030	-0.031	-0.041	-0.031	-0.030
β_4 (vps)	0.044	0.042	0.041	0.050	0.043	0.030	-0.031	-0.031	-0.033	-0.036	-0.030
β_5 (ltg)	0.046	0.043	0.040	0.043	0.072	0.030	-0.031	-0.030	-0.032	-0.029	-0.047
β_6 (partial)	0.030	0.030	0.030	0.030	0.030	0.044	-0.044	-0.044	-0.043	-0.044	-0.044
β_7 (partial*cbz)	-0.036	-0.031	-0.030	-0.031	-0.031	-0.044	0.052	0.045	0.044	0.044	0.044
β_8 (partial*phb)	-0.031	-0.052	-0.031	-0.031	-0.030	-0.044	0.045	0.074	0.045	0.044	0.044
β_9 (partial*pht)	-0.032	-0.032	-0.041	-0.033	-0.032	-0.043	0.044	0.045	0.057	0.044	0.043
β_{10} (partial*vps)	-0.030	-0.030	-0.031	-0.036	-0.029	-0.044	0.044	0.044	0.044	0.054	0.044
β_{11} (partial*ltg)	-0.030	-0.030	-0.030	-0.030	-0.047	-0.044	0.044	0.044	0.043	0.044	0.068

Table.E.2.4. Totality of evidence model including treatment + epilepsy type + treatment*type interactions (model 6.11)

Entries in table are values of –2logL for model 6.11 excluding individual interaction with treatment terms along with the change in -2logL, df and p-value, compared to the model including all interaction terms (model 6.11 for which -2logL=20725.387)

		Comparator drug				
		cbz	**oxc**	**phb**	**pht**	**vps**
Baseline drug	**oxc**	20725.727 **0.34, 1,** **p=0.56**	-			
	phb	20728.607 **3.22, 1,** **p=0.07**	20726.018 **0.63, 1,** **p=0.43**	-		
	pht	20727.892 **2.51, 1,** **p=0.11**	20727.620 **2.23, 1,** **p=0.14**	20733.164 **7.78, 1,** **p=0.005**	-	
	vps	20734.695 **9.31, 1,** **p=0.002**	20730.356 **4.97, 1,** **p=0.03**	20738.581 **13.19, 1,** **p=0.0003**	20726.644 **1.26, 1,** **p=0.26**	-
	ltg	20729.231 **3.84, 1,** **p=0.05**	20728.749 **3.36, 1,** **p=0.07**	20734.191 **8.80, 1,** **p=0.003**	20725.775 **0.39, 1,** **p=0.53**	20725.450 **0.06, 1,** **p=0.80**

BIBLIOGRAPHY

[1] Sackett DL, Richardson WS, Rosenberg W, Haynes RB. Evidence-Based Medicine. Edinburgh: Churchill Livingstone, 1998

[2] Egger M, Davey Smith G, Altman D. Systematic Reviews in Health Care: Meta-analysis in context. London: BMJ Publishing Group, 2001

[3] The Cochrane Library. Issue 4, 2003.Chichester, UK: John Wiley & Sons, Ltd.

[4] Fleiss J. The statistical basis of meta-analysis. Statistical Methods in Medical Research 1993; 2:121-45.

[5] Alderson P, Green S, Higgins JPT. Cochrane Reviewers' Handbook 4.2.2 [updated March 2004].

[6] Deeks J. Issues in the selection of a summary statistic for meta-analysis of clinical trials with binary outcomes. Statistics in Medicine 2002; **21**:1575-600.

[7] Whitehead A and Whitehead J. A general parametric approach to the meta-analysis of randomized clinical trials. Statistics in Medicine 1991; **10**:1665-77.

[8] Berlin JA, Laird NM, Sacks HS, Chalmers TC. A comparison of statistical methods for combining event rates from clinical trials. Statistics in Medicine 1989; 8:141-51.

[9] Yusuf S, Peto R, Lewis J, Collins R, Sleight P. Beta Blockade during and after myocardial infarction: An overview of the randomized trials. Progress in Cardiovascular Disease 1985; **27**:335-71.

[10] Hedges LV and Olkin I. Statistical Methods for Meta-analysis. London: Academic Press, 1985.

[11] Hardy RJ and Thompson SG. Detecting and describing heterogeneity in meta-analysis. Statistics in Medicine 1998; **17**:841-56.

[12] DerSimonian R and Laird N. Meta-analysis in clinical trials. Controlled Clinical Trials 1986; **7**:177-88.

