# Combining Multiple Survival Endpoints within a Single Statistical Analysis

Zakiyah Zain (BEng., MSc.)

Submitted for the degree of Doctor of Philosophy

at Lancaster University

November 2011



ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 – 1346

# Declaration

This is to declare that the work in this thesis has been done by Zakiyah Zain and has not been submitted elsewhere for the award of a higher degree.

Zakiyah Zain

Professor John Whitehead

## Abstract

The aim of this thesis is to develop methodology for combining multiple endpoints within a single statistical analysis that compares the responses of patients treated with a novel treatment with those of control patients treated conventionally. The focus is on interval-censored bivariate survival data, and five real data sets from previous studies concerning multiple responses are used to illustrate the techniques developed.

The background to survival analysis is introduced by a general description of survival data, and an overview of existing methods and underlying models is included. A review is given of two of the most popular survival analysis methods, namely the logrank test and Cox's proportional hazards model. The global score test methodology for combining multiple endpoints is described in detail, and application to real data demonstrates its benefits.

The correlation between two score statistics arising from bivariate interval-censored survival data is the core of this research. The global score test methodology is extended to the case of bivariate interval-censored survival data and a complementary log-log link is applied to derive the covariance and the correlation between the two score statistics. A number of common scenarios are considered in this investigation and the accuracy of the estimator is evaluated by means of extensive simulations.

An established method, namely the approach of Wei, Lin and Weissfeld, is examined and compared with the proposed method using both real and simulated data. It is concluded that our method is accurate, consistent and comparable to the competitor. This study marked the first successful development of the global score test methodology for bivariate survival data, employing a new approach to the

derivation of the covariance between two score statistics on the basis of an interval-censored model. Additionally, the relationship between the jackknife technique and the Wei, Lin and Weissfeld method has been clarified.

## Acknowledgements

I would like to say thank you to John Whitehead for his thorough and excellent supervision throughout this research. As for his always being there to help, I am so grateful for his patience and wisdom, not to mention the humour that has kept me sane. Both his professionalism and his personality have inspired me in many ways; a great supervisor he was, and a lifetime guru he is.

To my supportive and understanding family back home in Malaysia: my dear husband (best friend), Jasni, for trying his best at single parenting, and our wonderful girls, Afiqah and Damiya, who have grown so much so fast (while I could see them only on Skype) and to my beloved parents: Zain and Halimah for their endless prayers and encouragement.

I would like to thank also Paul Siney of Wrightington Hospital and Debbie Costain for providing the Young Patient's Hip study data. Many thanks are due also to the examiners, Alan Kimber and Debbie Costain, for their constructive feedback, and Alan Airth for his kind help with the English language. Gratitude and appreciation go to the Ministry of Higher Education (MOHE, Malaysia) and the Universiti Utara Malaysia (UUM) for the financial support.

# Preface

## Background

In clinical trials, the main purpose is often to compare efficacy between experimental and control treatments. These treatment comparisons often involve several responses or endpoints, and this situation complicates the analysis. For example, sets of responses concerned with survival times in a single clinical trial include: time to first cardiac event and time to death from any cause; time to loss of vision in the left eye and time to loss of vision in the right eye; and times from entry to a trial until the first, the second and the third asthma exacerbations.

One approach to simplifying the analysis would be to choose one of the survival times and to use that alone as a single primary endpoint. This is not always desirable, for example, in cases where the choice would be rather subjective or where the endpoints are of equal interest. A single parameter relating, to an overall assessment, is often required to give a solid justification of treatment advantage, and so separate analyses of more than one endpoint might likewise not be appropriate. Another alternative is to use a composite endpoint, such as the time until the first of the events.

The cumulative treatment advantage is usually measured by the score statistic for each endpoint. In survival analysis, the logrank test is one of the most popular methods for testing the equality of two treatment groups. It is routinely used in the analysis of clinical trials comparing the time-to-event distribution of a group of patients randomised to an experimental treatment with that of a control group. When prognostic factors are to be adjusted for, Cox's proportional hazards regression, which is a direct generalisation of the logrank test, is commonly employed. These two methods are extensively referred to throughout this thesis. In the case of bivariate

survival data, the score statistics can be summed directly, but the variance is now affected by the dependence structure between two endpoints.

To estimate the correlation coefficient, an approximate formula for the covariance between the two score statistics is derived in various settings of bivariate interval-censored survival data.

**Motivation**

In general, global test methodology can be defined as the use of a combined model to estimate a composite measure of treatment effect concerning multiple outcomes. A global null hypothesis, that the treatment has no effect on any of a number of patient responses, is tested. Global test methodology has been used successfully in major clinical trials involving binary data when multiple outcomes are concerned. To my knowledge thus far, it has been accepted only in stroke studies, for its ability to yield a single parameter of treatment advantage, which is easily interpreted, as well as for its cost-saving benefit in terms of the trial size.

In particular, use of a global test as a primary analysis for multiple binary outcomes, accompanied by secondary tests of individual outcomes, was implemented in the NINDS t-PA Stroke Trial (Tilley et al., 1996). Global testing was adopted also for the International Citicoline Trial in acUte Stroke (ICTUS) documented by Davalos (2007). Moreover, Bolland et al. (2009) concluded for larger samples that global tests gave accurate type I error rates and satisfactory power, even after adjustment for prognostic factors. Therefore, the global testing approach is attractive for research concerning situations in which two or more time-to-event responses are observed on each individual.

Previous work has successfully determined the correlation between two score statistics arising from binary data or from ordered categorical data (Whitehead, Branson and Todd, 2010), but the case of survival data has proved difficult. An existing method for combining two or more survival analyses is the method of Wei, Lin and Weissfeld (1989). Unlike the logrank test, their approach does not directly condition on risk sets and does not reproduce the familiar form of logrank variance. There would appear to be scope for increasing its power by taking advantage of conditioning on successive risk sets.

An earlier approach using the logrank test proved difficult (Whitehead et al., unpublished) and therefore a new strategy is now proposed. In this new approach, the survival data are summarised within categories and analysed as interval-censored survival data. Using such a formulation, it is possible to determine the correlation, which serves as an accurate approximation to the correlation of the logrank statistics. Correlations between score statistics arising from interval-censored forms of the Cox model are investigated. Once an estimate for the correlation between two score test statistics is available, it has many applications. For example, combined null hypotheses, testing whether a linear combination of effects is equal to zero, and global null hypotheses, testing whether all effects are equal to zero, can be addressed. Joint confidence regions for multiple hazard ratios can be determined. Multiple testing procedures and sequential testing approaches can be implemented.

## Aims

The aim of this thesis is to develop a methodology for combining multiple endpoints within a single statistical analysis focussing upon bivariate interval-censored survival data. A complete procedure is described by which to derive an estimator for the covariance of two score statistics and hence the correlation. Estimates of overall treatment effect, adjusting for the correlation, are also derived. Applications to real data and simulation studies will be performed. With the knowledge of such a correlation, an investigation into some of its various uses will be carried out. Most importantly, a detailed comparison between the present method and that of Wei, Lin and Weissfeld (1989) will be made.

## Outline

This thesis is divided into seven chapters. The first chapter introduces survival data and analysis, and provides an overview of the existing methods in general. The fundamental concepts of the logrank test, proportional hazards, and the derivation of the score statistics are given. A basic study design is constructed as a main reference for later use. In Chapter 2, the global test methodology central to this research study is described for binary data, with illustration using real data from previous clinical trials. Simulation is performed to investigate the accuracy of the method.

Chapter 3 focuses on the methods needed for analyzing interval-censored survival data, describing some model applications and providing a comparison of methods. These methods are illustrated using a real data set and the accuracy of each method is evaluated in a simulation study. In Chapter 4, the core component of this research is covered, namely the correlation between two score statistics. Bivariate survival data of various types and their associated modelling are described. Global

testing methodology is then applied to the case of main interest, bivariate interval-censored survival data. Using the selected model, the covariance between two statistics is estimated and hence an approximate formula for the correlation is made available. The methodology developed is illustrated using three real data sets.

Chapter 5 then investigates the accuracy of the proposed method using extensive simulation of six different cases, which are selected on the basis of their practical importance. The design of the simulation study is described at great length, with key points and limitations carefully annotated. Meanwhile, Chapter 6 presents the established method of Wei, Lin and Weissfeld from a different perspective compared to that in the original paper. Specifically, influence diagnostics are described and illustrated with clarity. To form the basis of comparison, the theories and computation are examined in detail. The same data sets (real and simulated) are analyzed using this method and compared to the results from our proposed method (Chapters 4 and 5). Finally, conclusions are drawn and potential areas for further work are put forward in Chapter 7 of this thesis.

# TABLE OF CONTENTS

**CHAPTER 1.BACKGROUND OF SURVIVAL ANALYSIS ..... 1**

1.1. INTRODUCTION TO SURVIVAL DATA ..... 1

*1.1.1. Special Features of Survival Data ..... 5*

*1.1.2. Survivor Function and Hazard Function ..... 6*

*1.1.3. Some Parametric Hazard Functions ..... 8*

*1.1.4. Censoring Mechanisms ..... 10*

1.2. UNIVARIATE SURVIVAL DATA ..... 13

1.3. INTERVAL-CENSORED SURVIVAL DATA ..... 14

1.4. SCORE STATISTIC AND FISHER'S INFORMATION ..... 16

*1.4.1. Z and V for Binary Data ..... 18*

*1.4.2. Z and V for Survival Data ..... 22*

1.5. PROPORTIONAL HAZARDS ..... 23

1.6. LOGRANK TEST ..... 25

1.7. SAMPLE SIZE CALCULATION ..... 28

**CHAPTER 2.GLOBAL TEST METHODOLOGY ..... 31**

2.1. INTRODUCTION TO GLOBAL TEST (BINARY DATA) ..... 31

*2.1.1. Global Z and V Approach ..... 33*

2.2. COMBINING MULTIPLE BINARY OUTCOMES ..... 34

*2.2.1. Analysis of Data from Trials of Citicoline ..... 37*

2.3. SAMPLE SIZE FOR A CLINICAL TRIAL ..... 39

2.4. SIMULATION AND RESULTS ..... 43

2.5. DISCUSSION ..... 51

**CHAPTER 3.METHODS FOR INTERVAL-CENSORED SURVIVAL DATA .... ..... 53**

3.1. MODELLING OF SURVIVAL DATA ..... 54

*3.1.1. Cox's Proportional Hazards Regression Model ..... 55*

*3.1.2. Modelling Interval-censored Survival Data ..... 64*

*3.1.3. Methods Using Log -Odds ratio Transformation ..... 67*

*3.1.4. Methods Using Complementary Log-Log Transformation ..... 69*

*3.1.5. Summary of Methods* ..... *73*
3.2. APPLICATION TO BONE MARROW TRANSPLANT DATA ..... 76
3.3. POWER SPECIFICATION OF A CLINICAL TRIAL ..... 79
3.4. SIMULATION AND RESULTS ..... 80
3.5. DISCUSSION ..... 84

**CHAPTER 4. THE CORRELATION BETWEEN TWO SCORE STATISTICS ..... 85**

4.1. BIVARIATE SURVIVAL DATA ..... 86
4.2. ADAPTATION OF THE GLOBAL TEST TO INTERVAL-CENSORED SURVIVAL DATA .. ..... 87
*4.2.1. Bivariate Interval-censored Survival Data* ..... *89*
4.3. THE CORRELATION BETWEEN TWO SCORE STATISTICS ..... 91
*4.3.1. Derivation of Estimators for Covariance and Correlation* ..... *92*
*4.3.2. Paired Organs* ..... *96*
*4.3.3. Generally Related Events or Indicators* ..... *97*
*4.3.4. Progression-Free Survival* ..... *99*
4.4. ESTIMATION OF AN OVERALL TREATMENT EFFECT ..... 100
4.5. APPLICATION TO REAL DATA: NON-RECURRENT EVENTS ..... 102
*4.5.1. Paired Organs: Hip Replacement Revision* ..... *103*
*4.5.2. Related Indicators and PFS: Cancer Data* ..... *113*
4.6. RECURRENT EVENTS ..... 118
*4.6.1. Key Model Components* ..... *119*
*4.6.2. Application to Bladder Cancer Data* ..... *124*
4.7. DISCUSSION ..... 132

**CHAPTER 5. SIMULATION STUDY ..... 133**

5.1. DESIGN OF SIMULATION STUDY ..... 133
*5.1.1. Key Performance Measures* ..... *142*
*5.1.2. Combined Hypothesis Tests* ..... *144*
5.2. SIMULATION AND RESULTS ..... 145
*5.2.1. Complete or Uncensored Data* ..... *145*
*5.2.2. Paired Organs* ..... *149*
*5.2.3. Related Indicators* ..... *150*
*5.2.4. Progression-Free Survival* ..... *151*

*5.2.5. Recurrent Events (Total Time)* ..... *153*
*5.2.6. Recurrent Events (Gap Time)* ..... *155*
*5.2.7. Summary of Correlation Ratios* ..... *157*
*5.2.8. Overall Results: Five vs Ten Intervals* ..... *159*
5.3. DISCUSSION ..... 162

**CHAPTER 6. COMPARISON TO THE WEI, LIN AND WEISSFELD METHOD ..... 165**
6.1. BACKGROUND ..... 166
*6.1.1. Marginal Models* ..... *167*
*6.1.2. Jackknife Method* ..... *170*
*6.1.3. Exact Delta Beta and its Approximation (DFBETA)* ..... *175*
*6.1.4. Comparison of Influences: Jackknife, Delta-beta and DFBETA* ..... *181*
6.2. THE WEI, LIN AND WEISSFELD (WLW) METHOD ..... 184
6.3. ESTIMATION OF AN OVERALL TREATMENT ADVANTAGE: WLW ..... 193
6.4. THEORETICAL COMPARISON: ZW VS. WLW ..... 196
6.5. APPLICATIONS TO REAL DATA: ZW VS. WLW ..... 199
*6.5.1. Recurrent Events: Bladder Cancer Data* ..... *199*
*6.5.2. Paired Organs: Hips Replacement Revision* ..... *203*
*6.5.3. Generally Related Indicators and PFS: Cancer Data* ..... *204*
6.6. SIMULATION AND RESULTS ..... 205
*6.6.1. Complete or Uncensored* ..... *206*
*6.6.2. Paired Organs* ..... *208*
*6.6.3. Related Indicators* ..... *209*
*6.6.4. Progression-Free Survival* ..... *210*
*6.6.5. Recurrent Events TT* ..... *211*
*6.6.6. Recurrent Events GT* ..... *212*
*6.6.7. Summary of Correlation Ratios* ..... *213*
6.7. OVERALL RESULTS: WLW VS ZW ..... 217
6.8. DISCUSSION ..... 219

**CHAPTER 7. CONCLUSIONS AND FURTHER WORK ..... 222**
**APPENDIX ..... 225**
**REFERENCES ..... 235**

## LIST OF TABLES:

Table 1.1: A 2 x 2 contingency table of binary response data (failure, success) in a two-group clinical trial comparing experimental and control. ..... 18
Table 1.2: A 2x2 contingency table of univariate survival responses at time $t_a$. ..... 22
Table 2.1: Counts of subjects for combined binary outcomes on two scales. ..... 35
Table 2.2: Summary of successes on individual and multiple scales from a meta-analysis study by Davalos et al. (2002). ..... 38
Table 2.3: Summary of the success probabilities, calculated at $\theta = 0.231$ for individual scales, with their corresponding sample sizes. ..... 41
Table 2.4: Summary of the success probabilities ($\theta = 0.231$), correlations and the corresponding values of $c_{uv}$ and $b_{uv}$ for two-scale. ..... 42
Table 2.5: Summary of the success probabilities calculated at $\theta = 0.231$ for individual and combined scales, with their corresponding sample sizes, $n$. ..... 43
Table 2.6: Simulation results under the null and alternative hypotheses for individual scales, each with a different sample size, $n$. ..... 45
Table 2.7: A 2 x 2 contingency table of outcomes for patients on placebo for the combined BI_mRS scales. ..... 46
Table 2.8: A 2 x 2 contingency table for combined outcomes on BI_mRS scales for patients on citicoline. ..... 47
Table 2.9: The probability for each combined outcome for patients on citicoline and placebo, under the alternative ($\theta = 0.231$). ..... 48
Table 2.10: Simulation results for all two-scale and three-scale combinations under the null and alternative hypotheses. ..... 48
Table 2.11: Summary of simulation results showing the power for each individual and combined scales under the alternative, with sample size $n = 3000$. ..... 49
Table 2.12: Summary of the p-values and powers from the simulations using unequal $\theta_u$ values, with a fixed sample size, $n = 3000$ ..... 50
Table 3.1: Statistics for a parallel group study with interval-censored survival responses (Whitehead, 1997). ..... 64
Table 3.2: A 2 x 2 contingency table for survival responses at a defined interval for experimental and control groups. ..... 65
Table 3.3: Summary of the various methods used to derive $Z$ and $V$: the derived formulae included. ..... 75

Table 3.4: Summary of outcomes for bone marrow transplant trial (Storb et al., 1986), used as illustration of interval-censored survival data..... 77
Table 3.5: Summary of the cumulative score statistic, $Z$ and information, $V$, using the five methods numbered 1 to 5 (Table 3.3)..... 78
Table 3.6: Sample sizes, $r$, determined from the power calculation approach, for each of the five methods..... 80
Table 3.7: Results for the average p-value and type I error using each of the five methods, simulated under the null for each sample size $r$..... 81
Table 3.8: Results for the average p-value and power using each of the five methods, simulated under the alternative for each sample size $r$..... 82
Table 3.9: Results for the average p-value and power using each of the five methods simulated under the alternative (half the treatment advantage), for each sample size, $r$. ..... 83
Table 4.1: A 2 x 2 contingency table of censored bivariate data for patients on control for each time interval..... 90
Table 4.2: Summary for hip replacement revision data of 342 bilateral patients, stratified by hip (left and right)..... 108
Table 4.3: Count of failures and numbers at risk for both patients on $E$ and $C$ for each pair of intervals $ij$ relating to both $T_1$ and $T_2$, and those relating to individual $i$ and $j$ intervals, for the hip replacement revision data..... 109
Table 4.4: Calculated values for covariance, score statistic, and Fisher's information for the hip replacement revision data, using $k = 2$ intervals..... 110
Table 4.5: Results for the hip replacement revision data, using different number of intervals ($k = 2$, 5 and 10)..... 111
Table 4.6: Example of data sets for related indicators and PFS. *(* indicates censoring)* ..... 115
Table 4.7: Summary of the cancer data for the indicators and PFS analyses..... 115
Table 4.8: Results for the cancer data using 2, 5 and 10 intervals based on the analysis of related indicators..... 116
Table 4.9: Results for the cancer data using 2, 5 and 10 intervals based on PFS. ... 117
Table 4.10: Tumour recurrence data extracted from Wei, Lin & Weissfeld (1989), presented in total time ($T_1$, $T_2$) and gap time ($T_{2G}$)..... 125
Table 4.11: Summary of outcomes for the bladder cancer data..... 128