[13] Thompson SG. Why sources of heterogeneity in meta-analysis should be investigated. British Medical Journal 1994; **309**:1351-5.

[14] Higgins JPT and Thompson SG. Quantifying heterogeneity in a meta-analysis. Statistics in Medicine 2002; **21**:1539-1558.

[15] Cochran WG . The combination of estimates from different experiments. Biometrics 1954; **10**:101-29.

[16] Gavaghan DJ, Moore RA, McQuay HJ. An evaluation of homogeneity tests in meta-analyses in pain using simulations of individual patient data. Pain 2000; 85:415-24.

[17] Takkouche B, Cadarso-Suarez C, Spiegelman D. Evaluation of old and new tests of heterogeneity in epidemiologic meta-analysis. American Journal of Epidemiology 1999; **150**:206-15.

[18] Jones MP, O'Gorman T, Lemke JH, Woolson R. Monte-Carlo investigation of homogeneity tests of the odds ratio under various sample size configurations. Biometrics 1989; **45**:171-81.

[19] Paul S and Donner A. Small sample performance of tests of homogeneity of odds ratios in K 2x2 table. Statistics in Medicine 1992; **11**:159-65.

[20] Hahn S, Williamson PR, Hutton JLGP, Flynn EV. Assessing the potential for bias in meta-analysis due to selective reporting of subgroup analyses within studies. Statistics in Medicine 2000; **19**:3325-36.

[21] Gelber RD and Goldhirsch A. Interpretation of results from subset analyses within overviews of randomized clinical trials. Statistics in Medicine 1987; **6**: 371-8.

[22] Yusuf S, Wittes J, Probstfielf J, Herman AT. Analysis and Interpretation of treatment effects in subgroups of patients in randomized clinical trials. JAMA 1991; **266**:93-8.

[23] Dickersin K, Chan S, Chalmers T. Publication bias and clinical trials. Controlled Clinical Trials 1987; **8**:343-53.

[24] Begg C and Berlin J. Publication bias: a problem in interpreting medical data. Journal of the Royal Statistical Society A 1988; **151**:419-63.

[25] Chalmers I. The Cochrane Collaboration: preparing, maintaining and disseminating systematic reviews of the effects of health care. Annals of the New York Academy of Science 1993; **703**:156-65.

[26] Stewart L and Clarke M. Practical methodology of meta-analyses (overviews) using updated individual patient data. Statistics in Medicine 1995; **14**:2057-79.

[27] Kaplan E and Meier P. Nonparametric estimation from incomplete observations. Journal of the American Statistical Association 1958; **53**:457-81.

[28] Mantel N and Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. Journal of the National Cancer Institute 1959; **22**:719-48.

[29] Collett D. Modelling Survival Data in Medical Research, First Edn. Great Britain: Chapman & Hall, 1994.

[30] Cox DR. Regression models and life-tables. Journal of the Royal Statistical Society Series B 1972; **34**: 187-220.

[31] Altman DG, De Stavola BL, Love SB, Stepnieweska KA. Review of survival analyses published in cancer journals. British Journal of Cancer 1995; **72**:511-8.

[32] Parmar MKB, Torri V, Stewart L. Extracting Summary Statistics to Perform Meta-analysis of the Published Literature for Survival End-points. Statistics in Medicine 1998; **17**:2815-34.

[33] Williamson PR, Tudur Smith C, Hutton J, Marson AG. Aggregate data meta-analysis with time-to-event outcomes. Statistics in Medicine 2002; **21**:3337-51.

[34] Tudur C, Williamson PR, Khan S, Best LY. The value of the aggregate data approach in meta-analysis with time-to-event outcomes. Journal of the Royal Statistical Society Series A 2001; **164**:357-70.

[35] Collette, L, Suciu, S, Bijnens, L, and Sylvester, R. Including literature data in individual patient data meta-analyses for time-to-event endpoints. *First symposium on systematic reviews: Beyond the basics.* 1998

[36] Pocock SJ, Clayton TC, Altman DG. Survival plots of time-to-event outcomes in clincal trials: Good practice and pitfalls. The Lancet 2002; **359**:1686-89.

[37] Armitage P, Berry G. Statistical Methods in Medical Research. Oxford: Blackwell Scientific Publications, 1994

[38] Mattson RH, Cramer JA, Collins JF. A comparison of valproate with carbamazepine for the treatment of complex partial seizures and secondarily generalized tonic-clonic seizures in adults. New England Journal of Medicine 1992; **327**:765-71.

[39] Marson AG, Williamson PR, Hutton JL, Clough HE, Chadwick DW. Carbamazepine versus valproate monotherapy for epilepsy (Cochrane Review). The Cochrane Library 2000; **Issue 3**, Oxford: Update Software.

[40] Hutton J, Cooke T, Pharaoh P. Life expectancy in children with cerebral palsy. British Medical Journal 1994; **309**:431-5.

[41] Khan S, Tudur Smith C, Williamson P, Sutton R. Shunts versus endoscopic therapy for long-term management of variceal haemorrhage (Cochrane Review). In: The Cochrane Library 2004; **Issue 4**: Chichester, UK: John Wiley & Sons, Ltd.

[42] Best L, Simmonds P, Baughan C *et al.* Palliative chemotherapy for advanced or metastatic colorectal cancer (Cochrane Review). In: The Cochrane Library 2000; **Issue 2**: Chichester, UK: John Wiley & Sons, Ltd.

[43] Galbraith RF. A note on the graphical presentation of estimated odds ratios from several clinical trials. Statistics in Medicine 1988; **7**:889-94.

[44] Boutitie F, Gueyffier F, Pocock SJ, Boissel JP. Assessing treatment-time interactions in clinical trials with time to event data: A meta-analysis of hypertension trials. Statistics in Medicine 1998; **17**:2883-903.

[45] De Silva M, MacArdle B, McGowan M *et al.* Randomised comparative monotherapy trial of phenobarbitone, phenytoin, carbamazepine, or sodium valproate for newly diagnosed childhood epilepsy. Lancet 1996; **347**:709-13.

[46] Verity CM, Hosking G, Easter DJ. A multicentre comparative trial of sodium valproate and carbamazepine in pediatric epilepsy. Developmental Medicine and Child Neurology 1995; **37**:97-108.

[47] Richens A, Davidson DLW, Cartlidge NEF, Easter DJ. A multicentre comparative trial of sodium valproate and carbamazepine in adult onset epilepsy. Journal of Neurology, Neurosurgery, and Psychiatry 1994; **57**:682-7.

[48] Heller AJ, Chesterman P, Elwes RDC *et al.* Phenobarbitone, phenytoin, carbamazepine, or sodium valproate for newly diagnosed adult epilepsy: a randomised comparative monotherapy trial. Journal of Neurology, Neurosurgery, and Psychiatry 1995; **58**:44-50.

[49] Williamson PR, Marson AG, Tudur C, Hutton JL, Chadwick DW. Individual patient data meta-analysis of randomized anti-epileptic drug monotherapy trials . Journal of Evaluation in Clinical Practice 2000; 6:205-14.

[50] Clarke M, Stewart L, Pignon J-P, Bijnens L. Individual patient data meta-analyses in cancer. British Journal of Cancer 1998; 77:2036-44.

[51] Early Breast Cancer Trialists' Collaborative Group. Treatment of early breast cancer, worldwide evidence 1985-90. 1990; 1.

[52] Stewart L and Parmar M. Meta-analysis of literature or of individual patient data: is there a difference? Lancet 1993; 341:418-22.

[53] Pignon J and Arriagada R. Role of thoracic radiotherapy in limited-stage small-cell lung cancer: quantitative review based on the literature versus meta-analysis based on individual data. Journal of Clinical Oncology 1993; 11:1819-20.

[54] Clarke M and Godwin J. Systematic reviews using individual patient data: A map for the minefields? Annals of oncology 1998; 9:827-33.

[55] Franxosi, MG, Santoro, E, and Santoro, L. Use of individual patient data versus published reports in a meta-analysis: the case of ace-inhibitors in myocardial infarction. *5th Annual Cochrane Colloquium*. 1997. Update Software, Oxford.

[56] Jeng G, Scott J, Burmeister L. A comparison of meta-analysis results using literature vs IPD: paternal cell immunisation for recurrent miscarriage. Journal of American Medical Association 1995; 274:830-6.

[57] Clarke M, Stewart L, Tierney J, Williamson P. Individual patient data meta-analyses compared with meta-analyses based on aggregate data [Protocol for Cochrane review]. The Cochrane Library 2001.

[58] Berry G, Kitchin RM, Mock PA. A comparison of two sample hazard ratio estimators based on the logrank test. Statistics in Medicine 1991; 10:749-55.