Table 4.12: Results for bladder cancer data analyzed @ 2, 5 and 10 intervals (TT). ..... 128
Table 4.13: Results for bladder cancer data analyzed @ 2, 5 and 10 intervals (GT). 130
Table 5.1: Average values for $V^*$ and $b$ computed at various sample sizes for the complete data ($d$ =10, $k$ = 5,). ..... 138
Table 5.2: Example of simulation setting for each case at $d$ = 10 and $k$ = 5 intervals. ..... 139
Table 5.3: Average values of $V$ and var $Z$ for the complete case under the null and alternative hypotheses. ..... 146
Table 5.4: Summary of results for the complete case under the null and alternative, using 5 and 10 intervals. ..... 147
Table 5.5: Summary of results for the paired case under the null and alternative. ... 149
Table 5.6: Results for the indicators case under the null and alternative. ..... 150
Table 5.7: Summary of results for the PFS case under the null and alternative. ..... 152
Table 5.8: Results for the recurrent events TT case under the null and alternative. . 154
Table 5.9: Results for the recurrent events GT case under the null and alternative.. 156
Table 6.1: Various estimated quantities for the bladder cancer data using the methods of jackknife, exact delta-beta and WLW for **total time** (TT). ..... 192
Table 6.2: Various estimated quantities for the bladder cancer data using the methods of jackknife, exact delta-beta and WLW for **gap time** (GT). ..... 193
Table 6.3: Results for the bladder cancer data using total time (TT) for WLW and ZW. ..... 200
Table 6.4: Results for the bladder cancer data using gap time (GT) for WLW and ZW *(ZW results correspond to Table 4.13 and WLW to Table 6.2 earlier).* ..... 201
Table 6.5: Results for the hip replacement revision data using WLW and ZW ..... 203
Table 6.6: Results for the cancer data using WLW and ZW (5 intervals) ..... 204
Table 6.7: Simulation results for the complete case under the null and alternative using WLW and ZW *(ZW results correspond to Table 5.4).* ..... 207
Table 6.8: Simulation results for the paired case under the null and alternative using WLW and ZW *(ZW results correspond to Table 5.5).* ..... 208
Table 6.9: Simulation results for the indicators under the null and alternative using WLW and ZW *(ZW results correspond to Table 5.6).* ..... 209
Table 6.10: Simulation results for the PFS under the null and alternative using WLW and ZW *(ZW results correspond to Table 5.7).* ..... 210

Table 6.11: Simulation results for recurrent events TT under the null and alternative using WLW and ZW *(ZW results correspond to Table 5.8)*..... 211
Table 6.12: Simulation results for recurrent events GT under the null and alternative using WLW and ZW *(ZW results correspond to* Table 5.9*)*..... 212

## LIST OF FIGURES:

Figure 1.1: An example of study time for five patients in a 10-year study period ..... 2
Figure 1.2: An example of patient time, corresponding to the study times for the five patients in Figure 1.1. ..... 3
Figure 1.3: Four common types of censoring in survival data..... 10
Figure 4.1: Various censorings for the paired case, an example for $T_1$ and $T_2$ of left and right hips respectively. ..... 96
Figure 4.2: Example of possible outcomes for the first failure, as measured by $T_1$ for generally related indicators. ..... 98
Figure 4.3: A schematic diagram of the basic components of Charnley's LFA. ..... 104
Figure 4.4: Survival distribution for time to left hip revision, $T_1$ in years, stratified by the cup position (cup_pos). ..... 105
Figure 4.5: Survival distribution for time to right hip revision, $T_2$ in years, stratified by cup position (cup_pos)..... 106
Figure 4.6: Survival distribution function for time to disease progression, $T_1$ in days, for the cancer data, stratified by treatment group (trt = 1, 2). ..... 113
Figure 4.7: Survival distribution function for time to death, $T_2$ in days, for the cancer data, stratified by treatment group. ..... 114
Figure 4.8: Example of three patients with recurrent events. ..... 120
Figure 4.9: Total time risk intervals for the scenario depicted in Figure 4.8. ..... 121
Figure 4.10: Gap time risk intervals for the scenario depicted in Figure 4.8. ..... 122
Figure 4.11: Survival distribution for time to first recurrence, $T_1$ in months, for the bladder cancer data, stratified by treatment group. ..... 126
Figure 4.12: Survival distribution for time to second recurrence, $T_2$ (total time) in months, for the bladder cancer data, stratified by treatment group. ..... 127
Figure 4.13: Survival distribution for time to second recurrence, $T_{2G}$ (gap time) in months, for the bladder cancer data, stratified by treatment group. ..... 127
Figure 5.1: Process flow charts for analysis of real data and simulation run ..... 136
Figure 5.2: The correlation ratio for each case at 5 and 10 intervals. ..... 158
Figure 5.3: Type I error rates, power and correlation ratio using 5 and 10 intervals. ..... 160
Figure 6.1: Schematic of the jackknife algorithm for estimating parameter, inspired by the bootstrap algorithm (Efron & Tibshirani, 1993 p48). ..... 171

Figure 6.2: Plot of the influence using the jackknife, exact and DFBETA methods for treatment against patient id for the bladder cancer data ..... 182
Figure 6.3: Part 1: SAS codes and output for analysis of the bladder cancer data using WLW. ..... 187
Figure 6.4: Part 2: SAS codes for summing DFBETA from the output of WLW earlier. ..... 189
Figure 6.5: Part 3: SAS codes and output for WLW to compute the estimated covariance and correlation *(Courtesy of Thomas Hamborg).* ..... 190
Figure 6.6: Plots of the sample correlation $\rho_{(\text{sample})}$ versus the derived correlation $\rho_{(\text{est})}$ for the six cases using WLW and ZW ($k = 5$). ..... 214
Figure 6.7: Correlation ratio against censoring proportion (percentage) for recurrent TT and recurrent GT ($d = 10$, under $H_0$) ..... 215
Figure 6.8: Type 1 error rates, powers and correlation ratios of ZW versus WLW. 217

# Chapter 1. Background of Survival Analysis

This chapter begins with an introduction to survival data, in particular, covering some special features of survival data such as censoring, which complicate the analysis. A theoretical description of the survivor function, hazard function, and their relationship is given in subsequent sections. In order to appreciate the distinct censoring mechanisms involved, illustrations and examples are presented.

Section 1.2 describes univariate survival data with a review of some well known statistical methods. Some issues pertaining to interval-censored survival data and existing methods for their analysis are introduced in the following section. Section 1.4 describes the score statistic, $Z$, and Fisher's information, $V$, as derived from the likelihood function. In Section 1.5, the proportional hazards assumption is reviewed; its application in the context of interval-censored survival data is illustrated in subsequent chapters.

The logrank test, which is one of the most widely used methods for testing the equality of event times of two groups, is described in Section 1.6. Finally, the basic principles of sample size determination in clinical trials are described in Section 1.7 for later reference.

## 1.1. Introduction to Survival Data

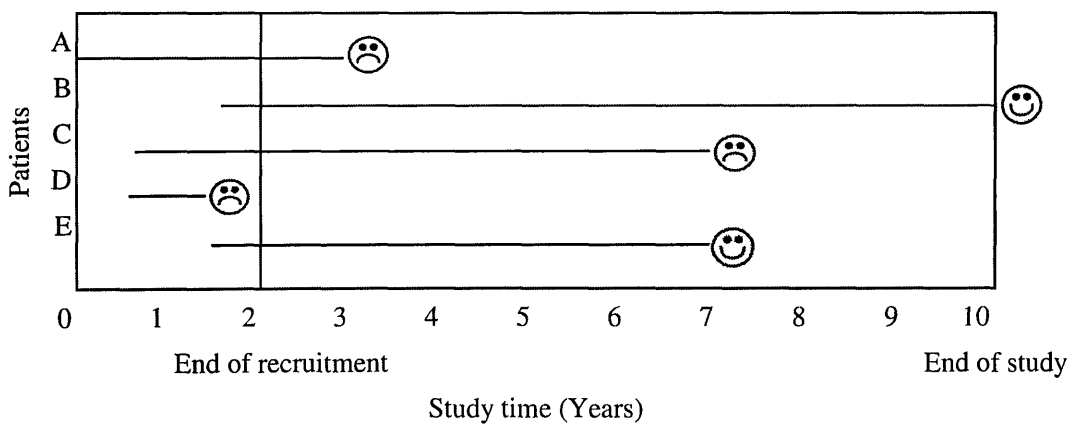
Survival data is a common term used for describing data which measure time to a certain event or endpoint. Originally, the term "survival" arose because in early developments, the event being referred to was death. However, in medical applications, an event may refer to an occurrence of a disease or condition, disease progression, tumour recurrence, and so on. In general, an event can be defined as a

transition from one state to another that can be indexed in time, for example, a transition from a state of being healthy to a state of being infected.

Apart from clarifying the definition of event, the origin of time and its measurement scale also need to be specified. For example, surgery time is taken as the time origin $t_0$, where $t = 0$, if the investigator is interested in the study of survival after surgery. Meanwhile, in clinical trials, it is the time of randomization to treatment that is usually designated $t_0$. The actual study time varies due to different entry times (say time origin $t_0$), as illustrated in Figure 1.1 (inspired by Collett, 2003 p3).

Figure 1.1: An example of study time for five patients in a 10-year study period

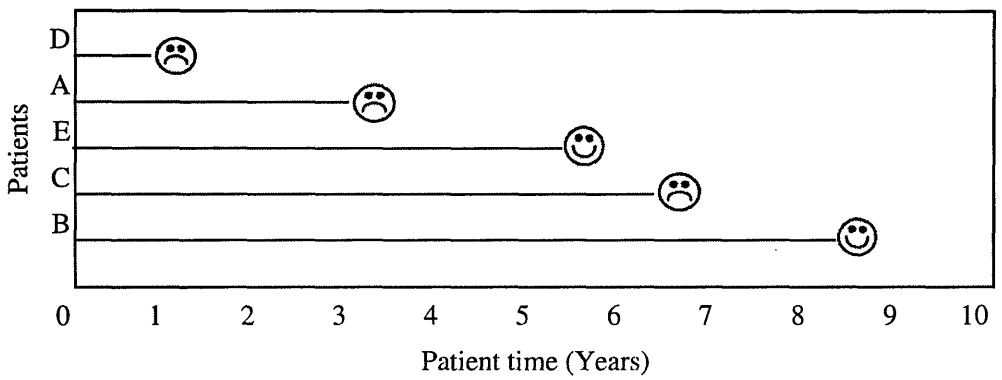
(☹ indicates died, ☺ indicates survived).




In a study involving survival outcomes, patients are observed until they reach a defined endpoint (for example, death). However, patients sometimes withdraw from a study, or the study is concluded before all patients reach the endpoint. In Figure 1.1, five patients (A to E) were recruited at various times over the first 2 years of a study which ended 10 years after the start of the study. Patients A, C and D died during the study, while patient B was still alive at the end of the study. Patient E was lost to follow-up and last known to be alive at 7 years after the start of the study. For each patient, the study begins at time $t_0$ which is the patient's time origin, and the time the

patient remains in the study is termed "patient time". This type of non-parallel data with regard to the time origin is quite common in survival data. Figure 1.2 illustrates the survival times in ascending order, of the same five patients, based on patient time (inspired by Collett, 2003 p4).

Figure 1.2: An example of patient time, corresponding to the study times for the five patients in Figure 1.1.




The transformation of data from study time to patient time is shown in Figures 1.1 and 1.2. For example, patient C who was recruited at about 6 months from the beginning of the study, failed after 7 years of the study time, therefore the actual patient time spent in the study is 6.5 years, as illustrated in Figure 1.2 above. It is to be clarified that in this thesis, patient time is used unless noted otherwise. The choice of time unit varies depending on the type of study and context. For example, in a radical type of cancer, the time may be weeks or months, whereas the length of treatment for a cardiovascular trial may be measured in years.

As shown in Figure 1.1, for patients B and E, the survival times, also known as failure times or event times, are censored: patients survived to a certain time beyond which their status is unknown. In the analysis of survival data, the absence of an event time is referred to as censoring and the patient for whom no event time is available is

referred to as censored. For a patient $i$, the survival time, $T_i$, is defined as the minimum time observed, either the failure time or death time, $D_i$, or the censored time, $C_i$; this definition is often represented by $T_i = \min(D_i, C_i)$. The most common type of censoring encountered in medical studies is right-censoring, followed respectively by interval-censoring and left-censoring of survival data. These censoring mechanisms are described in Section 1.1.4.

Theoretically, survival data are probabilistic in nature; the times at which events occur being assumed to be realizations of some random process. By definition, this means that, for a given individual, the time $T$ to the event is a random variable which has a probability distribution. Therefore, survival data can be described in terms of a cumulative distribution function (c.d.f.), probability density function (p.d.f.), survivor function or hazard function. The survivor function, $S(t)$, and the hazard function, $h(t)$, are paramount in survival data; their definitions and relationships are described in Section 1.1.2.

A good review of the development of survival analysis throughout the 20$^{th}$ century is provided by Oakes (2001), with emphasis on work since 1980. He regards the two landmark papers by Kaplan & Meier (1958), who formalized the product-limit estimator, and that by Cox (1972), who introduced the proportional hazards model, as primarily responsible for the present emphasis. The concepts of the latter are described and applied in several sections in this thesis.

### 1.1.1. Special Features of Survival Data

In general, survival data have three special features that are difficult to handle and that render conventional statistical methods inadequate: (i) censoring, (ii) non-normality and (iii) time dependence of covariates, which will now be described in turn.

Consider a study of patients with stage 3 or stage 4 prostatic cancers, randomized to form experimental and control treatment groups. The number of days between randomization until death is recorded, as well as the cause of death. A simple logistic regression can be used to analyze the current status of the patients: whether they are dead or alive after some specified follow-up time, say five years. However, this method ignores vital information on the timing of death if it occurs. Due to the nature of survival studies which typically involve a long period of study time, it is common to encounter censored observations. Discarding censored data may be a tempting option for simplicity, but it may only work if the proportion censored is small. Censored observations contain information about survival and thus should be accounted for in analysis. For example, in a cancer study, an observation censored at 15 years indicates better survival compared to that censored at 1 year.

In a survival study, a few individuals may experience the event much sooner or later than the majority of individuals under study, hence giving the survival distribution a skewed appearance and methods based on the normal distribution being unsuitable for use. Commonly used distributions are often either symmetric or right skewed, but survival distributions in many cases involve left skewed distribution of positive variables (Hougaard, 1999).

The most common goal of a clinical study is to determine a treatment effect. Additionally, when a study aims to estimate causal or predictive model parameters, in which the risk of an event depends on covariates, the analysis becomes more

interesting. While some covariates such as race and gender remain constant throughout the study period, some covariates may change with time, for example marital status and employment. The latter are called time-dependent covariates.

All methods of survival analysis should allow for censoring and non-normality. They may also need to accommodate time-dependent covariates. Information in the censored and uncensored observations should be combined to devise a procedure that can provide consistent estimates of the parameters of interest. This is often accomplished by the methods of maximum likelihood or partial likelihood, which can also be adjusted to incorporate the time-dependent covariates. Such methods are briefly described in Sections 1.4 and 1.5.

### 1.1.2. **Survivor Function and Hazard Function**

An analysis of survival data requires special techniques because the data are almost always incomplete. There exist many models for survival data, and each model is distinguished by its choice of the probability distribution for $T$, the non-negative random time variable, $f(t)$. The distribution of survival times can be summarized by a survivor function $S(t)$, which is the probability that the event occurs after time $t$: $S(t) = P(T > t)$, $0 < t < \infty$. Assuming $T$ is a continuous variable, its cumulative distribution function (c.d.f) gives the probability that the variable $T$ will be less than or equal to any value of $t$: $F(t) = P(T \leq t)$. The survivor function can thus be expressed in terms of c.d.f. of $T$: $S(t) = 1 - F(t)$. The slope or derivative of the c.d.f. gives the p.d.f, which can be written as $f(t) = dF(t)/dt = -dS(t)/dt$. Therefore, the c.d.f. and the survivor function respectively can be re-written as

$$F(t) = P(T \leq t) = \int_0^t f(u)du,$$

and

$$S(t) = 1 - P(T \le t) = 1 - \int_0^t f(u)du.$$

The hazard function $h(t)$ is defined as

$$h(t) = \lim_{\delta t \to 0} \frac{P(t \le T < t + \delta t \mid T \ge t)}{\delta t} = \lim_{\delta t \to 0} \frac{P(t \le T < t + \delta t)}{P(T \ge t)\delta t} = \lim_{\delta t \to 0} \frac{F(t + \delta t) - F(t)}{S(t)\delta t}, \tag{1.1}$$

where $\delta t$ is the time interval within which the subject fails, conditional on having survived up to time $t$. Therefore, $h(t) = f(t)/S(t)$ and $h(t) = -\frac{d}{dt}\log S(t)$. This implies that $f(t) = h(t)\exp\left\{-\int_0^t h(u)du\right\}$ which is the product of the hazard function and exponential of the minus accumulated hazards until time $t$, also known as the cumulative hazard. Therefore, the survivor function can be expressed as

$$S(t) = \exp\left\{-\int_0^t h(u)du\right\}. \tag{1.2}$$

The cumulative hazard function is also conventionally denoted by $H(t)$; thus $S(t) = \exp(-H(t))$. The survivor function has the following properties; (i) the probability of survival at time zero is 1: $S(0) = 1$, (ii) the probability of infinite survival is zero, $S(\infty) = 0$, and (iii) survivor function is non-increasing.

In survival analysis, discrete random variables often arise due to the rounding off of measurements, for example, time measured to the nearest week or month. Another situation arises where failure times are grouped into intervals in studies where patients are not monitored continuously, but rather are followed-up only at pre-defined time points. In this thesis, grouping of failure times into specified intervals is used.

Consider $T$ as a discrete random variable with a probability function $f(t_j) = P(T = t_j)$; $j = 1, 2, ..., n$, where $t_1 < t_2 < ... < t_n$. The survivor function is given by the summation of this probability function: $S(t) = 1 - \sum_{j:t<t_j} f(t_j)$. The hazard at time $t_j$ is defined as the conditional probability that the event occurs at $t_j$ given that the event has not occurred before $t_j$: $h_j = P(T = t_j \mid T \geq t_j) = f(t_j) / S(t_j)$. The survivor function and the hazard function, are given by $S(t) = \prod_{j:t \geq t_j} (1 - h_j)$ and $f(t_j) = h_j \prod_{k=1}^{j-1} (1 - h_k)$ respectively.

The probability density function, $f(t)$, the survivor function, $S(t)$, and the hazard function, $h(t)$, are regarded as equivalent ways of describing the probability distribution of the survival time, $T$; if any one of them is known, the other two can be recovered. The fundamental equations in this section are very useful in various representations of models in survival analysis.

### 1.1.3. **Some Parametric Hazard Functions**

The hazard function, $h(t)$, as expressed in Section 1.1.2, is an unobserved function, yet it controls both the occurrence and the timing of the event. This hazard function or rate can only be estimated from the data, and it is often helpful to envisage hazard as a characteristic of an individual, not of a sample or population. It represents the instantaneous or immediate risk of the event (or death) at time $t$ for an individual who has survived until time $t$. It is a fundamental dependent variable in survival analysis. It must be non negative, $h(t) \geq 0$, and its integral over $[0, \infty)$ must be infinite, but it is not otherwise constrained; the hazard function may be increasing or decreasing, non-monotonic, or discontinuous.

The simplest hazard function is that with a constant rate over time: $h(t) = \lambda$ or equivalently, $\log h(t) = \mu$, where $\lambda = e^{\mu}$, for all $t > 0$. The corresponding survivor function is $S(t) = e^{-\lambda t}$ and the density, $f(t) = \lambda e^{-\lambda t}$ is indeed the well known exponential distribution with parameter $\lambda$. This relationship illustrates the importance of the exponential distribution in survival analysis. However, in reality the hazard is not always constant and therefore the assumption of an exponential distribution has its limitations. Increasing hazard rates often arise when there is disease progression of a patient or natural ageing. Although less common, decreasing hazard rates are sometimes observed, for example in patients experiencing organ transplant whereby the hazard rates are high before and just after surgery (due to infections or other surgical complications), but gradually decrease as the patients recover.

An alternative model which accommodates increasing, and decreasing hazard rates is the Weibull distribution. The hazard function is then given by $h(t) = \alpha\lambda t^{\alpha-1}$, and an exponential distribution is indeed a special case of Weibull distributions when the shape parameter $\alpha = 1$. The Weibull model is widely applicable in industrial applications, most importantly in engineering reliability analysis.

The hazard function has many alternative names in other fields. For example, it is also known as the conditional failure rate in reliability analysis, the force of mortality in demography, the intensity function in stochastic processes, the age-specific failure rate in epidemiology, and the inverse of the Mill's ratio in economics (Klein and Moeschberger, 1997). In this thesis, the term hazard function is used.

Many survival studies involve comparison between the hazard functions of two or more groups; hence a model for the relationship between them is needed. Some common relationships are namely proportional, non-proportional and accelerated failure time (AFT). Central to this research is the proportional hazards model (Cox,

1972), which will be briefly introduced in Section 1.5 and further described in Section 3.1.1. Other regression models such as proportional odds, additive hazards, and AFT are described in survival texts such as Sun (2006).

### 1.1.4. **Censoring Mechanisms**

Censoring occurs for many different reasons and may take different forms, as described in Section 1.1. Observations may be censored, and thus survival data are typically described by parameters which include a censoring indicator $\delta$ with a zero value when censored, otherwise unity or vice versa: $\delta \in \{0,1\}$ Figure 1.3 shows types of censoring often encountered include left-censored, right-censored and interval-censored.

Figure 1.3: Four common types of censoring in survival data

(X *indicates an occurrence of event*).

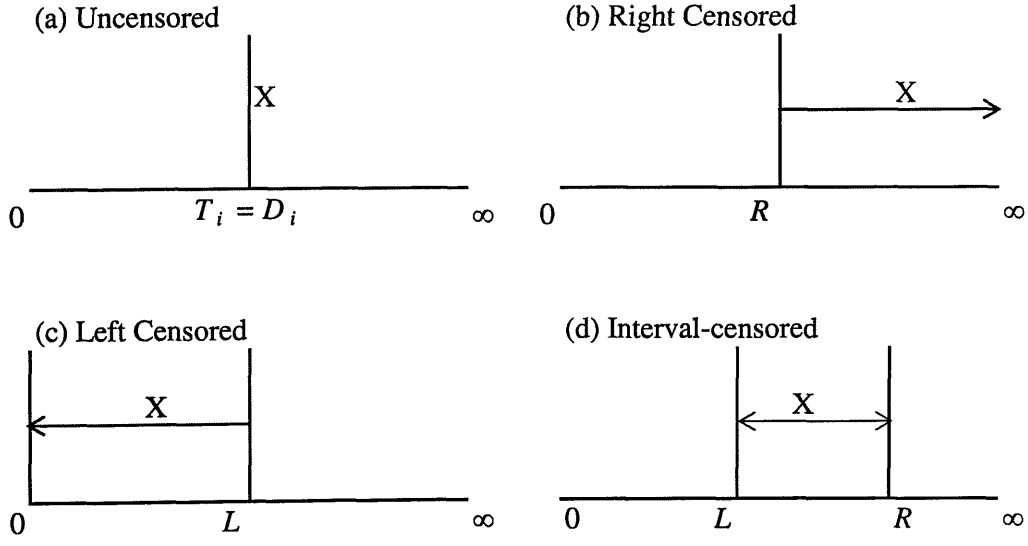


Figure 1.3 (a) shows the uncensored or complete survival data. This scenario occurs when the exact event time of patient $i$ is known, $T_i = D_i$, which is the actual survival time of the patient. However, as shown in Figure 1.3 (b), when the study ended before an event occurred, the actual event time is unknown, but it is known that the patient was still alive at that time: the survival time is given by the censoring time, $R$. The fact that the actual event has not occurred and may occur only to the right of the observed time, gives rise to the name right censored data. The third scenario involving left censoring is shown in Figure 1.3 (c). Left censored data arise when it is known that the event occurred prior to a certain time, $L$ (to the left), but the exact time is unknown. For example in a study of tumour recurrence, the examination is at three months after surgery where the exact event date is unknown, but is certainly earlier than the detection time.

Figure 1.3 (d) portrays interval censoring which occurs when the event of interest is known to have occurred within a defined interval ($L$, $R$). Interval censoring often arises in studies of non-fatal endpoints requiring regular follow-ups or inspections. Consider the case of tumour recurrence where no recurrence had been observed at a three months examination, but one was detected at a six months check-up. It is known that the event time is greater than three months and less than or equal to six months: $3 < T \leq 6$. Another common scenario of interval-censored data is present when continuous survival times are grouped into defined intervals prior to analysis. From Figure 1.3, left and right censoring are indeed special cases of interval censoring. The left, right and interval censoring can be represented by $(0, L]$, $(R, \infty]$ and $(L, R]$ respectively, where $(0, L]$ indicates that $0 < T \leq L$.

A critical assumption made in survival analysis is that failure times are independent of censoring times, which means that the hazard rates for the patients who are still at risk and those who have been censored are the same. A common model, specifically that known as random censoring, asserts that each subject satisfies this assumption. A specific condition requires that censoring be non-informative (Cox, 1984). This means that an individual who is censored at time $t$ should be representative of all those subjects who, having the same values of explanatory variables, survive up to that time. For example, if a patient was lost to follow-up because of migration, this may be non-informative censoring. However, if a patient was indeed lost to follow-up because of his deteriorating condition; the censoring becomes informative with respect to the patient's survival. Such informative censoring can lead to severe biases in survival analysis, and it is often difficult to determine the magnitude or direction of those biases. In situations where informative censoring is suspected, such as those involving long trial periods, sensitivity analysis is often conducted to assess the degree of bias (Allison, 2001).

To appreciate the importance of the censoring mechanisms, consider the following scenario. Suppose a trial shows similar results for two treatments, A and B in the treatment of basal cell carcinoma, with a primary endpoint of 5-year disease-free survival. By the end of five years, about 30 percent of both treatment groups had withdrawn. For treatment A, withdrawal is mainly due to treatment failure or side effects. Meanwhile, patients on treatment B withdraw because they are completely cured and prefer not to be followed-up further. In a standard analysis, disease-free survival rates would be overestimated for treatment A and underestimated for treatment B because they would both be based on the patients who remained in the study. A way to minimize the consequences of this problem is to perform a sensitivity

analysis of the main conclusions due to various assumptions when data on censored patients are scarce.

When survival curves are derived for endpoints other than death, the latter may be considered as a censoring event. For instance, if the endpoint is progression of breast cancer, then we may treat death due to heart disease as a censored event since there is no established relationship between these two diseases. However, in the case of lung cancer and smoking, treating death from heart disease as censoring might bias the result since smoking increases the risk of both cancer morbidity and cardiovascular mortality (Dupont, 2009). Whatever the type may be, censoring complicates the likelihood function, and hence the estimation of parameters of interest in survival analysis. Further reading on censoring is available from survival texts such as Anderson & Keiding (2006).

## 1.2. Univariate Survival Data

Univariate survival data occur when observations concern the time to a single event, for example, time to death, and individuals are assumed to be independent. As mentioned in Section 1.1.2, analysis of survival data requires methods that are able to accommodate both skewed and censored observations. Suppose there are two distinct groups of individuals in a study. A natural method to explore univariate survival data is to compute survival curves, $S(t)$, for each group and compare the proportions surviving over time. The non-parametric method of Kaplan Meier (Kaplan and Meier, 1958) allows a visual comparison of the survival experiences of the two groups, and it is often supplemented by a test statistic summarizing the overall survival comparison.

The logrank test is a popular method for comparing the survival of groups, which takes the whole follow-up period into account. It has an advantage that it does not require any knowledge about the shape of the survival curves and is a powerful

significance test if the assumption of proportional hazards is true. The proportional hazards assumption asserts that the ratio of hazard functions is the same at all time points. This means that for any $t \geq 0$, the relationship between the hazard functions of patients on treatment $E$ and $C$ is given by $h_E(t) = \psi h_C(t)$, where $\psi$ is a constant known as the hazard ratio.

For continuous predictor variables, a univariate Cox's proportional hazards regression (Cox, 1972), which is a semi-parametric model is often the popular choice. The proportional hazards assumption and the logrank test are described in Sections 1.5 and 1.6 respectively, while Cox's proportional hazards regression model is described in Section 3.1.1.

### 1.3. Interval-censored Survival Data

Interval-censored survival data commonly occur in medical or health studies that entail periodic follow-up examinations. Another situation giving rise to interval-censored data occurs when continuous survival times for all subjects are grouped into specific intervals for analysis purposes, as described in Chapter 3. In practice, survival data are often observed to the nearest time unit: day, month or year, and hence the analyses are generally based on interval-censored data. Consequently, it is natural to consider the underlying survival variables as discrete in developing methods for their analysis.

As described in Section 1.1.4, the interval-censoring mechanism involves two related variables $L$ and $R$ which define the interval. From this juncture onwards, the notation $T \in (L,R]$ indicates that $L < T \leq R.$. This is consistent with the definition of interval-censored observations viewed as a union of several non-overlapping windows or intervals (Turnbull, 1976). In this thesis, an assumption of independent interval-

censoring is made. This simply means that the censoring mechanism is independent of the survival times. Mathematically, this condition can be represented as $P(L < T \leq R \mid L = l, R = r) = P(l < T \leq r)$, such that the joint distribution of $L$ and $R$ is free of the parameters concerning the survival function of $T$. This is also known as non-informative interval censoring.

Suppose that patients, whose tumours have been removed, are randomized to experimental and control treatments. To compare the treatment effect, the patients are assessed at regular intervals to detect whether a tumour recurrence has occurred. In reality, it is not uncommon for patients to visit either earlier or later than at the scheduled date. Instead of being given for fixed intervals, the data now refer to occurrences at times varying from those scheduled for screenings, so varying also the interval between screenings. A simple way to analyse such data is by ignoring the interval censoring. However, this may not be appropriate if the time interval between screenings is short relative to the average time to recurrence, due to the interval-detected recurrences being fewer.

Another issue is often that such interval-detected recurrences contain information about patients' conditions: those experiencing symptoms visit earlier, while those feeling better visit later or miss altogether. Such a situation renders the standard survival analysis inappropriate since the length of intervals may vary widely between individuals and the assumption of independent censoring is thereby violated.

Unlike right-censored survival data analysis, which has been highly developed, its interval-censored counterpart has gained attention only quite recently. Comprising left and right-censoring, the analysis of interval-censored data generally cannot be achieved with the methods established for right-censored data. Earlier methods for grouped survival data have been established by Kalbfleisch & Prentice (1973) and

Prentice & Gloeckler (1978). Meanwhile, Finkelstein (1986) proposed a method for fitting the proportional hazards model to interval-censored data. The use of a complementary log-log transformation in the analysis of interval-censored survival data has been described by Whitehead (1989). Further, Whitehead (1997) adapted the log-rank test for comparing two treatment groups for use in the analysis of interval-censored survival data. A survival text by Sun (2006) provides a comprehensive coverage of the topic of interval-censored survival data.

The properties of the many proposed methods, however, remain unknown and there is no approach as simple as the partial likelihood method for right-censored data (Anderson & Keiding, 2006). In our research, a methodology is developed and validated to improve the analysis of interval-censored survival data.

### 1.4. Score Statistic and Fisher's Information

In an investigation of a treatment effect $\theta$, an important sample statistic is the cumulative measure of the advantage of the experimental treatment, often denoted by $Z$. Its companion, denoted by $V$, indicates the amount of information about $\theta$ contained in $Z$. Statistically termed as the efficient score statistic, and Fisher's information, $Z$ and $V$, respectively, they can be calculated at any stage of a clinical trial. As shown in this sub-section, both statistics can be derived from an appropriate likelihood function. It is to be noted that the term "treatment advantage" is used to denote a positive treatment effect when the parameterization is arranged so that $\theta > 0$ implies that the patients receiving the experimental treatment do better than those on the control.

Suppose we have a data set $x = x_i$, $i = 1, ..., n$ and a single unknown parameter $\theta$, for example a treatment effect. The likelihood of observing $\theta$, given $x$, is the probability of observing $x$, given $\theta$: $L(\theta; x) = P(\theta; x)$. The log likelihood, $\ell(\theta; x)$ is

defined as log $\{L(\theta; x)\}$. The first derivative of the log likelihood evaluated at $\theta = 0$ gives the efficient score statistic, $Z = \ell'(0)$, and its minus second derivative yields Fisher's information, $V = -\ell''(0)$. By Taylor's expansion, for small $\theta$, $\ell(\theta) \approx \ell(0) + \theta\ell'(0) + \frac{1}{2}\theta^2\ell''(0)$ which is approximately equal to $C + \theta Z - \frac{1}{2}\theta^2 V$, where $C$ is a constant.

According to Scharfstein, Tsiatis and Robins (1997), for large sample sizes, $Z$ is normally distributed with mean $\theta V$ and variance $V$. This assumption that $Z \sim N(\theta V, V)$ forms the basis for many statistical methods such as Pearson's chi-squared test, the logrank test and the Wilcoxon test; as it leads to asymptotically efficient methods. $Z$ and $V$ are also known as the logrank statistics since they are used to construct the logrank test, as will be shown in Section 1.5.

By definition, the maximum likelihood estimate (MLE) $\hat{\theta}$ of $\theta$ satisfies $\ell'(\hat{\theta}) = 0$. Since the log likelihood can be approximated, $\ell(\theta) \approx C + \theta Z - \frac{1}{2}\theta^2 V$, its first derivative $\ell'(\theta) \approx Z - \theta V$, and thus gives the MLE $\hat{\theta} \approx Z / V$. In large samples, the MLE is unbiased and linear in $Z$: it is therefore efficient. This formulation is central to this research, thus its recall shall be anticipated throughout this thesis. The derivation of $Z$ and $V$ for binary and survival data based upon the likelihood function is covered in Sections 1.4.1 and 1.4.2 respectively.

### 1.4.1. *Z* and *V* for Binary Data

Consider a study to assess the effectiveness of a new treatment comprising $n$ subjects, randomized to experimental or control treatment. For example, subjects are followed-up for a year after surgery and their binary status (either alive or dead) is recorded. At the end of the study, a total of $f$ subjects died (failed) while $s$ subjects are still alive (succeeded). Subscripts $E$ and $C$ are used to indicate experimental and control treatment respectively; this convention remains throughout this thesis, unless noted otherwise. Table 1.1 summarizes the binary outcomes of this hypothetical study.

Table 1.1: A 2 x 2 contingency table of binary response data (failure, success) in a two-group clinical trial comparing experimental and control.

| Response | Experimental | Control | Total |
|---|---|---|---|
| Failure | $f_E$ | $f_C$ | $f$ |
| Success | $s_E$ | $s_C$ | $s$ |
| Total | $n_E$ | $n_C$ | $n$ |

Suppose the probabilities of failure on $E$ and $C$ are denoted by $p_E$ and $p_C$ respectively, and the likelihood can be derived by conditioning on the right hand margin of Table 1.1. The likelihood is thus the probability of observing the outcomes, as in Table 1.1, given the condition that the total number of failures $F$ observed is $f$, and given $p_E$ and $p_C$. Simply, the conditional likelihood is given by, $L(p_E, p_C)$ = $P$(observed table| $F = f$). Based on conditional probability, $L(p_E, p_C)$ = $P$(observed table)/$\sum P$(table with $F = f$). This can be written as

$$L(p_E, p_C) = \frac{P(F_E = f_E, F_C = f_C)}{P(F = f)}, \quad f_E + f_C = f.$$

Suppose $j$ is an index denoting the various possible values that the random variable $f_E$ can take. The likelihood is given by the product of the likelihood components of $E$ and $C$,

$$L(p_E, p_C) = \frac{\binom{n_E}{f_E} {p_E}^{f_E}(1-p_E)^{s_E}\binom{n_C}{f_C}{p_C}^{f_C}(1-p_C)^{s_C}}{\sum_{j=\max(0,f-n_C)}^{\min(f,n_E)}\binom{n_E}{j}{p_E}^{j}(1-p_E)^{n_E-j}\binom{n_C}{f-j}{p_C}^{f-j}(1-p_C)^{n_C-(f-j)}},$$

which can then be simplified to

$$L(p_E, p_C) = \frac{\binom{n_E}{f_E}\binom{n_C}{f_C}}{\sum_{j=\max(0,f-n_C)}^{\min(f,n_E)}\binom{n_E}{j}\binom{n_C}{f-j}\left\{\frac{p_E(1-p_C)}{p_C(1-p_E)}\right\}^{-(f_E-j)}}. \tag{1.3}$$

Further, a measure of treatment effect is defined as minus the log-odds ratio given by

$$\theta = -\log\left\{\frac{p_E(1-p_C)}{p_C(1-p_E)}\right\}, \tag{1.4}$$

such that $E$ is better than $C$ if $\theta > 0$. The likelihood in equation (1.3) can now be expressed in terms of the log-odds ratio,

$$L(p_E, p_C) = L(\theta) = \frac{\binom{n_E}{f_E}\binom{n_C}{f_C}e^{-\theta f_E}}{\sum_{j=\max(0,f-n_C)}^{\min(f,n_E)}\binom{n_E}{j}\binom{n_C}{f-j}e^{-\theta j}}. \tag{1.5}$$

Under the null hypothesis of zero treatment effect, $p_E = p_C$ and $\theta = 0$. The likelihood is given by

$$L(0) = \frac{\binom{n_E}{f_E}\binom{n_C}{f_C}}{\binom{n}{f}}, \tag{1.6}$$

where the random variable $f_E$ follows a hypergeometric distribution with parameters $n$, $n_E$ and $f$. Upon taking the derivatives of equation (1.5) with respect to $\theta$, using the quotient rule, we get

$$L'(\theta) = -\frac{\binom{n_E}{f_E}\binom{n_C}{f_C} f_E e^{-\theta f_E}}{\sum_{j=\max(0,f-n_C)}^{\min(f,n_E)} \binom{n_E}{j}\binom{n_C}{f-j} e^{-\theta j}} + \frac{\binom{n_E}{f_E}\binom{n_C}{f_C} e^{-\theta f_E} \sum_{j=\max(0,f-n_C)}^{\min(f,n_E)} j\binom{n_E}{j}\binom{n_C}{f-j} e^{-\theta j}}{\left\{\sum_{j=\max(0,f-n_C)}^{\min(f,n_E)} \binom{n_E}{j}\binom{n_C}{f-j} e^{-\theta j}\right\}^2}. \tag{1.7}$$

Evaluating the derivative of the likelihood in equation (1.7) under the null, we get

$$L'(0) = -\frac{\binom{n_E}{f_E}\binom{n_C}{f_C}}{\binom{n}{f}}\left(f_E - \frac{fn_E}{n}\right), \tag{1.8}$$

where the expected number of failures on experimental treatment is given by $fn_E/n$. By definition, the score statistic is given by the first derivative of the log likelihood evaluated at $\theta = 0$: $Z = \ell'(0)$.

Using the chain rule, $Z = L'(0)/L(0)$ and using equations (1.6) and (1.8), the score statistic

$$Z = -f_E + \frac{fn_E}{n}, \tag{1.9}$$

which is the difference between the expected number of events under the null hypothesis, given the proportion of subjects on $E$, and the observed number of events on $E$. Simply stated, $Z$ is given by the expected failures minus the observed failures on $E$. Its variance is given by $V = -\ell''(0)$, which is then given by the quotient rule as $\frac{L(0)L''(0) - \{L'(0)\}^2}{\{L(0)\}^2}$. Upon differentiating the log likelihood for the second time, and evaluating the expression under the null, the Fisher's information is expressed as

$$V = \frac{n_E n_C fs}{n^2(n-1)}. \tag{1.10}$$

Both equations (1.9) and (1.10) are fundamental to this research and therefore will be recalled or re-expressed accordingly in subsequent chapters.