[59] Turner RM, Omar RZ, Yang M, Goldstein H, Thompson SG. A multilevel

model framework for meta-analysis of clinical trials with binary outcomes. Statistics in Medicine 2000; **19**:3417-32.

[60] Higgins JPT, Whitehead A, Turner RM, Omar RZ, Thompson SG. Meta-analysis of continuous outcome data from individual patients. Statistics in Medicine 2001; **20**:2219-41.

[61] Whitehead A, Omar RZ, Higgins JPT, Savaluny E, Turner RM, Thompson SG. Meta-analysis of ordinal outcomes using individual patient data. Statistics in Medicine 2001; **20**:2243-60.

[62] Yamaguchi T and Ohashi Y. Investigating centre effects in a multi-centre clinical trial of superficial bladder cancer. Statistics in Medicine 1999; **18**:1961-71.

[63] Andersen PK, Klein JP, Zhang M-J. Testing for centre effects in multi-centre survival studies: a monte-carlo comparison of fixed and random effects tests. Statistics in Medicine 1999; **18**:1489-500.

[64] Matsuyama Y, Sakamoto J, Ohashi Y. A Bayesian hierarchical survival model for the institutional effects in a multi-centre cancer clinical trial. Statistics in Medicine 1998; **17**:1893-908.

[65] Gustafson P. A bayesian analysis of bivariate survival data from a multicentre cancer clinical trial. Statistics in Medicine 1995; **14**:2523-35.

[66] Greenland S and Salvan A. Bias in the one-step method for pooling study results. Statistics in Medicine 1990; **9**:247-52.

[67] National Institute for Clinical Excellence. The diagnosis and the management of the epilepsies in adults and children in primary and secondary care. Draft for first consultation, 2003. http://www.nice.org.uk .

[68] Commission on Antiepileptic Drugs of the International League Against Epilepsy. Guidelines for clinical evaluation of antiepileptic drugs. Epilepsia 1989; **30**:400-8.

[69] Chadwick D. Valproate in the treatment of epilepsies. Epilepsia 1994; **Suppl 5**:S96-8.

[70] Nicholson A, Appleton R, Chadwick D, Smith D. The relationship between treatment with valproate, lamotrigine, and topiramate and the prognosis of the idiopathic generalized epilepsies. Journal of Neurology, Neurosurgery and Psychiatry 2004; **75**:75-9.

[71] Marson AG. Systematic reviews of randomized controlled trials of antiepileptic drugs. 2000. MD Thesis, University of Liverpool.

[72] Tudur Smith C, Marson AG, and Williamson PR. Carbamazepine versus phenobarbitone monotherapy for epilepsy (Cochrane Review). In: The Cochrane Library . **Issue 1**: 2003. Oxford: Update Software

[73] Tudur Smith C, Marson AG, Clough HE, Williamson PR. Carbamazepine versus phenytoin monotherapy for epilepsy (Cochrane Review). In: The Cochrane Library . **Issue 2**: 2002 Oxford: Update Software.

[74] Taylor S, Tudur Smith C, Williamson PR, Marson AG. Phenobarbitone versus phenytoin monotherapy for partial onset seizures and generalized onset tonic-clonic seizures (Cochrane Review). In: The Cochrane Library. **Issue 4**: 2001. Oxford: Update Software.

[75] Tudur Smith C, Marson AG, Williamson PR. Phenytoin versus valproate monotherapy for partial onset seizures and generalized onset tonic-clonic seizures (Cochrane Review). In: The Cochrane Library. **Issue 4**: 2001 Oxford: Update Software.

[76] Tudur Smith C, Marson AG, Williamson PR. Valproate versus phenobarbitone monotherapy for epilepsy [Cochrane protocol]. 2004.

[77] Muller MM, Marson AG, Williamson PR. Oxcarbazepine versus phenytoin monotherapy for epilepsy (Protocol for a Cochrane Review). In: The Cochrane Library. **Issue 3**: 2004. Oxford: Update Software.

[78] Gamble CL, Marson AG, Williamson PR. Lamotrigine versus carbamazepine monotherapy for epilepsy (Protocol for a Cochrane Review). In: The Cochrane Library. **Issue 3**: 2004. Oxford: Update Software.

[79] ILAE Commission on Antiepileptic Drugs. Considerations on designing clinical trials to evaluate the place of new antiepileptic drugs in the treatment of newly diagnosed and chronic patients with epilepsy. Epilepsia 1998; **39**(7); 799-803.