The derivation of the two logrank statistics $Z$ and $V$ have been shown above. These expressions are useful in the analysis of univariate data. In the case of bivariate data of two endpoints, the covariance between the two score statistics is needed. The derivation of such covariance for a binary data is given in Section 2.2.

### 1.4.2. *Z* and *V* for Survival Data

In this section, univariate survival data are considered. Suppose events are observed at times $t_a$ where $a = 1, 2,..., m$ and the number of patients at risk of an event (at time $t_a$) is $r_a$. An event is counted as a failure, and the total number of failures at each time is denoted by $o_a$; the total number of successes is then given by $r_a - o_a$. The outcomes at each failure time $t_a$ can be tabulated as shown in Table 1.2; note that the form is identical to Table 1.1 earlier, but specific to survival at time $t_a$.

Table 1.2: A 2x2 contingency table of univariate survival responses at time $t_a$.

| Response | Experimental | Control | Total |
|---|---|---|---|
| Failed at $t_a$ | $o_{aE}$ | $o_{aC}$ | $o_a$ |
| Survived beyond $t_a$ | $r_{aE} - o_{aE}$ | $r_{aC} - o_{aC}$ | $r_a - o_a$ |
| Total | $r_{aE}$ | $r_{aC}$ | $r_a$ |

Now let the discrete hazard function, $h_G(t_{a^*}) = P(T_G \in (t_{a^*}, t_{a^*+1}) \mid T_G \geq t_{a^*})$, $a^* = 0, ..., m+1$, $t_0 = 0$, where $t_{m+1} = \infty$, and $T_G$ is the survival time of a subject on treatment $G$, where $G = E, C$. The log-odds ratio of failing during $(t_{a^*}, t_{a^*+1})$, given survival past $t_a$ for $E$ relative to $C$ is given by

$$\theta_a = -\log\left\{\frac{h_E(t_a)(1-h_C(t_a))}{h_C(t_a)(1-h_E(t_a))}\right\}. \tag{1.11}$$

From equation (1.9), the score statistic is $Z_a = - o_{aE} + (o_a r_{aE}/r_a)$, which can be expressed as

$$Z_a = \frac{r_{aE} o_{aC} - r_{aC} o_{aE}}{r_a}, \tag{1.12}$$

and from equation (1.10), Fisher's information is now given by

$$V_a = \frac{r_{aC} r_{aE} o_a (r_a - o_a)}{r_a^2 (r_a - 1)}. \tag{1.13}$$

In the next section, we will see that equation (1.11) corresponds to the proportional hazards model, and equations (1.12) and (1.13) are summed over multiple time points. As mentioned earlier, the assumption that $Z_a \sim N(\theta_a V_a, V_a)$ for large $V_a$ and small $\theta_a$, is very useful and is continually referred to within this thesis. Based on this fundamental assumption, the derivation of $Z$ and $V$ using several methods for survival data, are described also in Chapter 3. Further in Chapter 4, the derivation of the covariance between two score statistics is described.

## 1.5. Proportional Hazards

In Table 1.2, the numbers of failures and successes were counted at defined event times. In this section, the assumption of proportional hazards is described, while the logrank test which considers all time points with an event is described in Section 1.6.

Suppose there is a finer grid which includes all the observed failure times (regardless of an event occurring). Now we may have intervals with zero failure; $o_E = o_C = o = 0$ and these intervals do not contribute anything to $Z_a$ and $V_a$. The hazard function is now, $h_G(t_a) = P(T_G \in (t_a, t_a + \delta t) \mid T_G \geq t_a)$ and as $\delta t \rightarrow 0$, $h_G(t_a) \approx \delta t h_G(t_a)$. From equation (1.11), $\theta$ is now given by

$$\theta_a = -\log\left\{\frac{\delta t h_E(t_a)(1-\delta t h_C(t_a))}{\delta t h_C(t_a)(1-\delta t h_E(t_a))}\right\},$$

$$\text{and as } \delta t \to 0,\ \theta_a \to -\log\left\{\frac{h_E(t_a)}{h_C(t_a)}\right\}.$$

The proportional hazards assumption is expressed as,

$$\theta = -\log\left\{\frac{h_E(t)}{h_C(t)}\right\} \tag{1.14}$$

for all $t > 0$ and implies that $h_E(t) = \exp(-\theta)h_C(t)$. The importance of this assumption is that the logrank test is efficient for the detection of a proportional hazards alternative, which is illustrated in Section 1.6.

The proportional hazards assumption is also central to the widely known Cox's model (1972). Cox's paper launched an explosion of statistical and applied research on the effects of individual patient characteristics on the survival process, with obvious extension to prognosis (Armitage & Gehan, 1974). Cox's method does not require any knowledge or assumption of the survival time distribution, and hence it is non-parametric with respect to time. Perhaps this flexibility is the primary reason for its popularity: the paper has been cited over 25,000 times as of June 2011. A detailed description of this model is covered in Section 3.1.1.

## 1.6. Logrank Test

The logrank test (Peto and Peto, 1972) is a non-parametric test to compare two samples of right-censored survival data. It is based on the assumptions that censoring is unrelated to prognosis, and that the survival probabilities are the same regardless of when a subject was recruited. Deviations from these assumptions matter most if they occur differentially in the groups being compared, for example, if censoring is more likely in one group than another. In summary, the logrank test statistic is constructed by computing the observed and expected number of events in one of the groups at each observed event time and then summing over all time points where there is an event, which is then divided by the square root of the cumulative variance.

Suppose two treatments are being compared, namely control and experimental. Since the intention of the experimental treatment is to reduce hazard, a positive $\theta$ indicates its superiority. Using equation (1.2), the survivor function for the experimental group is given by, $S_E(t) = \exp\left\{-\int_0^t h_E(u)du\right\}$, and similarly for the control group. Based on their relationships with treatment advantage and the assumption of proportional hazards,

$$S_E(t) = \exp\left\{-\int_0^t e^{-\theta} h_C(u)du\right\} = \left[\exp\left\{\int_0^t -h_C(u)du\right\}\right]^{e^{-\theta}} = \{S_C(t)\}^{e^{-\theta}}. \tag{1.15}$$

Taking logs of equation (1.15), it follows that $\log\{S_E(t)\} = e^{-\theta}\log\{S_C(t)\}$. Reversing the sign and taking logs again, gives an expression for the log hazard ratio, $\theta$ in terms of survivor functions,

$$\theta = -\log\{-\log S_E(t)\} + \log\{-\log S_C(t)\}. \tag{1.16}$$

For illustration of the logrank test, the example of survival data with $m$ events (assuming one event at each time) from Section 1.4.2 is considered here. Suppose there is a common treatment advantage given by the log hazard ratio in equation (1.14), or equivalently equation (1.16), such that $\theta_1 = ... \theta_m = \theta$. Using equations (1.12) and (1.13), the overall score statistic $Z$ and its null variance $V$ are given by summing the $Z_a$ and $V_a$ values over all time points, and respectively taking the forms,

$$Z = \sum_{a=1}^{m} Z_a = \sum_{a=1}^{m} \frac{r_{aE} o_{aC} - r_{aC} o_{aE}}{r_a} \tag{1.17}$$

and

$$V = \sum_{a=1}^{m} V_a = \sum_{a=1}^{m} \frac{r_{aE} r_{aC} o_a (r_a - o_a)}{r_a^2 (r_a - 1)}. \tag{1.18}$$

As noted in Section 1.4, the $Z$ and $V$ are known also as the logrank statistics, and the test based on them is called the logrank test or the score test. Similar assumption applies here: $Z \sim N(\theta V, V)$ when $\theta$ is small and $V$ is large. To test the null hypothesis that $\theta = 0$, the value of $Z/\sqrt{V}$ can be compared with the critical value of the standard normal distribution $N(0, 1)$, or equivalently comparing $Z^2/V$ with the critical value of the chi-squared distribution $\chi^2_1$. This is the popular logrank test which is commonly used to test the null hypothesis that there is no difference between the populations in the probability of an event over all time points.

The logrank test is most likely to detect a difference between groups when the risk of an event is consistently greater for one group than another. It is efficient for alternatives satisfying proportional hazards, but valid under the null with no further assumption. However, it is unlikely to detect a difference when survival curves cross: non-proportional hazards ratio. Since the logrank test is purely a test of significance, it cannot provide an estimate of the size of the difference between the groups or a

confidence interval. For these, we need to make some assumptions about the data. Common survival methods often use the assumption of proportional hazards and then estimate the ratio, as described in Section 1.5.

In many clinical trials, there exist factors associated with each patient which are anticipated to influence the patient response. For example, in multi-centre trials in which two treatment regimes are to be compared in terms of their effects on the survival times of cancer patients. The survival data are recorded for each centre and logically stratified by centre, with each centre termed a stratum. Instead of taking the individual log rank test for each stratum separately, a more realistic summary of the treatment effect over all strata is often necessary. This is achieved by a stratified logrank test which combines information about the treatment effect in each stratum. From above, denote the logrank statistic for the $s^{th}$ stratum as $Z_s$, and its variance as $V_s$. The stratified logrank test is then based on the statistic $(\sum_{s=1}^{S} Z_s)^2 / \sum_{s=1}^{S} V_s$, which has the chi-squared distribution $\chi^2_1$. Other variables often requiring stratification are gender, age group and a classification of diseases such as late stage or early stage for cancer studies.

In general, stratification allows a less restrictive assumption whereby the hazards need not be proportional between all strata (for example: centres or events), but more realistically the hazards are assumed to be proportional within each stratum. It is worth noting that the stratified logrank test is equivalent to the test of treatment effect in a stratified proportional hazards model when treatment is the only factor being fitted. This topic of stratification is further discussed in Chapter 4.

## 1.7. Sample Size Calculation

In the design of a clinical trial, one fundamental aspect is its sample size: the number of patients required to detect a clinically relevant difference reliably. Apart from making the main purpose of the trial clear, it is also important to identify the principal measure of outcome, type of data analysis, type of results anticipated with standard treatment, and the magnitude of treatment advantage to detect and its degree of certainty (Pocock, 1983). Based on current knowledge as well as previous trials, the magnitude of treatment advantage, sometimes referred to as the 'reference improvement' and denoted as $\theta_R$, can be set.

In any trial, there exist two possible errors: type I error and type II error. The former consists in falsely detecting a significant difference when the treatments are really equally effective, that is when $\theta = 0$. Its probability of occurrence is commonly denoted by $\alpha$, representing the risk of a false-positive result. A type II error arises when the trial fails to detect, as significant, a difference when there really is such a treatment difference of $\theta = \theta_R$. The probability of a type II error represents the risk of a false-negative result, and is commonly denoted by $\beta$. The 'power' to detect the desired treatment advantage is often denoted by $1-\beta$, which represents the degree of certainty that $\theta_R$, if present would be detected.

The result of a significance test is expressed as a 'p-value' ($p$), and a value of $p < 0.05$ would indicate that such an extreme observed difference or greater could be expected to have arisen by chance alone less than 5% of the time, when the null hypothesis is true; thus it is likely that the treatment difference really is present. Meanwhile, 'detecting a difference' is usually taken to mean 'obtaining a statistically significant difference with $p < 0.05$' (Machin, 1997).

In determining sample size and in the analysis of results, there exist two options for significance levels: one-sided and two-sided. The former seeks to quantify the evidence that the experimental treatment is better than control, with no interest in a difference in the other direction. The two-sided equivalent is considered a justifiable safeguard against prejudging the direction of treatment difference (Pocock, 1983). The p-value for a one-sided test is therefore half that for a two-sided test. For example, the type I error for the former might typically be set so that $p < 0.025$ (one-sided), while for the latter $p < 0.05$ (two-sided) would be required. In this thesis, both types of test are employed for illustration and convenience purposes.

The design of a clinical trial is dependent on the relationship between sample size, $n$, and information, $V$. It is logical that more subjects give more information about the unknown parameter, such that the relationship between $n$ and $V$ is directly proportional: $V = bn$ where $b$ is a constant. Meanwhile, for a two-sided type I error, the amount of information is given by a well-known expression (Whitehead, 1997),

$$V = \left\{ \frac{u_{\alpha/2} + u_{\beta}}{\theta_{\mathrm{R}}} \right\}^2, \tag{1.19}$$

where $u_\gamma$ is the upper $100_\gamma$ percentage point of the standard normal distribution: $\Phi(u_\gamma) = 1-\gamma$, and $\theta_R$ is the treatment advantage at the required power. The method of sample size calculation for a clinical trial is based on a power calculation. By specifying the desired power to detect a certain amount of treatment effect at the targeted significance level, the required information, $V$, can be obtained. The relationship between $V$ and $n$ can be quantified by a constant value, $b$, which will then lead to the sample size needed for the trial: $n = V/b$. The value of $b$ can be determined from the approximation of $V$ in terms of model parameters and sample size $n$; usually the result $V \propto n$ is found. For survival data, generally $V = e/4$ for 1:1 treatment

allocation, where $e$ is the total number of events, but interval-censoring leads to the relationship $V \propto n$ as usual. Equation (1.19) is fundamental to a design study and is referred to throughout this thesis.

This chapter has introduced the basic concepts of survival analysis that are used later in this thesis, described some of the important models and highlighted the difficulties in dealing with interval-censored survival data. In describing the proportional hazards model, the basic likelihood function was introduced in deriving the logrank statistic, and its relationship to the logrank test has been described. The principles governing sample size calculation too, were explained as part of the crucial requirement in clinical trials.

# Chapter 2. Global Test Methodology

This chapter presents the fundamental theory and application of global test methodology which will be used to combine multiple survival endpoints in later chapters. The global test methodology is first demonstrated for binary data in this chapter. Section 2.1 begins with an introduction to the global test approach in the evaluation of treatment efficacy. The established method of combining multiple binary endpoints is described and illustrated in Section 2.2 using a real data set from previous clinical trials involving stroke patients.

A study design of a clinical trial is described with illustrations of multiple stroke scales deployed. Section 2.4 presents the results of simulations conducted. The purpose of this investigation is two-fold: it serves as a foundation on which the case of interval-censored survival data will be built, and also forms a basis for comparison with results to be presented in that case. Finally, a discussion on the applicability of the method in general is provided.

## 2.1. Introduction to Global Test (Binary Data)

In clinical trials, the primary objective is often to evaluate the relative efficacy of two or more treatments. This may involve multiple assessment tools employed on the same group of patients, with each tool measuring certain aspects of each patient's condition. For the same patient, the outcomes on each tool or measurement scale are usually correlated with those on the other scales. The conventional method of analysis to evaluate treatment efficacy is one that controls the risk of detecting a significant difference when the treatments are really equally effective, that is a type I error, while maximising the power to detect a significant difference when there really is a difference. To find such a method is a challenging task especially when combining

outcomes that are correlated. O'Brien (1984) and Pocock, Geller and Tsiatis (1987) have combined multiple binary endpoints and reported that global tests may increase the power to detect differences between groups, but the appropriateness of their use should be carefully considered. Use of a global test as a primary analysis for binary outcomes, accompanied by secondary tests of individual outcomes was implemented in the NINDS t-PA Stroke Trial (Tilley et al., 1996). The success of this trial has increased awareness and acceptance of the approach within stroke research, as is evident by its adoption for the International Citicoline Trial in acUte Stroke (ICTUS) documented by Davalos (2007).

The existence of a variety of stroke scales poses a statistical opportunity in analysing the clinical trial, as each scale is correlated with the others and none captures all of the information on a patient's condition. Therefore, the multiple scales employed in a trial may be analysed in order to yield the most information. The method of combining the stroke scales works well for ordinary data analysis and can be adopted in meta-analysis and interim analyses of such trials. This methodology, also termed a 'global test' is useful when the outcome, such as clinical recovery from stroke, is difficult to measure and a combination of correlated outcomes (each measuring recovery from stroke) would be informative (Tilley et al., 1996). The global test statistic assumes a common treatment effect in terms of magnitude and direction on all outcomes. If the assumption is met, the power of the global test is equal to, or greater than, that of the individual outcomes.

In this chapter, the investigation is based on the data from four trials of citicoline by Davalos et al. (2002) that motivated the ICTUS trial. The outcome on each scale is categorized as a success for Barthel index (BI) $\geq 95$ (favourable outcome), modified Rankin scale (mRS) $\leq 1$ (no symptoms or no significant

disability) and National Institutes of Health Stroke Scale (NIHSS) $\leq 1$ (common or minor impact) as measured after 12 weeks from randomization. Previously, Bolland et al. (2009) concluded for larger samples that global tests gave accurate type I error rates and satisfactory power, even after the adjustment for prognostic factors. This chapter now investigates the simpler setting without any covariates, and also the effect, on the power, of deviation from the assumption of a common log-odds ratio. First, the underlying theories and assumptions are to be described.

### 2.1.1. Global $Z$ and $V$ Approach

The basic definitions of the score statistic $Z$ and Fisher's information $V$ were given in Section 1.4. These statistics are now described within the context of the global test approach. Suppose that each patient is measured on $m$ stroke scales and subscript $u$ is used to indicate the quantities associated with the $u^{th}$ of the stroke scales, where $u = 1, 2, ..., m$. The score statistics $Z_u$ are distributed approximately as $N(\theta_u V_u, V_u)$. Let $Z^+$ denote the sum of the score statistics and similarly $V^+$ the sum of its null variances. The global test statistic, which is the combination of all $Z_u$ values on each $u^{th}$ scale, $Z^+ = Z_1 + Z_2 + ... + Z_m$ is deployed. However, due to the correlation among the scales, the variance of $Z^+$ is no longer equal to the sum of the information, that is $\text{var}(Z^+) \neq V^+$ where $V^+ = V_1 + V_2 + ..... + V_m$. A correction factor of twice the sum of the covariances between $u^{th}$ and $v^{th}$ scales (when $u < v$) needs to be considered as $\text{var}(Z^+) = V^+ + 2(\sum_{u<v}^{m} C_{uv})$, where $C_{uv} = \text{cov}(Z_u, Z_v)$, $u, v = 1, 2, ..., m$.

The global test statistics are $Z^* = Z^+V^+/\text{var}(Z^+)$ and $V^* = V^{+2}/\text{var}(Z^+)$ such that the expected value of $Z^*$ is given by

$$E(Z^*) = E(Z^+V^+/\text{var}(Z^+)) = E(Z^+)V^+/\text{var}(Z^+) = \theta V^+ V^+/\text{var}(Z^+) = \theta V^*.$$

Consequently, the variance of $Z^*$ is given by

$$\text{var}(Z^*) = \text{var}(Z^+V^+/\text{var}(Z^+)) = \text{var}(Z^+)(V^+/\text{var}(Z^+))^2 = V^{+2}/\text{var}(Z^+),$$

which equals the global information, $V^*$. Assuming the log-odds ratios $\theta_u$ are all equal to $\theta$, $Z^*$ is distributed approximately as $N(\theta V^*, V^*)$. Under the null hypothesis of zero treatment effect on any scale, $\theta_1 = \theta_2 = ... = \theta_m = 0$, thus $Z_u \sim N(0, V_u)$ and similarly, $Z^* \sim N(0, V^*)$.

A global score test can be based on comparing the quantity $Z^{*2}/V^*$ with the critical value of the chi-squared distribution on 1 degree of freedom. Under the alternative hypothesis, when the log-odds ratios $\theta_1 = \theta_2 = ... = \theta_m = \theta \neq 0$, then the common log-odds ratio $\theta$ represents the global log-odds ratio of success for experimental treatment relative to control. The global test is always valid under the null hypothesis ($\theta_u = 0$) and is efficient for an alternative hypothesis in which the log-odds ratios $\theta_u$ are all equal. Since an assumption of equality of the $\theta_u$ is deployed in this study design, any deviation from it may degrade the power of the test. An illustration of this global test methodology using a real data set is given in the subsequent sections.

## 2.2. Combining Multiple Binary Outcomes

In Section 1.4.1, a 2 x 2 table for a single binary response has been described. This section extends the application of the derived quantities to binary outcomes evaluated on multiple stroke scales. The covariance between the score statistics which is required in combining analyses of bivariate binary data is also derived. The possible outcomes on an individual stroke scale such as BI, mRS or NIHSS, for both citicoline and placebo can be summarized in a 2 x 2 contingency table such as that in Table 1.1 on Page 18. Based on the conditional likelihood, $Z$ and $V$ are respectively given by

equations (1.9) and (1.10), except that $Z$ is now expressed in terms of $S$'s for successes, as is more natural for many forms of binary data,

$$Z = \frac{(n_C S_E - n_E S_C)}{n}. \tag{2.1}$$

Suppose that outcomes on $m$ stroke scales are available from each patient. The total number of patients is denoted by $n_{**}$ where $n_{**} = n_{E*} + n_{C*}$. A subscript $u$ is added to all quantities associated with the $u^{\text{th}}$ of the stroke scales, where $u = 1, 2, ..., m$. For combined outcomes on two scales: $m = 2$, a table can be constructed as shown in Table 2.1, where 0 and 1 respectively indicate success and failure according to the scales, while $Q_E^{(01)}$ denotes, for example, the number of counts for subjects on $E$ with success according to $u^{\text{th}}$ and failure according to $v^{\text{th}}$ scales.

Table 2.1: Counts of subjects for combined binary outcomes on two scales.

| Scale $u$ | Scale $v$ | $E$ | $C$ | Total |
|---|---|---|---|---|
| 0 | 0 | $Q_E^{(00)}$ | $Q_C^{(00)}$ | $Q_*^{(00)}$ |
| 0 | 1 | $Q_E^{(01)}$ | $Q_C^{(01)}$ | $Q_*^{(01)}$ |
| 1 | 0 | $Q_E^{(10)}$ | $Q_C^{(10)}$ | $Q_*^{(10)}$ |
| 1 | 1 | $Q_E^{(11)}$ | $Q_C^{(11)}$ | $Q_*^{(11)}$ |
| | Total | $n_{E*}$ | $n_{C*}$ | $n_{**}$ |

It is to be noted that the number of double successes is given by $Q_*^{(00)}$, while the number of successes on scale $u$ is given by $Q_*^{(00)} + Q_*^{(01)}$. Based on the notation in Table 2.1, the binary score statistic $Z_u$ is given by

$$Z_u = (Q_E^{(00)} + Q_E^{(01)}) - \frac{n_{E*}(Q_*^{(00)} + Q_*^{(01)})}{n_{**}},$$

and similarly

$$Z_v = (Q_E^{(00)} + Q_E^{(10)}) - \frac{n_{E*}(Q_*^{(00)} + Q_*^{(10)})}{n_{**}}. \tag{2.2}$$

It is to be noted that the binary score statistic in equation (2.2) is simply given by the difference between the observed successes and the expected successes. Conditional on the margins of Table 2.1, the covariance between the two score statistics is given by

$$\begin{aligned} C_{uv} &= \operatorname{cov}(Z_u, Z_v) = \operatorname{cov}\left\{(Q_E^{(00)} + Q_E^{(01)}), (Q_E^{(00)} + Q_E^{(10)})\right\} \\ &= \operatorname{var}\left(Q_E^{(00)}\right) + \operatorname{cov}\left(Q_E^{(00)}, Q_E^{(01)}\right) + \operatorname{cov}\left(Q_E^{(00)}, Q_E^{(10)}\right) + \operatorname{cov}\left(Q_E^{(01)}, Q_E^{(10)}\right). \end{aligned} \tag{2.3}$$

To find the variance and covariances in equation (2.3), further notation is necessary. Let $N_G$ denote the vector of category counts $(n_{G1}, ..., n_{Ga})'$ for $G = E, C, *$ and $q, r = 1, ..., a$. The conditional density of $N_E$, given $N_*$ is expressed as

$$f\left(n_E \mid N_*\right) = \frac{\binom{n_{*1}}{n_{E1}} \cdots \binom{n_{*a}}{n_{Ea}}}{\binom{n_{**}}{n_{E*}}}, \tag{2.4}$$

where $n_{**} = n_{E*} + n_{C*}$. It follows that $E(n_{Eq} \mid N_*) = n_{E*}\, n_{*q}/n_{**}$ and $E\{(n_{Eq}(n_{Eq}-1) \mid N_*\} = n_{E*}(n_{E*}-1)n_{*q}(n_{*q}-1)/\{(n_{**}(n_{**}-1)\}$, such that var $(n_{Eq}) = n_{E*}n_{C*}n_{*q}(n_{**} - n_{*q})/\{(n_{**}^2(n_{**}-1)\}$, as for hypergeometric sampling in a 2 x 2 table. Using similar arguments, $E(n_{Eq}n_{Er}) = n_{E*}(n_{E*}-1)(n_{*q}n_{*v})/\{(n_{**}(n_{**}-1)\}$ and $\operatorname{cov}(n_{Eq}, n_{Er}) = -n_{E*}n_{C*}n_{*q}n_{*r}/\{(n_{**}^2(n_{**}-1)\}$. Putting these expressions back into equation (2.3), we get,

$$\begin{aligned} C_{uv} &= \frac{n_{E*}n_{C*}}{n_{**}^2(n_{**}-1)}\left\{Q_*^{(00)}(n_{**} - Q_*^{(00)}) - Q_*^{(00)}Q_*^{(01)} - Q_*^{(00)}Q_*^{(10)} - Q_*^{(01)}Q_*^{(10)}\right\} \\ &= \frac{n_{E*}n_{C*}}{(n_{**}-1)}\left\{\frac{Q_*^{(00)}}{n_{**}} - \left(\frac{Q_*^{(00)} + Q_*^{(01)}}{n_{**}}\right)\left(\frac{Q_*^{(00)} + Q_*^{(10)}}{n_{**}}\right)\right\}. \end{aligned} \tag{2.5}$$

It is to be noted that $Q_*^{(00)}$ is the total count of *"double"* successes according to *both* $u^{\text{th}}$ and $v^{\text{th}}$ scales over both treatments, now denoted by $S_{uv}$, while the total number of patients $n_{**}$ can be represented by the conventional $n$, for simplicity. Similarly, the

total number of patients succeeding according to scale $\Sigma_u$ is denoted by $S_u$ which is given by the expression $Q_*^{(00)} + Q_*^{(01)}$ and likewise, $S_v$ for the quantity according to scale $\Sigma_v$. Therefore, the covariance between two binary score statistics $Z_u$ and $Z_v$ is written as

$$C_{uv} = \text{cov}(Z_u, Z_v) = \frac{n_E n_C}{n^2(n-1)}(nS_{uv} - S_u S_v). \tag{2.6}$$

This equation follows from the conditional likelihood approach, resulting in a denominator of $n^2(n-1)$, as derived by Whitehead et al. (2010), while that from the unconditional approach leads to a denominator of $n^3$, as in Pocock et al. (1987).

The general expressions for $m$ scales were given in Section 2.1.1. Specific for the case of $m = 2$, the global test statistic $Z^+ = Z_1 + Z_2$, while $V^+ = V_1 + V_2$, and var($Z^+$) $= V^+ + 2C_{12}$, where $C_{uv} = \text{cov}(Z_u, Z_v)$, $u, v = 1, 2$. Furthermore, for three scales, $m = 3$, hence $Z^+ = Z_1 + Z_2 + Z_3$, $V^+ = V_1 + V_2 + V_3$ and var($Z^+$) $= V^+ + 2(C_{12} + C_{23} + C_{13})$, where $C_{uv} = \text{cov}(Z_u, Z_v)$, $u, v = 1, 2, 3$. The expressions for the $Z$'s, $V$'s and $C$'s can be used to give the global score statistics $Z^*$ and the information $V^*$ needed to conduct a global score test on any given $m$ scales of stroke outcomes.

### 2.2.1. **Analysis of Data from Trials of Citicoline**

A previous study by Davalos et al. (2002) involved 1372 patients with 789 randomized to citicoline and 583 to placebo. The primary analysis was a global test for multiple outcomes of patients treated with citicoline for 6 weeks, and the efficacy of treatment was measured on Barthel index, modified Rankin scale and NIH stroke scale at 12 weeks after randomization. Bolland et al. (2009) evaluated a sequential global test design and concluded that the global score test can be used, even with adjustment for multiple covariates. In their study, treatment comparisons on multiple

scales were conducted using similar expressions as given earlier. The trial was designed to have a power of 0.80 to detect significance at the level 0.05 (two-sided) if all three scales have a true log-odds ratio equal to $\theta$ = log (1.26) = 0.231 which then led to fixing the power of ICTUS for the same odds-ratio value. A summary of the results is presented in Table 2.2.

Table 2.2: Summary of successes on individual and multiple scales from a meta-analysis study by Davalos et al. (2002).

| Treatment | Total no. of patients | Success on scale (s): (Proportions of success on scale(s)) | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | BI | mRS | NIHSS | BI_ mRS | mRS_ NIHSS | BI_ NIHSS | BI_mRS_ NIHSS |
| Citicoline | 789 | 283 (0.359) | 325 (0.412) | 325 (0.412) | 165 (0.209) | 279 (0.354) | 149 (0.189) | 122 (0.155) |
| Placebo | 583 | 186 (0.319) | 223 (0.383) | 217 (0.372) | 103 (0.177) | 189 (0.324) | 87 (0.149) | 72 (0.123) |

**BI_mRS denotes success on both the BI and mRS scales.*

In Table 2.2, the numbers of successes on each individual and combined scales are tabulated from the original data sets used by Davalos et al. (2002) and Bolland et al. (2009). Columns 3 to 5 present counts (and proportions) of successes on each individual scale, columns 6 to 8 counts of patients who succeeded on both of the two scales mentioned and column 9 counts of patients who succeeded on all three scales. With $p_E$ and $p_C$ denoting the probabilities of success on citicoline (experimental) and placebo (control) respectively, the advantage of citicoline relative to placebo can be expressed as the log-odds ratio, $\theta = \log\{p_E(1-p_C)/p_C(1-p_E)\}$, similar to equation (1.4) earlier. For each individual scale, the corresponding log-odds ratio is calculated: $\theta_1$ = 0.178, $\theta_2$ = 0.123 and $\theta_3$ = 0.166. Note that these $\theta_u$ values are not equal, but are close to each other. From equation (2.2), the values of $Z_1$, $Z_2$, and $Z_3$ are obtained,

leading to $Z^* = 20.335$, and similarly for the $V$ counterparts, resulting in $V^* = 131.704$. The calculated p-value is 0.038 indicating that the treatment difference is statistically significant at the 5% level (two-sided).

### 2.3. Sample Size for a Clinical Trial

The basic principles for sample size determination as described in Section 1.7 apply here. Say the probability of success on $u^{th}$ scale is $p_u$, where $u = 1$, the information needed can be estimated as follows. Put $w_u = p_u(1-p_u)$, $V^+ = V_1 \approx (n_E n_C / n)w_1$. Since $V^* = V^{+2}/\text{var}(Z^+)$ and $\text{var}(Z^+) = V^+ = V_1$, therefore $V^* = V_1 \approx (n_E n_C / n)w_1$. For equal sample sizes, $n_E = n_C = n/2$, then $V^* = \frac{1}{4}(nw_1)$ so that $b = \frac{1}{4}(w_1)$. For $m$ scales, $V^+ = V_1 + V_2 + ....+ V_m \approx (n_E n_C / n)(w_1 + w_2 + ...+ w_m)$. Each covariance, $C_{uv}$ can be approximated by $(n_E n_C / n)c_{uv}$ where $c_{uv} = p_{uv} - p_u p_v$ and $p_{uv}$ denotes the overall probability of simultaneous success on both $u^{th}$ and $v^{th}$ scales. The required $V^*$ and $b$ could be derived more generally first, before specialising to the binary case, and will be given later. From what was shown earlier, since $V^* = V^{+2}/\text{var}(Z^+)$ and $\text{var}(Z^+) = V^+ + 2(\sum_{u<v}^{m} C_{uv})$, then

$$V^* \approx \left(\frac{n_E n_C}{n}\right)\left(\sum_{u=1}^{m} w_u\right)^2 \left\{\left(\sum_{u=1}^{m} w_u\right) + 2\left(\sum_{u<v}^{m} c_{uv}\right)\right\}^{-1} \tag{2.7}$$

and $V = bn$ where

$$b = \frac{1}{4}\left(\sum_{u=1}^{m} w_u\right)^2 \left\{\left(\sum_{u=1}^{m} w_u\right) + 2\left(\sum_{u<v}^{m} c_{uv}\right)\right\}^{-1}. \tag{2.8}$$

Suppose all the binary responses on the $m$ scales were independent, then $c_{uv} = 0$ for all $u$ and $v$, and hence $V^* \approx (n_E n_C / n)(w_1 + w_2 + ....+ w_m)$, such that the quantity of information provided by the global test would be represented by the sum of those

provided by all $m$ scales. This can be written as $V^* \approx (n_E n_C / n)\left(\sum_{u=1}^{m} w_u\right)$, where $u = 1$, 2, .., $m$ scales. On the other hand, if all three responses were measuring exactly the same phenomenon, then the values of $w_u$ and $c_{uv}$ would lead to a single common value, $w$, for whatever value of $m$, and hence $V^* \approx (n_E n_C / n)w$, for a test based on any of the $m$ scales. The advantage of using a global test should reside somewhere between these two extreme situations.

The expressions for $V^*$ and $b$ are now given for two and three scales, as they will be required in the following investigations. For the case of two scales, $m = 2$, the information $V^*$ and $b$ can be estimated as

$$V^* \approx (n_E n_C / n)(w_1 + w_2)^2\{(w_1 + w_2) + 2(c_{12})\}^{-1}$$

$$\text{and} \quad b = \tfrac{1}{4}(w_1 + w_2)^2\{(w_1 + w_2) + 2(c_{12})\}^{-1} \tag{2.9}$$

while for three scales, $m = 3$,

$$V^* \approx (n_E n_C / n)(w_1 + w_2 + w_3)^2\{(w_1 + w_2 + w_3) + 2(c_{12} + c_{23} + c_{13})\}^{-1}$$

$$\text{and} \quad b = \tfrac{1}{4}(w_1 + w_2 + w_3)^2\{(w_1 + w_2 + w_3) + 2(c_{12} + c_{23} + c_{13})\}^{-1}. \tag{2.10}$$

Certain assumptions are made in the sample size calculation presented here. Firstly, the average of estimates (of the probabilities of success for patients on citicoline and placebo) from Bolland et al. (2009) are used as the true probabilities of success for patients on placebo, $p_{uC}$ on $u^{th}$ scale where individual success probabilities are taken to be $p_{1C} = 0.347$, $p_{2C} = 0.207$ and $p_{3C} = 0.195$, and the double success probabilities are taken as $p_{12C} = 0.200$, $p_{23C} = 0.147$ and $p_{13C} = 0.175$. Under the null hypothesis of zero treatment effect, the same values are deployed for citicoline. Say the alternative hypothesis is based on the $\theta_R$ value in the same study, which is $\theta_R = 0.231$ with an assumption of equality $\theta_1 = \theta_2 = \theta_3 = 0.231$. Type I error is set as $\alpha/2 = 0.025$ and the power of the test $(1-\beta)$ is aimed at 0.90; an increase in power compared to 0.80 in the earlier study by Bolland et al. (2009).

For each individual scale, the average success probability is $p_u = (p_{uC} + p_{uE})/2$, where $p_{uC}$ is the probability of success for patients on placebo according to $u^{th}$ scale, and that for citicoline, $p_{uE}$, is derived from the log-odds ratio assuming $\theta = 0.231$. The derived values are $p_1 = 0.374$, $p_2 = 0.227$, and $p_3 = 0.215$. For equal sample sizes for each treatment arm, $n_E = n_C = n/2$, $V \approx \{np_u(1-p_u)\}/4$. Since $V^* = \{(u_{\alpha/2} + u_\beta)/\theta_R\}^2$, where $u_{\alpha/2} = u_{0.25} = 1.960$ and $u_\beta = u_{0.10} = 1.282$, and $\theta_R = 0.231$, the sample size $n$ can then be calculated for each individual scale as displayed in Table 2.3.

Table 2.3: Summary of the success probabilities, calculated at $\theta = 0.231$ for individual scales, with their corresponding sample sizes.

| Scale | $p_{uC}$ | $p_{uE}$ | $p_u$ | $w_u$ | $n$ |
|---|---|---|---|---|---|
| BI | 0.347 | 0.401 | 0.374 | 0.234 | 3366 |
| mRS | 0.207 | 0.247 | 0.227 | 0.175 | 4490 |
| NIHSS | 0.195 | 0.234 | 0.215 | 0.169 | 4668 |

For two scales, computation of $b$ involves covariance, $c_{uv}$ as described in equation (2.9). In order to predict $c_{uv}$, we need the average double success probabilities given by $p_{uv} = (p_{uvC} + p_{uvE})/2$. While $p_{uvC}$ for placebo is given, its citicoline counterpart, $p_{uvE}$ needs to be determined. One way is to assume the same value of the correlation coefficient for both placebo and citicoline. This correlation coefficient, $\rho$, describes the association between both scales that measure different variables, which can be computed as

$$\rho = (p_{uv} - p_u p_v) / \sqrt{p_u(1-p_u)p_v(1-p_v)}. \tag{2.11}$$

Taking the individual success probabilities for placebo, $p_{uC}$ as per Table 2.3, and the double success probabilities for placebo as $p_{12C} = 0.200$, $p_{23C} = 0.147$ and $p_{13C} = 0.175$, we get $\rho_{12} = 0.665$, $\rho_{23} = 0.664$ and $\rho_{13} = 0.569$ in this case. These values indicate that each scale adds information to the others. If the value of $\rho$ is 1, then one

scale provides exactly the same information as the other and the power would equal that of a single test. The closer $\rho$ is to 1, the higher the correlation, which may lead to reduction in the advantage from combining scales. By preserving the correlation $\rho$, for example on BI and mRS scales, $\rho_{12}$ = 0.665, the corresponding double success probability for citicoline, $p_{12E}$ can be found by rearranging equation (2.11),

$$p_{12E} = \rho_{12}\left(\sqrt{p_{1E}(1-p_{1E})p_{2E}(1-p_{2E})}\right)+(p_{1E}p_{2E}) = 0.240,$$

and the average double success probability, $p_{12}$ = 0.220. From Table 2.3, $w_1$ = 0.234, $w_2$ = 0.175, and thus $c_{12}$ = 0.1351. Using equation (2.9), we get $b_{12}$ = 0.0616. A similar procedure is followed for the other combined scales, giving $c_{23}$ = 0.1142, $c_{13}$ = 0.1136, $b_{23}$ = 0.0517 and $b_{13}$= 0.0644. These values of $b$ are required to determine the sample sizes.