[80] Annegers J, Hauser W, Elveback L. Remission of seizures and relapse in patients with epilepsy. Epilepsia 1979; **20**:729-37.

[81] Hauser W. Incidence of epilepsy and unprovoked seizures in Rochester, Minisota 1935 through 1984. Epilepsia 1993; **34**:353-68.

[82] MacDonald B, Johnson A, Goodridge D, Cockerell O, Sander J, Shorvon S. Factors predicting prognosis of epilepsy after presentation with seizures. Annals of Neurology 2000; 48:833-41.

[83] Cockerell O, Johnson A, Sander J, Hart Y, Shorvon S. Remission of epilepsy: results from the National General Practice Study of Epilepsy [see comments]. Lancet 1995; **346**:140-4.

[84] Cockerell O, Johnson A, Sander J, Hart Y, Goodridge D, Shorvon S. Mortality from epilepsy: results from a prospective population-based study [see comments]. Lancet 1994; **344**:918-21.

[85] Hart Y, Sander J, Johnson A, Shorvon S. National General Practice Study of Epilepsy: recurrence after a first seizure [see comments]. Lancet 1990; **336**: 1271-4.

[86] Mattson RH, Cramer JA, Collins JF *et al.* Comparison of carbamazepine, phenobarbital, phenytoin, and primidone in partial and secondarily generalized tonic-clonic seizures. New England Journal of Medicine 1985; **313**:145-51.

[87] Brodie MJ, Overstall PW, Giorgi L. Multicentre, double-blind, randomised

comparison between lamotrigine and carbamazepine in elderly patients with newly diagnosed epilepsy. Epilepsy Research 1999; **37**:81-7.

[88] Whitehead A. Meta-Analysis of Controlled Clinical Trials. Chichester: Wiley, 2003.

[89] Lambert PC, Sutton AJ, Jones DR. A comparison of summary patient-level covariates in meta-regression with individual patient data meta-analysis. Journal of Clinical Epidemiology 2002; **55**:86-94.

[90] Williamson PR, Clough HE, Hutton JL, Marson AG, Chadwick DW. Statistical issues in the assessment of the evidence for an interaction between factors in epilepsy trials. Statistics in Medicine 2002; **21**:2613-22.

[91] Thompson SG and Sharp SJ. Explaining heterogeneity in meta-analysis: a comparison of methods. Statistics in Medicine 1999; **18**:2693-708.

[92] Aalen OO. Heterogeneity in survival analysis. Statistics in Medicine 1988; **7**:1121-37.

[93] Sargent DJ. A general framework for random effects survival analysis in the Cox proportional hazards setting. Biometrics 1998; **54**:1486-97.

[94] Vaupel JP, Manton KG, Stallard E. The impact of heterogeneity in individual frailty and the dynamics of mortality. Demography 1979; **16**:439-54.

[95] Hougaard P. Life table methods for heterogenous populations: Distributions describing the heterogeneity. Biometrika 1984; **71**:75-83.

[96] Therneau TM and Grambsch PM. Modeling Survival Data: extending the Cox model. New York: Springer, 2000.

[97] Therneau TM, Grambsch PM, Pankratz VS. Penalized Survival Models and Frailty. Technical Report 66. 2000.

[98] McGilchrist CA and Aisbett CW. Regression with frailty in survival analysis.

Biometrics 1991; **47**: 461-466.

[99] McGilchrist CA. REML Estimation for survival models with frailty. Biometrics 1993; **49**: 221-225.

[100] Breslow N. Contribution to the discussion of a paper by D.R.Cox. Journal of the Royal Statistical Society, Series B 1972; **34**:216-7.

[101] Efron B. The efficiency of Cox's likelihood function for censored data. Journal of the American Statistical Association 1977; **72**:557-65.

[102] Malafosse A, Genton P, Hirsch E, Marescaux C, Broglin D, Bernasconi R. Idiopathic Generalised Epilepsies. London: John Libbey and Company, 1994.

[103] Berlin JA, Santanna J, Schmid CH, Szczech LA, Feldman HI. Individual patient versus group-level data meta-regressions for the investigation of treatment effect modifiers: Ecological bias rears its ugly head. Statistics in Medicine 2002; **21**: 371-87.

[104] O'Quigley J and Stare J. Proportional hazards models with frailties and random effects. Statistics in Medicine 2002; **21**: 3219-3233.