Table 2.4: Summary of the success probabilities ($\theta$ = 0.231), correlations and the corresponding values of $c_{uv}$ and $b_{uv}$ for two-scale.

| Scales | $p_{uvC}$ | $p_{uvE}$ | $p_{uv}$ | $\rho_{uv}$ | $c_{uv}$ | $b_{uv}$ |
|---|---|---|---|---|---|---|
| BI_mRS | 0.200 | 0.240 | 0.220 | 0.665 | 0.1351 | 0.0616 |
| mRS_NIHSS | 0.147 | 0.179 | 0.163 | 0.664 | 0.1142 | 0.0517 |
| BI_NIHSS | 0.175 | 0.212 | 0.194 | 0.569 | 0.1136 | 0.0644 |

As per equation (1.19) earlier, the global sample size, $n$, can then be calculated from the information required, $V^* \approx bn$ whereby $V^* = \{(u_{\alpha/2} + u_{\beta})/\theta_R\}^2$ when $Z^*$ is approximately distributed as $N(\theta V^*, V^*)$, and $\theta_R = \log(1.26) = 0.231$, leading to $V^*$ = 196.96. The sample size for a fixed sample study on combining BI and mRS scales (denoted as BI_mRS in Table 2.4) is then $n_{12} = V^*/b_{12} = 3198$. Similar calculations are repeated for each two-scale combination. Finally, for $m$ = 3, using $b$ = 0.0641 from equation (2.10), the required sample size is 3074 in order to achieve a power of 0.90 to detect significance at a level of 0.05 (two-sided), if all three scales have a true log-

odds ratio equal to $\theta = 0.231$. The resulting values of sample sizes $n$ for individual and combined scales ($m = 2, 3$) are displayed in Table 2.5.

Table 2.5: Summary of the success probabilities calculated at $\theta = 0.231$ for individual and combined scales, with their corresponding sample sizes, $n$.

| Scale(s) | $p_{uC}$ | $p_{uE}$ | $p_u$ | $n$ |
|---|---|---|---|---|
| BI | 0.347 | 0.401 | 0.374 | 3366 |
| mRS | 0.207 | 0.247 | 0.227 | 4490 |
| NIHSS | 0.195 | 0.234 | 0.215 | 4668 |
| BI_mRS | 0.200 | 0.240 | 0.220 | 3198 |
| mRS_NIHSS | 0.147 | 0.179 | 0.163 | 3810 |
| BI_NIHSS | 0.175 | 0.212 | 0.194 | 3060 |
| BI_mRS_NIHSS | 0.147 | 0.179 | 0.163 | 3074 |

*N.B. The values of $p_{uC}$ are extracted from Bolland et al. (2009).*

As displayed in Table 2.5, the required sample size is reduced for the combined scales, indicating the usefulness of the global test approach in minimizing patient recruitment for a clinical trial. The fundamental relationship between $V$ and $n$ (hence $\theta_R$) is applicable throughout the designing of clinical trials in subsequent chapters which reference to this section.

## 2.4. Simulation and Results

To investigate the accuracy of the sample sizes derived earlier, a simulation study is conducted to verify the type I error rate and the power of the test to detect treatment effect at the 2.5% (one-sided) significance level. A power of 0.90 is targeted in this simulation. It is to be recalled from the previous section that Bolland et al. (2009) had set type I error rate at 5% (two-sided) and power of 0.80 in their study. The data sets are generated from a random uniform distribution, $U$~(0, 1) from which the binary outcomes of a clinical trial are obtained. The proportion of successes is set by

specifying $U < p_u$ accordingly in the SAS codes, where $p_u$ is the average success probability for patients on $u^{th}$ scale. Under the null hypothesis of no treatment effect, the same success probabilities are used, $p_{uE} = p_{uC}$, for example on the BI scale, $p_{uE} = p_{uC} = 0.347$. Under the alternative hypothesis, $p_{uE} > p_{uC}$ and the values shown in Table 2.5 are used; the proportion of successes on $E$ needs to be included in the SAS codes accordingly since it is different than that on $C$.

Investigation begins with data for the individual fixed sample sizes for the targeted power to detect the desired significance level and simulated under both the null and alternative hypotheses. For each data set, simulations of 20,000 replicates are conducted to verify the type I error and the power of the test, with results shown in Table 2.6 below. Under the null hypothesis, $Z$ is distributed approximately as $N(0, V)$, while $\Phi(Z/\sqrt{V})$ follows a uniform distribution. Hence, the p-value which is given by $1- \Phi(Z/\sqrt{V})$ also follows $U$ (0, 1). Under the null, the average p-value should be very close to 0.5 as the mean for $U$ (0, 1) equals ½. The proportion of p-values $\leq 0.025$ gives an estimate of the type I error rate which represents the risk of a false-positive finding that is the probability of rejecting the null hypothesis when the treatments on both arms are indeed equal. Under the alternative hypothesis, the average p-value $\leq$ 0.025 illustrates the power of the test; that is the degree of certainty that the treatment difference, $\theta$, if present, would be detected.

Table 2.6: Simulation results under the null and alternative hypotheses for individual scales, each with a different sample size, $n$.

| $\theta$ | Scale | $n$ | p-value | p-value $\leq$ 0.025 | within 95% PI? |
|---|---|---|---|---|---|
| 0 | BI | 3366 | 0.497 | **0.029** | **No** |
| 0 | mRS | 4490 | 0.500 | 0.025 | Yes |
| 0 | NIHSS | 4668 | 0.501 | 0.025 | Yes |
| 0.231 | BI | 3366 | 0.011 | 0.898 | Yes |
| 0.231 | mRS | 4490 | 0.012 | **0.892** | **No** |
| 0.231 | NIHSS | 4668 | 0.011 | 0.899 | Yes |

*N.B. Texts in bold highlight out-of-limit situations.*

As illustrated in Table 2.6, under the null hypothesis, type I error rates are found to be within the 95% probability interval of (0.022, 0.028), except for that of the BI scale which marginally exceeds the upper limit by 0.001. This implies that the type I error rate when using BI is not exactly 0.025, as it is indeed an asymptotic result, whereby in this case, the actual $\alpha$ is perhaps slightly elevated. It is also noted that the 95% PI itself is approximate. Quite possibly, the proportion of p-value $\leq$ 0.025 = 0.029 (for BI) is simply a chance result. Under the alternative, where $\theta = 0.231$, the p-value for mRS lies slightly outside the 95% probability interval of (0.894, 0.906) based on the power target of 0.90. The results also show that large sample sizes are needed to achieve the desired power to detect significant treatment advantage if a single scale is being used.

For the combined scales, the individual and double success probabilities are now used to construct the appropriate 2 x 2 tables for each pair of scales. As usual, under the null hypothesis of no treatment difference, the same success probabilities are deployed on each treatment arm. The individual success probabilities are $p_{1E} = p_{1C} = 0.347$, $p_{2E} = p_{2C} = 0.207$, $p_{3E} = p_{3C} = 0.195$, while the double success probabilities

for placebo and citicoline are as displayed in Table 2.4 earlier. With knowledge of these values, a 2 x 2 contingency table of outcomes for two scales $u$, $v$, can be constructed as shown in the example for placebo on combined BI_mRS scales in Table 2.7.

Table 2.7: A 2 x 2 contingency table of outcomes for patients on placebo for the combined BI_mRS scales.

| Placebo | | mRS | | Total |
|---|---|---|---|---|
| | | Failure | Success | |
| BI | Failure | 0.646 | 0.007 | 0.653 |
| | Success | 0.147 | 0.200 | 0.347 |
| | Total | 0.793 | 0.207 | 1.000 |

Four categories are formed, *SS* = success on both scales, *SF* = success on $u^{th}$ scale, but failure on $v^{th}$ scale, *FS* = failure on $u^{th}$ scale, but success on $v^{th}$ scale, and *FF* = failure on both scales. From earlier, the double success probability for patients on placebo according to both BI and mRS scales, $p_{12C} = 0.200$, is represented by the probability for *SS*. Given that the total success probabilities for patients on placebo according to the BI scale, $p_{1C} = 0.347$ (last column, $2^{nd}$ row), the probability for success on BI but failure on mRS, represented by *SF*, can be computed by $p_{1C} - p_{12C} = 0.347 - 0.200 = 0.147$. Similarly, the probability for *FS* can be calculated from Table 2.7. The desired treatment advantage of $\theta = 0.231$ is then introduced by imposing the success probabilities for patients on citicoline ($p_{1E} = 0.240$, $p_{2E} = 0.179$, $p_{3E} = 0.212$, as in Table 2.4) to the same data sets. The method used to calculate these by preserving the correlation coefficient, has earlier been explained in Section 2.3. The case under the alternative hypothesis is illustrated in Table 2.8.

Table 2.8: A 2 x 2 contingency table for combined outcomes on BI_mRS scales for patients on citicoline.

| Citicoline | | mRS | | Total |
|---|---|---|---|---|
| | | Failure | Success | |
| BI | Failure | 0.592 | 0.007 | 0.599 |
| | Success | 0.161 | 0.240 | 0.401 |
| | Total | 0.793 | 0.247 | 1.000 |

Similar contingency tables are constructed for each combination of two scales under both the null and alternative hypotheses, and then simulation is run accordingly. The same process is extended to the combined three-scale, with 8 categories of probabilities. These are *SSS* = triple successes, *SSF*, *SFS*, *FSS* = double successes, *SFF*, *FSF*, *FFS* = single successes, and *FFF* = triple failures. Say *SSF* denotes successes on both BI and mRS scales, but failure on NIHSS scale. The total of success probabilities for patients on placebo according to the BI scale, calculated as $p_{1C}$ = 0.347, represents the summation of probabilities for *SSS*, *SSF*, *SFS*, and *SFF*. With the known double success probabilities, $p_{12C}$, $p_{23C}$, and $p_{13C}$, each category can be assigned with the correct probability, as shown in Table 2.9.

Table 2.9: The probability for each combined outcome for patients on citicoline and placebo, under the alternative ($\theta = 0.231$).

| Combined outcomes | Citicoline | Placebo |
|---|---|---|
| *SSS* | 0.179 | 0.147 |
| *SSF* | 0.061 | 0.053 |
| *SFS* | 0.033 | 0.028 |
| *SFF* | 0.128 | 0.119 |
| *FSS* | 0.000 | 0.000 |
| *FSF* | 0.007 | 0.007 |
| *FFS* | 0.022 | 0.020 |
| *FFF* | 0.570 | 0.626 |

Table 2.9 shows the probabilities of outcomes on combined scales for patients on citicoline and placebo under the alternative hypothesis: $H_1$: $\theta = 0.231$. Under the null, the probabilities of outcome for patients on citicoline are taken as those values for patients on placebo. Based on these probabilities, the outcomes on each scale can be simulated, and the required statistics and p-values can then be computed. The corresponding results for all the combined scales are displayed in Table 2.10.

Table 2.10: Simulation results for all two-scale and three-scale combinations under the null and alternative hypotheses.

| $\theta_u = \theta$ | Combined scales | $n$ | p-value | p-value $\leq 0.025$ |
|---|---|---|---|---|
| 0 | BI_mRS | 3198 | 0.494 | 0.028 |
| 0 | mRS_NIHSS | 3810 | 0.498 | 0.024 |
| 0 | BI_NIHSS | 3060 | 0.497 | 0.024 |
| 0 | BI-mRS-NIHSS | 3074 | 0.497 | 0.024 |
| 0.231 | BI_mRS | 3198 | 0.011 | 0.894 |
| 0.231 | mRS_NIHSS | 3810 | 0.011 | 0.894 |
| 0.231 | BI_NIHSS | 3060 | 0.011 | 0.898 |
| 0.231 | BI-mRS-NIHSS | 3074 | 0.011 | 0.900 |

Table 2.10 shows that the type I error rates are well within the 95% probability interval of (0.022, 0.028) under the null and similarly for the power of the test. Comparing with the power for each individual scale earlier (Table 2.6), it is evident that the combined scales yield higher power than those using an individual scale. Equivalent power can be achieved on combined scales with smaller sample sizes than that achieved on individual scales. For example, to give a power of 0.898, the required sample size for combined BI_NIHSS is 3060 (Table 2.10), whereas a sample size of 4668 on a single NIHSS is required (Table 2.6). Furthermore, the global test on combined three scales yields the highest power; 0.90 which is the set target for this study, attainable with a reasonably small sample size of 3074. It is to be noted that the slightly smaller sample size for BI_NIHSS ($n$ = 3060), compared to the global sample size ($n$ = 3074), could be a peculiarity of this particular data set.

Another simulation, fixing the same sample size ($n$ = 3000) on each individual and combined scales also proves the advantage of the global test over any single test as shown in Table 2.11 below. Consistent with earlier results, the combined all-three-scale procedure again gives the highest power of test of all.

Table 2.11: Summary of simulation results showing the power for each individual and combined scales under the alternative, with sample size $n$ = 3000.

| $\theta_u = \theta$ | Scale(s) | Power |
|---|---|---|
| 0.231 | BI | 0.864 |
| 0.231 | mRS | 0.746 |
| 0.231 | NIHSS | 0.739 |
| 0.231 | BI_mRS | 0.878 |
| 0.231 | mRS_NIHSS | 0.818 |
| 0.231 | BI_NIHSS | 0.891 |
| 0.231 | BI_mRS_NIHSS | 0.892 |

To investigate the effect of any deviation from the assumption of log-odds ratio equality ($\theta_u$ values are all equal), simulation is extended for situations in which $\theta_1 \neq \theta_2 \neq \theta_3$, with results summarized in Table 2.12. It is to be noted that this table contains the results for the cases of two-scale and three-scale and "N/A" indicates that the individual scale is not applicable.

Table 2.12: Summary of the p-values and powers from the simulations using unequal $\theta_u$ values, with a fixed sample size, $n$ = 3000.

| Case | $\theta_1$ (BI) | $\theta_2$ (mRS) | $\theta_3$ (NIHSS) | p-value | p-value $\leq$ 0.025 |
|---|---|---|---|---|---|
| 1 | 0.231 | 0.116 | N/A | 0.040 | 0.703 |
| 2 | N/A | 0.116 | 0 | 0.298 | 0.115 |
| 3 | 0.231 | N/A | 0 | 0.090 | 0.477 |
| 4 | 0.231 | 0.116 | 0 | 0.100 | 0.441 |
| 5 | 0.231 | 0.462 | N/A | 0.001 | 0.996 |
| 6 | N/A | 0.462 | 0.116 | 0.004 | 0.967 |
| 7 | 0.231 | N/A | 0.116 | 0.036 | 0.720 |
| 8 | 0.231 | 0.462 | 0.116 | 0.000 | 1.000 |

For Cases 1 to 4, the log-odds ratios are set such that $\theta_2 = ½(\theta_1)$, where $\theta_1$ = 0.231, and $\theta_3 = 0$. As anticipated, the power of test (last column) deteriorates upon departure from the equality assumption for log-odds ratio $\theta_u$. Only Case 1 shows quite a high power of 70.3%, largely offset by the unaltered treatment advantage on the BI scale. In Case 2, the power of the test suffers the most due to the absence of any treatment effect on the NIHSS scale which cannot be salvaged by only half of the desired log-odds ratio on the mRS scale. The effect of no treatment advantage on the NIHSS scale ($\theta_3 = 0$) is clearly illustrated in Cases 3 and 4 as the powers are both below 50% for this study.

To investigate further whether a reduction in treatment advantage on one scale could be recovered by an increase on another, Cases 5 to 6 follow with $\theta_1 = 0.231$, $\theta_2 = 2x\ \theta_1$, and $\theta_3 = ½ \text{ x } \theta_1$. Case 6 demonstrates that half of the treatment effect on one scale can be compensated for by a double on the other to achieve the desired power. On all three scales, the power of the test further improves as shown in Case 8. This serves as an additional confirmatory result to emphasize the importance of adherence to the equality assumption in the global test approach.

It is worth mentioning that Bolland et al. (2009) suggested, for the case where $\theta_u$ values are not equal, the power can be approximated by taking $\theta$ to be a weighted average: $\theta_w = (\theta_1 w_1 + \theta_2 w_2 + \theta_3 w_3)/(w_1 + w_2 + w_3)$, where $w_u = p_u(1-p_u)$, as defined in Section 2.3. Using equation (1.19), $V^* = \{(u_{\alpha/2} + u_\beta)/\theta_R\}^2$, $u_\beta$ can be found from the standard normal distribution table and so it is $1-\beta$ which gives the power. Using this approximation method for Cases 4 and 8 above, the powers are respectively 43.7% and 97.8%. Both values are very close to those displayed in Table 2.12, which are 44.1% and 100%. Hence, the approximation method works well for the cases explored here.

## 2.5. Discussion

The method of combining binary assessments has achieved the intended type I error rate under the null hypothesis of zero treatment advantage, and yielded improved power under the alternative hypothesis of equal log-odds ratio, $\theta = 0.231$, compared to that for individual scales or even the two-scale combinations. The fundamental equations in Section 2.3 simplify the procedure for deriving sample size and power for a clinical trial. This illustrates the benefit of the global test approach for the correlations established between BI, mRS and NIHSS.

The global test assumes a common treatment advantage, whereby a treatment has the same direction and size of effect on all outcomes. If the assumption is not valid, such that some outcomes for the experimental show benefits and others disadvantages, the power of the test is reduced. This is indeed a desirable feature in such clinical trials which demand consistent and persuasive evidence among the outcome measures. The basic requirement of this global test method is that the treatment advantage on each component scale can be summarized in terms of a score statistic $Z$, which is approximately normally distributed with the mean $\theta V$ and variance $V$. Many types of responses satisfy this requirement, including survival data under proportional hazards. Thus, there is huge potential for the application of this global test approach. The simulations conducted have shown evidence for large samples that global tests based on the score method yielded satisfactory type I error rates and accurate power in this binary assessment.

To conclude, the overall result in this chapter justifies further investigation into survival data and combination of responses to explore the capability of this promising method. Using similar methodology, Whitehead et al. (2010) have developed a global score test for binary and ordinal endpoints, which centred on the derivation of the correlation between two score statistics. In Chapter 4, the global score test methodology is extended to bivariate interval-censored survival data.

# Chapter 3. Methods for Interval-censored Survival Data

Interval-censored survival data were earlier introduced in Section 1.3. Such data may arise from the grouping of continuous survival data or genuinely discrete data. The aim of this chapter is to identify within a few derived models, the model that yields the best approximations to the score statistic $Z$, and Fisher's information $V$; the selected model will be used in subsequent work.

Section 3.1 begins with an overview of the modelling of survival data. The very basic form of Cox's proportional hazards model is described, followed by the application of the proportional hazards model in the context of interval-censored survival data. Two familiar models, namely log-odds ratio and complementary log-log transformation are explored; $Z$ and $V$ are derived from each model, and each leads to a different version depending on approximations made concerning information. Therefore, it is necessary to examine whether these variations are important before deciding on the model of choice. The resulting equations are then applied to a real interval-censored survival data set in Section 3.2.

A study design is formulated in Section 3.3 and simulation is performed to evaluate the accuracy of each method under the assumption of proportional hazards, and to investigate the effect of any deviation from it on the accuracy in Section 3.4. A discussion is given in the last section.

## 3.1. Modelling of Survival Data

The modelling of survival data needs considerable attention to the underlying distribution as well as to the censoring mechanism. While the non-parametric methods, such as the logrank test (Section 1.6), can be useful in a simple comparison between two or more groups of survival times, they may not be suitable for more complex data. For example, often in a clinical trial to evaluate treatment advantage, other variables such as gender, smoking status and physiological measurements are recorded as well as the survival times. In such a situation where these explanatory variables must be included in the analysis, a statistical modelling approach renders solutions.

As survival analysis is mainly centred on the hazard of death (or event) at any time after a study has begun, the hazard function itself needs to be modelled. Although hazard models are distinct from linear models, many principles of the latter are applicable to the modelling of survival data. In general, such is performed to identify which of the explanatory variables affect the hazard function and consequently to estimate the hazard function. Such knowledge is useful in evaluating treatment efficacy since the survivor function can then be obtained from the hazard function: their unique relationship has been described in Section 1.1.2.

The famous Cox's proportional hazards regression model is now described in greater detail than in Section 1.5. Cox's model is first explained in the situation where each survival time is distinct, before considering the case where two or more survival times might be tied at the same value. The latter case which involves tied failures or ties is relevant to interval-censored data. The log-odds ratio and complementary log-log transformation are then employed in the derivation of $Z$ and $V$. It is to be noted

that the scope here is limited to fixed covariates, such as treatment group, which do not vary over time.

### 3.1.1. Cox's Proportional Hazards Regression Model

Cox's proportional hazards regression model (hereafter Cox's PH) is very popular for many reasons, as already documented in many texts. Two key concepts of Cox's PH model are the proportional hazards model and the estimation procedure using maximum partial likelihood, which are described in this section. The former is a direct generalization of the Weibull and Gompertz models while the latter is genuinely a novel approach in statistics. The Cox's PH assumes a parametric form for the effects of the explanatory variables, but it allows an unspecified form for the underlying survivor function, hence it is a semi-parametric model.

In the following example, the comparison of two survival curves is expressed in the form of a proportional hazards model. This approach is essentially the same as the log-rank test when there is only one covariate: a treatment indicator. In fact, if there are no ties in the survival times or when Cox's discrete approach to ties is applied, the likelihood score test in Cox's regression analysis is identical to the log-rank test. The advantage of Cox's regression approach is its ability to adjust for the other variables by including them in the model, as will be shown later.

Consider a group of individuals randomized to treatments $E$ and $C$, and their hazards of failure at time $t$ which are given by $h_E(t)$ and $h_C(t)$ respectively. As described in Section 1.5, the hazard for an individual on $E$ is taken to be proportional to the hazard for a patient on $C$, at time $t$. For any $t \geq 0$, this proportional hazards model can be expressed as

$$h_E(t) = \psi h_C(t), \tag{3.1}$$

where $\psi$ is a constant known as the relative hazard or hazard ratio. This is the true meaning of proportional hazards. The assumption of proportional hazards implies that the survivor functions for individuals on *E* and *C* do not cross.

More generally, suppose there are *n* individuals randomized between *E* and *C*, $i = 1, \ldots, n$. The hazard for individual *i* at time *t*, denoted by $h_i(t)$, is simply expressed as the product of a baseline hazard function, $h_0(t)$ and the constant hazard ratio $\psi_i$. Since the hazard ratio cannot be negative, it can be written as $\psi_i = \exp\eta_i$ such that any value of the parameter $\eta_i$ always gives a positive value of $\psi_i$. To consider the different treatments, enter an indicator variable *X* with a value 1 if an individual is on *E* and 0 otherwise. Denoting by $x_i$ for the value of *X* for the $i^{th}$ individual and $\eta_i = \beta x_i$, the hazard function can be written as

$$h_i(t) = h_0(t)\exp(\beta x_i). \tag{3.2}$$

Note that the baseline hazard function, $h_0(t)$ is left unspecified apart from being a non-negative function; it can be envisaged as the hazard of individuals with $x_i = 0$. Equation (3.2) implies two conditions: (i) the value of any covariate is measured at a fixed point in time, for example, at the beginning of the study, and (ii) the effect of each covariate is constant throughout the individual's survival. The above model can be re-written by taking logarithms of both sides of equation (3.2) giving

$$\log h_i(t) = \alpha(t) + \eta_i, \tag{3.3}$$

where $\alpha(t) = \log h_0(t)$. This baseline hazard can be specified to cater for any function: for example, $\alpha(t) = \alpha$ defines the exponential model, while $\alpha(t) = \alpha t$ and $\alpha(t) = \alpha \log t$

describe the Gompertz and Weibull models respectively. Such a specification, however, may be redundant as the hazard for any individual is a fixed proportion of the hazard for any other individual. For example, in a clinical trial comparing treatment advantage of experimental over control, the hazard for a patient on $E$ is proportional to that for another patient on $C$. In the Cox's model, no form is specified for the baseline hazard.

In reality, the hazard of failure at a given time often depends on many variables which are termed explanatory variables. The above expression can be generalized to suit this situation accordingly. Suppose there are $k$ explanatory variables with all values recorded at baseline when a patient enters the study: $x_1, x_2, .., x_k$. Let these values be denoted by the vector $\boldsymbol{x} = (x_1, x_2, .., x_k)'$, so that the hazard function of the individual $i$ can be expressed as

$$h_i(t) = h_0(t)\exp(\boldsymbol{\eta}_i). \tag{3.4}$$

This is indeed the product of the baseline hazard function and the exponential of a linear function of a set of $k$ fixed covariates $\mathbf{x}_i$ with coefficient $\boldsymbol{\beta}$ where the constant, $\boldsymbol{\eta}_i = \boldsymbol{\beta}_1 \boldsymbol{x}_{i1} + ... + \boldsymbol{\beta}_k \boldsymbol{x}_{ik}$. In medical contexts, this constant is also known as the risk score or prognostic index. Applying equation (3.2) to two individuals $i$ and $j$, then their hazard ratio is simply

$$\frac{h_i(t)}{h_j(t)} = \exp\left\{\beta_1(x_{i1} - x_{j1}) + ... + \beta_k(x_{ik} - x_{jk})\right\}, \tag{3.5}$$

since the baseline hazard for one individual cancels out with that of the other, hence giving a constant hazard ratio over time. It is to be noticed that the linear combination of the explanatory variables resembles a linear model, hence making this approach semi-parametric. It is assumed that these explanatory variables act multiplicatively on

the hazard, and conditionally on the covariates $x_i$ and $x_j$; the failure times of individuals $i$ and $j$ are independent. The key point is that the baseline hazard need not be specified in order to obtain meaningful interpretation of the coefficients being estimated. The $\beta$ coefficients can be estimated using the method of maximum likelihood which will be covered next.

The basic expression for likelihood has been introduced in Section 1.4; partial likelihood is now described in the context of Cox's estimation method (Cox, 1975). The partial or conditional likelihood approach enables the $\beta$ coefficients of the PH model to be estimated without having to specify the baseline hazard, as earlier mentioned. To construct the partial likelihood function, it is assumed that the intervals between the successive failure times, where there are no failures occurring, do not contribute any information about the effect of the explanatory variables on the hazard of failure.

Suppose we have a data set comprising $n$ patients, $P_{j,1}$, $P_{j,2}$, ..., $P_{j,n}$ who failed at each time $t_j$ and those censored between any two times $t_j$ and $t_{j+1}$, $j = 1, 2, \ldots, m$.

The probability that the individual $i$ fails at time $t_j$ is conditional on $t_j$ being included in the set of $m$ failure times. Since the failure times are assumed to be independent, the partial likelihood is a product of the partial likelihoods, one at each time point, denoted by $L_j(\beta)$, which is the probability of patients $P_{j,1}$, $P_{j,2}$, ..., $P_{j,n}$ failing at time $t_j$ given $\beta$ and how many failed, and who were at risk of failing at time $t_j$. This quantity can be written as

$$L(\beta) = \prod_{j=1}^{m} L_j(\beta). \tag{3.6}$$

Suppose only one patient fails at each $t_j$, denoted by $ij^{th}$ and let $R(t_j)$ denote the risk set at time $t_j$, which consists of all the patients who are at risk of failing at time $t_j$. The partial likelihood over all failure times, is written as

$$L(\boldsymbol{\beta}) = \prod_{j=1}^{m} \frac{h_0(t_j)\exp(\boldsymbol{\eta}_{ij})}{\sum_{l \in R(t_j)} h_0(t_j)\exp(\boldsymbol{\eta}_l)}, \tag{3.7}$$

where the individuals who are at risk of failure at time $t_j$, are indexed by $l$. Since the baseline hazards cancel out, the partial likelihood for $m$ events can be generalized as

$$L(\boldsymbol{\beta}) = \prod_{j=1}^{m} \frac{\exp(\boldsymbol{\eta}_{ij})}{\sum_{l \in R(t_j)} \exp(\boldsymbol{\eta}_l)}. \tag{3.8}$$

It is to be noted that the baseline hazard is not required for model fitting, thanks to elimination by conditioning. The name "partial likelihood" implies that only part of the data is directly used in its construction; the probabilities for subjects experiencing an event are considered, while those for censored subjects are taken into account only when summing over the risk sets at event times that occur before the subjects being censored, as shown in equation (3.8).

In a conventional conditional likelihood, the probability of observing all of the data, conditional on a single set of ancillary data, is considered. Instead, for partial likelihood, only part of the data is used and the conditioning is progressive, considering each failure time separately and conditioning on the risk set associated with that particular time. This simply means that we only make use of the identity of the patient who failed amongst all those who were at risk at any failure time, whereas the actual failure times are not used. Nevertheless, the partial likelihood is treated in the same manner as an ordinary likelihood, and the general MLE theory still applies. This provides an asymptotic distribution for the maximizing value of $\boldsymbol{\beta}$, and provides also a hypothesis testing and confidence interval framework (Cox and Hinkley, 1974).

So far, only unique failure times have been considered, but in reality several patients may fail at the same time. These tied failures occur in part because survival times are recorded coarsely, perhaps to the nearest day. On the other hand, interval censored survival data (Section 1.3) arise from a slightly different formulation, whereby a coarse but precise representation of the data in time intervals, is given. Interval censored survival data often involve tied failures or ties within the failure time intervals. In order to incorporate these ties, the failure set at time $t_j$ (comprising all $o_j$ patients who failed at that time) and the set of all groups of $o_j$ patients who might have failed at time $t_j$ is included. Denoting these sets by $D(t_j)$ and $R(t_j)$ respectively, in Cox's method for ties, the partial likelihood can now be re-written as

$$L(\boldsymbol{\beta}) = \prod_{j=1}^{m} \frac{\prod_{u \in D(t_j)} \exp(\eta_u)}{\sum_{G \in R(t_j)} \prod_{v \in G(t_j)} \exp(\eta_v)}, \tag{3.9}$$

where $G(t_j)$ is the group of $o_j$ patients who might have failed at time $t_j$. Suppose patients $P_1$ and $P_2$ failed at time $t_j$, while patient $P_3$ is also at risk of failing. The likelihood considering Cox's method for handling ties is then given by

$$L(\boldsymbol{\beta})_{Cox} = \frac{\exp(\eta_1)\exp(\eta_2)}{\exp(\eta_1)\exp(\eta_2) + \exp(\eta_1)\exp(\eta_3) + \exp(\eta_2)\exp(\eta_3)}.$$

Cox's method as outlined above is accurate but can be computationally slow; hence approximation methods have emerged as offering more viable options. In general, ties are dealt with by considering the event times separately, such that $o$ subjects fail at $o$ separate times. Using equation (3.9), the product is now realized over the patients who failed, instead of over the failure times.

How the risk sets are defined for these separate event times is what distinguishes one method from another. For example, in the case of the failure time $t_j$ above, Breslow's method considers two separate event times whereby $P_1$, $P_2$ and $P_3$ are at risk, $P_1$ failed and then $P_2$ failed for the same risk set. The likelihood can be expressed as

$$L(\boldsymbol{\beta})_{Breslow} = \frac{\exp(\eta_1)\exp(\eta_2)}{\{\exp(\eta_1)+\exp(\eta_2)+\exp(\eta_3)\}^2}.$$

It is to be noticed that the denominator for the Breslow's method for handling tied observations is slightly different from that for the Cox's method. The former method is a default in SAS which will be used in analyses performed in Chapter 6.

A common goal in clinical trials is to compare two treatment groups by evaluating the treatment advantage, $\theta$. Relating to the likelihood in equation (1.5), the partial likelihood can be expressed as

$$L(\theta) = \prod_{j=1}^{m} \frac{\exp(-o_{jE}\theta)}{\sum_{q=\max(0,o_j-r_{jC})}^{\min(o_j,r_{jE})} \binom{r_{jC}}{o_j-q}\binom{r_{jE}}{q}\exp(-q\theta)}, \tag{3.10}$$

where $o_j$ is the total number of failures and $r_{jG}$ is the total number of patients on treatment $G$ at risk of $j^{th}$ event, $G = E$, $C$, and $q$ is an index denoting the various possible values of the random variable $o_{jE}$. By differentiating the logarithm of equation (3.10) and using equation (1.5), it can be shown that the efficient score is

$$Z = -o_E + \sum_{j=1}^{m}\left\{\frac{o_j r_{jE}}{r_j}\right\}, \tag{3.11}$$

where $o_E$ is the total number of failed patients on $E$; a positive value of $Z$ indicates that treatment $E$ is better than $C$. Upon differentiating a second time, we get

$$V = \sum_{j=1}^{m} \frac{o_j(r_j-o_j)r_{jC}r_{jE}}{r_j^2(r_j-1)}, \tag{3.12}$$

which is indeed identical to equation (1.18), the Fisher's information as described in Section 1.6. The proportional hazards model, with Cox's method for ties yields the logrank test as its associated score test.

As demonstrated above, the ordering or ranking of event time $t_j$ is used, but not its magnitude. Hence any monotonic transformation of the event times will neither change the coefficient estimates $\hat{\beta}$ nor alter the conclusion of the analysis. Since the baseline hazard is a nuisance parameter, it seems logical that there is not much information beyond the ranking information regardless of its underlying function. Indeed, it has been reported that for a wide range of hazard functions, Cox's partial likelihood yields inferences that are asymptotically equivalent to those obtained from the full likelihood based on all the data (Efron, 1977).

Partial likelihood estimates still retain two standard properties of ML estimates: consistency and asymptotic normality. This means that in large samples, they are approximately unbiased and their sampling distribution is approximately normal. Their standard errors are, however, larger than those for estimates using the full likelihood; but the robustness gained from the simplification of employing the partial likelihood method of estimation may outweigh this shortcoming.

Although its formulation appears rather complicated, Cox's PH regression model is available in many standard software packages. For example, SAS/STAT deploys it as PROC PHREG with a simple basic coding such as: "PROC PHREG DATA=DEATH; MODEL TIME*STATUS(0)=TREAT; RUN; ". This coding will fit a model of data set (DEATH) with only one covariate of treatment (TREAT) and two variables (TIME, STATUS). The MODEL statement specifies the variables that define the survival time (TIME), the censoring variable (STATUS: {0, 1} where 0

typically means censored), and the explanatory variables (TREAT). Further description and samples of typical output are given in Chapter 6.

The merits of the proportional hazards model and the estimation method have probably over-shadowed the other aspects of Cox's regression model. Among other attractive features of Cox's regression model are: (i) facility to incorporate time-dependent covariates, (ii) admissibility of stratified analysis that is effective in controlling nuisance variables, (iii) adjustability when a subject is not at risk of an event, and (iv) readiness to accommodate both discrete and continuous measurement of event times (Allison, 2001).

All in all, Cox's PH regression model has opened doors to many possibilities and thus numerous methods have built on it since its introduction. Examples of new extensions to the original Cox's model include the analysis of residuals, time-dependent coefficients, multiple or correlated observations, multiple time scales, time-dependent strata, and estimation of underlying hazard functions (Therneau and Grambsch, 2000). One drawback of the popular Cox's model is that the assumption of proportional hazards may be violated in the presence of heterogeneity which affects the hazard ratio (Hougaard, 2000). Wei et al. (1989) proposed the well-known Wei, Lin and Weissfeld (WLW) model which assumes that the marginal distributions of the multiple events time follow Cox's models; this method is later described in Chapter 6.

### 3.1.2. **Modelling Interval-censored Survival Data**

In Section 1.3, various types of interval-censored data have been described. From this section onwards, only the case of fixed intervals for all subjects is considered. For illustration purposes, interval-censored survival responses can be summarized as in Table 3.1. This research is developed partly on the basis of previous work by Whitehead and Thomas (1997), which concerns trial design for data of interval-censored nature.

Table 3.1: Statistics for a parallel group study with interval-censored survival responses (Whitehead, 1997).

| Treatment | Experimental | Control | Overall |
|---|---|---|---|
| Number of events | $e_E$ | $e_C$ | $e$ |
| Number of events in the interval | | | |
| $(0, t_1)$ | $o_{1E}$ | $o_{1C}$ | $o_1$ |
| $(t_1, t_2)$ | $o_{2E}$ | $o_{2C}$ | $o_2$ |
| . | | | |
| . | | | |
| $(t_{k-1}, t_k)$ | $o_{kE}$ | $o_{kC}$ | $o_k$ |
| Number of patients being followed-up for time $t$ | | | |
| $t_1$ | $r_{1E}$ | $r_{1C}$ | $r_1$ |
| $t_2$ | $r_{2E}$ | $r_{2C}$ | $r_2$ |
| . | . | . | . |
| . | . | . | . |
| $t_k$ | $r_{kE}$ | $r_{kC}$ | $r_k$ |

In viewing Table 3.1, suppose patients who have been cured of ulcers are recruited into a clinical trial to compare two treatments for suppressing recurrence and that endoscopies are performed at defined times, $t_1$, ...., $t_k$. A positive result indicates ulcer recurrence and is considered as an event within that time interval. The exact

times of events are not known since the events could have happened at any time during the interval between the last visit when the patient was determined to be negative for the outcome and the first visit with positive outcome.

In the trial, say $e$ events are observed, for a total of $r$ patients who are recruited over time $t$, randomized to experimental and control treatments giving $r_E$ and $r_C$ respectively. The patients are observed at fixed intervals between $t_{i-1}$ and $t_i$, $i = 1, 2, .., k$, and the number of events occurred, $o_i$ in the intervals is recorded, as shown above. The number of patients on experimental and control treatments who are at risk during each interval, $r_{iE}$ and $r_{iC}$ respectively are recorded, and their summation gives the total number of patients at risk for that interval, $r_i$. Note that the number of successes, that is no recurrence, is given by $r_i - o_i$. If an event has occurred for the patient, the associated failure time is recorded, and there are no further examinations. Meanwhile, patients may withdraw from the study or not have had any occurrence until the end of the study, thus their data are censored.

Interval-censored survival data can also be considered as a series of familiar 2 x 2 tables, one for each of the intervals $(t_{i-1}, t_i)$ under study as shown in Table 3.2. It is to be noted that this table is similar to Table 1.2, but particular to a defined interval.

Table 3.2: A 2 x 2 contingency table for survival responses at a defined interval for experimental and control groups.

| Responses during interval $(t_{i-1}, t_i)$ | Experimental | Control | Total |
|---|---|---|---|
| Number of events (Failure) | $o_{iE}$ | $o_{iC}$ | $o_i$ |
| Number of patients survived (Success) | $r_{iE} - o_{iE}$ | $r_{iC} - o_{iC}$ | $r_i - o_i$ |
| Total | $r_{iE}$ | $r_{iC}$ | $r_i$ |

The number of events that occurred during any interval $(t_{i-1}, t_i)$ is recorded as $o_{iE}$ and $o_{iC}$ for patients on experimental and control treatments respectively. Given that the total number of patients in the interval is $r_i$, the total number of successes is simply $r_i - o_i$, and similarly for those on experimental and control. The $k$ distinct 2 x 2 tables are assumed to have a common value for treatment effect $\theta$, measuring the advantage of experimental over control, regardless of the failure pattern. Hence, under the null where $H_0 : \theta = 0$, p-values are guaranteed to be valid.

The logrank test for interval-censored data is a special case of the Mantel-Haenszel test (Mantel, 1966) for combining 2 x 2 tables and the two are often used interchangeably. As already described in Section 1.6, the efficient score $Z$ is the sum of $Z_i$'s from 2 x 2 tables, $Z = \sum_i Z_i$, and similarly, $V = \sum_i V_i$. However, the derivation of $V$ is not quite straightforward due to the presence of nuisance parameters. Such parameters are not of immediate interest but need to be accounted for in analysis of the parameter of interest, $\theta$. An example of the parameterization is introduced in Section 3.1.3 and illustrated further in Section 3.1.4.