[105] Walker AS and Babiker AG. A frailty test for heterogeneity of treatment effect in meta-analyses. 21st ISCB Conference Proceedings, Trento. 2000.

[106] Gray RJ. Tests for variation over groups in survival data. Journal of the American Statistical Association 1995; **90**:198-203.

[107] Marson AG, Kadir ZA, Hutton JL, Chadwick DW. The new antiepileptic drugs: A systematic review of their efficacy and tolerability. Epilepsia 1997; **38**:859-80.

[108] Bucher HC, Guyatt GH, Griffith LE, Walter SD. The results of direct and indirect treatment comparisons in meta-analysis of randomized controlled trials. Journal of Clinical Epidemiology 1997; **50**:683-91.

[109] Berkey CS, Anderson JJ, Hoaglin DC. Multiple-outcome meta-analysis of clinical

trials. Statistics in Medicine 1996; **15**:537-57.

[110] Dominici F, Parmigiani G, Wolpert R, Hasselblad V. Meta-analysis of migraine headache treatments: Combining information from heterogeneous designs. Journal of the American Statistical Association 1999; **94**:16-28.

[111] Hasselblad V. Meta-analysis of multitreatment studies. Medical Decision Making 1998; **18**: 37-43.

[112] Higgins JPT and Whitehead A. Borrowing Strength From External Trials In A Meta-analysis. Statistics In Medicine 1996; **15**:2733-49.

[113] Hirotsu C and Yamada Y. Estimating odds ratios through the connected comparative experiments. Communications in Statistics - Theory and Methods 1999; **28**:905-29.

[114] Fisher LD, Gent M, Buller HR. Active-control trials: How would a new agent compare with placebo? A method illustrated with clopidogrel, aspirin and placebo. American Heart Journal 2001; **141**:26-32.

[115] Gleser LJ and Olkin I. Meta-analysis for 2 x 2 tables with multiple treatment groups. In: Berry D, Editor. Meta-analysis in medicine and health policy. New York: Marcel Dekker, 2001.

[116] Hasselblad V and Kong DF. Statistical methods for comparison to placebo in active-control trials. Drug Information Journal 2001; **35**:435-49.

[117] Packer M, Antonopoulos GV, Berlin JA, Chittams J, Konstam MA, Udelson JE. Comparative effects of carvedilol and metroprolol on left ventricular ejection fraction in heart failure: Results of a meta-analysis. American Heart Journal 2001; **141**:899-907.

[118] Begg CB and Pilote L. A model for incorporating historical controls into a meta-analysis. Biometrics 1991; **47**:899-906.

[119] Li Z and Begg CB. Random effects models for combining results from

controlled and uncontrolled studies in a meta-analysis. Journal of the American Statistical Association 1994; 89:1523-7.

[120] Raghunathan TE. Pooling controls from different studies. Statistics in Medicine 1991; 10:1417-26.

[121] Baker SG and Kramer BS. The transitive fallacy for randomized trials: If A bests B and B bests C in separate trials, is A better than C? BMC Medical Research Methodology 2002; 2:13.

[122] Turnbull DM, Howel D, Rawlins MD, Chadwick DW. Which drug for the adult epileptic patient: phenytoin or valproate? British Medical Journal 1985; 290: 815-9.

[123] Craig I and Tallis R. Impact of valproate and phenytoin on cognitive function in elderly patients: results of a single-blind randomized comparative study. Epilepsia 1994; 35:381-90.

[124] Scottish Intercollegiate Guidelines Network. Diagnosis and Management of epilepsy in adults: A national clinical guideline. 2003 http://www.sign.ac.uk/

[125] Lumley T. Network meta-analysis for indirect treatment comparisons. Statistics in Medicine 2002; 21: 2313-2324.

[126] Thompson SG and Higgins JPT. How should meta-regression analyses be undertaken and interpreted? Statistics in Medicine 2002; 21:1559-73.

[127] Ades AE. A chain of evidence with mixed comparisons: models for multi-parameter synthesis and consistency of evidence. Statistics in Medicine 2003; 22: 2995-3016.

[128] Scurrah KJ, Palmer LJ, Burton PR. Variance components analysis for pedigree-based censored survival data using generalized linear mixed models (GLMMs) and gibbs sampling in BUGS. Genetic Epidemiology 2000; 19:127-148.

[129] Zahl P and Harris JR. Cancer incidence for Swedish twins studied by means of

bivariate frailty models. Genetic Epidemiology 2000; **19**:354-365.