### 3.1.3. **Methods Using Log -Odds ratio Transformation**

A model using a logit link function expresses the log-odds of failure as a linear function of regression parameters. Suppose $p_{iE}$ and $p_{iC}$ denote the probabilities of occurrence for experimental and control respectively, and that $p_{iE}$ can be estimated by $(o_{iE} / r_{iE})$ for the $i^{th}$ interval, $i = 1, 2, ..., k$. Similar expressions can be derived for $p_{iC}$ accordingly. For independent Bernoulli random variables $o_{iE}$ and $o_{iC}$ with parameters $p_{iE}$ and $p_{iC}$ respectively, the likelihood functions for experimental and control are given by

$$\begin{aligned} L(p_{iE}) &= p_{iE}^{\,o_{iE}}(1-p_{iE})^{r_{iE}-o_{iE}} \\ L(p_{iC}) &= p_{iC}^{\,o_{iC}}(1-p_{iC})^{r_{iC}-o_{iC}}. \end{aligned} \tag{3.13}$$

Therefore, the log likelihood functions for each component are derived as,

$$\begin{aligned} \ell(p_{iE}) &= o_{iE}\log p_{iE} + (r_{iE}-o_{iE})\log(1-p_{iE}) \\ \ell(p_{iC}) &= o_{iC}\log p_{iC} + (r_{iC}-o_{iC})\log(1-p_{iC}). \end{aligned} \tag{3.14}$$

Based on the log-odds ratio approach, the advantage of experimental relative to control can be expressed as $\theta = -\log\left[\{p_{iE}(1-p_{iC})\}/\{p_{iC}(1-p_{iE})\}\right]$, as given in Section 1.4.1 earlier. In Table 3.2, there were two parameters concerning the probabilities of failures in the two groups ($p_{iE}$ and $p_{iC}$), conditional on surviving up to that time interval. Each of these probabilities can be re-expressed in terms of $\theta$, but to complete the parameterisation, a second parameter is required. This second parameter could be one of the treatment-specific probabilities, or the sum of the logit probabilities or the baseline hazards which may be strata-specific; regardless of the form, it is called the nuisance parameter. In order to derive $Z$ and $V$, parameterization

of the nuisance parameter is necessary. For example, let the nuisance parameter $\phi_i$ for all $i$, be expressed as $\phi_i = -\log\left(\frac{p_{iE}}{1-p_{iE}}\right) - \log\left(\frac{p_{iC}}{1-p_{iC}}\right)$.

For survival responses of the patients such as in Table 3.2, the efficient score for $\theta$ based on log-odds ratio, $Z_{OR}$ can be derived by the unconditional profile likelihood approach, giving

$$Z_{OR} = \sum_{i=1}^{k} \frac{r_{iE}o_{iC} - r_{iC}o_{iE}}{r_i} \tag{3.15}$$

for $k$ intervals, and the Fisher's information is,

$$V_{OR} = \sum_{i=1}^{k} \frac{r_{iC}r_{iE}o_i(r_i - o_i)}{r_i^3}. \tag{3.16}$$

It is to be noted that equations (3.15) and (3.16) are similar to equations (1.17) and (1.18) earlier; they are displayed here as a reminder. Both equations (3.15) and (3.16) are used as the first method to derive $Z$ and $V$ respectively (Method 1). Conditioning of the procedure with regard to the totals of $o_i$ events and $r_i - o_i$ successes can be applied to eliminate the nuisance parameters; this is referred to as Method 2. Based on this conditional likelihood, the efficient score, $Z$ is similar to equation (3.15) while the Fisher's information, $V_{OR(C)}$ is given by (Whitehead, 1997):

$$V_{OR(C)} = \sum_{i=1}^{k} \frac{r_{iC}r_{iE}o_i(r_i - o_i)}{r_i^2(r_i - 1)}. \tag{3.17}$$

When events are few relative to the numbers surviving in each interval, $r_{iE}/r_i \approx n_E/n = R/(R+1)$ where $n_E$ is the number of patients on experimental and $n$ is the total number of patients in the study. Similarly, $r_{iC}/r_i \approx n_C/n = 1/(R+1)$. Therefore, the amount of information can be approximated by $Re/(R+1)^2$, where $e$ is the total number of events observed. In the case of equal treatment allocation, $R = 1$, $V \approx e/4$ as

mentioned in Section 1.7 earlier. However, this approximation is unreliable if the number of events is large relative to the number of survivors.

In the case of the logrank test, only the relative ordering of survival time is used instead of the actual times, thus no assumption is required for the individual survival distribution. In general, when $\theta$ is small for large sample sizes, $Z$ is normally distributed with mean $\theta V$ and variance $V$ and the maximum likelihood of $\theta$ can be approximated by $\hat{\theta} = Z / V$.

The logit model above presumes that events can occur only at discrete points in time. For most applications, however, ties occur because event times are measured coarsely at defined intervals. Aside from the implausibility of the logit model for such data, the model suffers from a lack of invariance to the length of the time interval. For example, switching from person-months to person-years changes the model in a fundamental way, so that coefficients are not directly comparable across intervals of different length (Allison, 2001). An alternative model is complementary log-log link which is described next.

### 3.1.4. **Methods Using Complementary Log-Log Transformation**

A model using a complementary log-log (cloglog) link function expresses the log negative log survival probability, $-\log(-\log S(t))$, as a linear function of regression parameters. Like the logit function, the complementary log-log function takes a quantity that varies between 0 and 1 and associates it with a quantity that varies between minus and plus infinity. Unlike the logit function, however, the complementary log-log function is asymmetrical. For example, after taking the logit transformation, a change in probability from 0.25 to 0.50 is the same as one from 0.50 to 0.75 ($\theta$ = 1.10). On the complementary log-log scale, however, the difference between probabilities of 0.25 and 0.50 ($\theta$ = 0.88) is larger than the difference between

0.50 and 0.75 ($\theta$ = 0.69). This difference has an important practical implication in situations where the magnitudes being compared matter too, not just the difference between them.

From Section 1.6, treatment advantage can be expressed in terms of survival functions: $\theta = -\log\{-\log S_E(t)\} + \log\{-\log S_C(t)\}$. For interval-censored survival data (illustrated in Table 3.2), equation (1.16) can be written as

$$\frac{S_E(t_i)}{S_E(t_{i-1})} = \frac{S_C(t_i)}{S_C(t_{i-1})}\exp(-\theta),$$

for all $i$, and therefore, the log hazard ratio is

$$\theta = -\log\left[-\log\left\{\frac{S_E(t_i)}{S_E(t_{i-1})}\right\}\right] + \log\left[-\log\left\{\frac{S_C(t_i)}{S_C(t_{i-1})}\right\}\right].$$

Alternatively, it can also be expressed in terms of $p_{iE}$ and $p_{iC}$, the probability of occurrence within interval ($t_i$, $t_{i-1}$), for subjects on $E$ and $C$ respectively. The probability of occurrence for patients on the experimental during the interval ($t_i$, $t_{i-1}$), $p_{Ei}$ is given by the number failed at $t_i$ conditional on surviving at $t_{i-1}$: $p_{iE}$ = [1-$\{S_E(t_i)/S_E(t_{i-1})\}$] and similarly for the control group. Upon re-parameterization by these probabilities,

$$\theta = -\log\{-\log(1-p_{iE})\} + \log\{-\log(1-p_{iC})\} \tag{3.18}$$

$$\phi_i = -\log\{-\log(1-p_{iE})\} - \log\{-\log(1-p_{iC})\}, \tag{3.19}$$

for all $i$. From equations (3.18) and (3.19), the probability of occurrence for experimental can be given by $p_{iE} = 1-\exp\left[-\exp\left\{-\frac{1}{2}(\phi_i+\theta)\right\}\right]$ and similarly for control, $p_{iC} = 1-\exp\left[-\exp\left\{-\frac{1}{2}(\phi_i-\theta)\right\}\right]$. Under the null, $\theta$ = 0 and the maximum likelihood estimate, $\hat{\phi}_0$, of $\phi_i$ is given when $\ell_\phi(\phi_0) = 0$, hence giving

$$\exp(-\frac{\phi_0}{2}) = -\log(1-\frac{o_i}{r_i}) = q_i. \tag{3.20}$$

From equation (3.14), the full log likelihood in terms of $\phi_i$ and $\theta$, can now be expressed as,

$$\begin{aligned}\ell(\phi_i,\theta) = o_{iE}\log\left[1-\exp(-\exp(-(\frac{\phi_i+\theta}{2})))\right] - (r_{iE}-o_{iE})\exp(-(\frac{\phi_i+\theta}{2})) \\ +o_{iC}\log\left[1-\exp(-\exp(-(\frac{\phi_i-\theta}{2})))\right] - (r_{iC}-o_{iC})\exp(-(\frac{\phi_i-\theta}{2})).\end{aligned} \tag{3.21}$$

Differentiating equation (3.21) with respect to $\theta$ and $\phi_i$ respectively, we get $\ell_\theta = G + H$ and $\ell_\phi = G - H$, where

$$G = -\frac{o_{iE}}{2}\frac{\exp(-(\frac{\phi_i+\theta}{2}))\exp(-\exp(-(\frac{\phi_i+\theta}{2})))}{\left[1-\exp(-\exp(-(\frac{\phi_i+\theta}{2})))\right]} + \frac{(r_{iE}-o_{iE})}{2}\exp(-(\frac{\phi_i+\theta}{2})),$$

and $$H = \frac{o_{iC}}{2}\frac{\left[\exp(-(\frac{\phi_i-\theta}{2}))\exp(-\exp(-(\frac{\phi_i-\theta}{2})))\right]}{\left[1-\exp(-\exp(-(\frac{\phi_i-\theta}{2})))\right]} - \frac{(r_{iC}-o_{iC})}{2}\exp(-(\frac{\phi_i-\theta}{2})).$$

Under the null, $\theta$ = 0, and taking $\ell_\phi(\phi_0) = 0$, we get $\exp(-\frac{\phi_0}{2}) = -\log(1-\frac{o_i}{r_i}) = q_i$.

Substituting this expression back into $\ell_\theta$, the efficient score can be obtained from $\ell_\theta(0,\phi_0)$,

$$Z_{CL} = \sum_i^k \frac{q_i}{o_i}\{r_{iE}o_{iC} - r_{iC}o_{iE}\}. \tag{3.22}$$

Upon taking the second derivatives, the Fisher's information can be given by $V = -\ell_{\theta\theta} + \ell_{\theta\phi}(\ell_{\phi\phi})^{-1}\ell_{\phi\theta}$, and without conditioning on the margins, we get

$$V_{CL} = \sum_{i}^{k} \frac{q_i^2 (r_i - o_i) r_i}{4o_i} + \left\{ \begin{array}{l} \dfrac{-q_i^2 (r_i - o_i)(o_{iE} - o_{iC}) r_i}{4o_i^2} + \\ \dfrac{q_i}{4o_i} \left[ r_i (o_{iE} - o_{iC}) - o_i (r_{iE} - r_{iC}) \right] \end{array} \right\} \left\{ - \frac{4o_i}{q_i^2 (r_i - o_i) r_i} \right\}. \tag{3.23}$$

The unconditional $Z$ and $V$ given by equations (3.22) and (3.23) respectively, form Method 3. Similarly under the null hypothesis, but conditional on the margins, and

$\ell_{\theta\phi} = \ell_{\phi\theta} = K - L$, where $K = \dfrac{-q_i^2 (r_i - o_i)}{4o_i} r_{iE}$ and $L = \dfrac{-q_i^2 (r_i - o_i)}{4o_i} r_{iC}$,

$$V_{CL(C)} = \sum_{i}^{k} \frac{4KL}{K + L} = \sum_{i}^{k} \frac{q_i^2 (r_i - o_i) r_{iE} r_{iC}}{o_i r_i}. \tag{3.24}$$

The conditional expressions for $Z$ and $V$ are given by equations (3.22) and (3.24) respectively, form Method 4. Under the null hypothesis that failures are independent of treatment ($H_0$: $\theta = 0$), then conditional on $O_i = o_i$, the random variable $o_{iE}$ follows the hypergeometric distribution. This leads to an alternative expression for $V$, with a slight difference in the denominator,

$$V_{CL(H)} = \sum_{i}^{k} \frac{q_i^2 (r_i - o_i) r_{iE} r_{iC}}{o_i (r_i - 1)}. \tag{3.25}$$

Both equations (3.22) and (3.25) respectively are conditional expressions for $Z$ and $V$, here called Method 5. All these five methods are summarized and compared in the next section.

### 3.1.5. **Summary of Methods**

It is apparent that the statistics derived from the complementary log-log approach (cloglog link) are in the form of a multiplier to those derived from the log-odds ratio approach (logit link). On comparing equations (3.15) to (3.22) for $Z$, and equations (3.17) to (3.25) for $V$, it appears that $Z_{CL} = (q_i r_i / o_i) Z_{OR}$, while $V_{CL} = (q_i r_i / o_i)^2 V_{OR}$. Notice that from equation (3.22), when the $o_i$ are small relative to $r_i$, then $q_i \approx o_i / r_i$. Substituting this approximation into equation (3.22) gives the same expression as in equation (3.15). Similarly, equation (3.25) is reduced to the form as in equation (3.17). This situation occurs only when there are few events in each interval, implying that in the case of finer intervals, the method of cloglog is approximately the same as that of log-odds ratio. However, when the intervals are coarse, the former method is more appropriate.

The resulting expressions for $Z$ and $V$, as derived by the various methods described are summarized in Table 3.3. Methods 1 and 2 use the familiar likelihood function to derive both statistics based on the log-odds ratio, with the latter method taking the conditional likelihood approach resulting in $r^2(r-1)$ in the denominator for $V$. Since these two link functions often provide similar fits, the choice may depend upon whether inference should be in terms of odds ratios or discrete hazards ratios (TenHave, 1996) and should be based primarily on ease of interpretation (McCullagh, 1980). For survival data, an inference based on the hazards ratios is an obvious choice.

As explained earlier, Methods 3 to 5 take advantage of a more appropriate complementary log-log approach, which should be better suited for the case of interval-censored survival data. The statistics $Z$ and $V$ can be calculated for any single interval using equations in Table 3.3, while for $k$ intervals, $Z = Z_1 + Z_2 + ...+ Z_k$, and similarly $V = V_1 + V_2 + ...+ V_k$. The applications of both logit and cloglog link using these five methods are demonstrated in the next section.

Table 3.3: Summary of the various methods used to derive $Z$ and $V$: the derived formulae included.

| Method | Description | $Z$ | $V$ |
|---|---|---|---|
| 1 | Log-odds ratio (unconditional) | $\frac{(r_{iE}o_{iC} - r_{iC}o_{iE})}{r_i}$ | $\frac{r_{iE}r_{iC}o_i(r_i - o_i)}{r_i^3}$ |
| 2 | Log-odds ratio (conditional) | $\frac{(r_{iE}o_{iC} - r_{iC}o_{iE})}{r_i}$ | $\frac{r_{iE}r_{iC}o_i(r_i - o_i)}{r_i^2(r_i - 1)}$ |
| 3 | Complementary log-log (unconditional) | $\frac{q_i}{o_i}\{r_{iE}o_{iC} - r_{iC}o_{iE}$ | $\frac{q_i^2(r_i - o_i)r_i}{4o_i} + \left\{\frac{-q_i^2(r_i - o_i)(o_{iE} - o_{iC})r_i}{4o_i^2} + \frac{q_i}{4o_i}[r_i(o_{iE} - o_{iC}) - o_i(r_{iE} - r_{iC})]\right\}\left\{-\frac{4o_i}{q_i^2(r_i - o_i)r_i}\right\}$ |
| 4 | Complementary log-log (conditional) | $\frac{q_i}{o_i}\{r_{iE}o_{iC} - r_{iC}o_{iE}$ | $\frac{q_i^2(r_i - o_i)r_{iE}r_{iC}}{o_i r_i}$ |
| 5 | Complementary log-log (hypergeometric) | $\frac{q_i}{o_i}\{r_{iE}o_{iC} - r_{iC}o_{iE}$ | $\frac{q_i^2(r_i - o_i)r_{iE}r_{iC}}{o_i(r_i - 1)}$ |

## 3.2. Application to Bone Marrow Transplant Data

In this section, the methods listed in Table 3.3 are applied to data from a trial reported by Storb et al. (1986) concerning leukaemia patients. The patients were randomized between immunosuppression with cyclosporine (control) or with cyclosporine and methotrexate (experimental): abbreviated as CSP and CSP + MTX respectively. At the beginning of recruitment, 17 and 24 patients were randomized to experimental and control treatments respectively. The aim of the trial was to investigate whether the addition of methotrexate would reduce the incidence of acute graft-versus-host-disease (GVHD) upon the bone marrow transplantation. The survival times of patients on CSP and CSP + MTX in days are listed below, where an asterisk indicates a censored observation.

| | |
|---|---|
| CSP: | 1*, 9, 9, 10, 11*, 13, 14, 18*, 19, 20, 21, 21, 21, 23, 24, 25, 29, 30, 33, 34, 34, 62, 218*, 251*. |
| CSP + MTX: | 7*, 10, 12, 15*, 21, 21*, 24, 25, 26*, 34, 69*, 89*, 167*, 201*, 205*, 237*, 257*. |

The survival times are then split into three intervals with cut-off points at weeks 2, 3 and 4. The summary of occurrence of the acute GVHD during the defined intervals following transplantation is displayed in Table 3.4. As an example, interval (2, 3] indicates the third week after transplantation.

Table 3.4: Summary of outcomes for bone marrow transplant trial (Storb et al. 1986), used as illustration of interval-censored survival data.

| Treatment | Experimental | Control | Total |
|---|---|---|---|
| | Interval (0, 2] | | |
| Failure ($o_i$) | 2 | 5 | 7 |
| Success ($r_i - o_i$) | 15 | 19 | 34 |
| No. at risk ($r_i$) when $t = 0$ | 17 | 24 | 41 |
| | Interval (2, 3] | | |
| Failure ($o_i$) | 1 | 5 | 6 |
| Success ($r_i - o_i$) | 13 | 12 | 25 |
| No. at risk ($r_i$) when $t = 2$ | 14 | 17 | 31 |
| | Interval (3, 4] | | |
| Failure ($o_i$) | 2 | 3 | 5 |
| Success ($r_i - o_i$) | 10 | 8 | 18 |
| No. at risk ($r_i$) when $t = 3$ | 12 | 11 | 23 |

During the first two weeks of observation for the experimental group, 2 had suffered acute GVHD (failure) while 1 died without suffering from acute GVHD, hence recorded as success along with 14 others who were still alive and free from acute GVHD. Notice that the number of patients on the experimental treatment remaining at risk during the third week was only 14, out of which one patient suffered from acute GVHD and another died without suffering acute GVHD. After a fourth week of follow-up, 1 of the remaining 12 patients (at risk) had died without suffering from acute GVHD, 2 had suffered from acute GVHD and 9 were still alive and free from the acute disease. Similar details can be read off from Table 3.4 for patients on the control treatment. Applying each of the five methods described above, $Z_i$ and $V_i$ for each interval are calculated and the resultant sums of $Z$ and $V$ are displayed in Table 3.5.

Table 3.5: Summary of the cumulative score statistic, $Z$ and information, $V$, using the five methods numbered 1 to 5 (Table 3.3).

| Method | $Z$ | $V$ | $Z^2/V$ | $Z/\sqrt{V}$ | p-value | $\hat{\theta} = Z/V$ |
|---|---|---|---|---|---|---|
| 1 | 3.22 | 3.58 | 2.895 | 1.701 | 0.089 | 0.899 |
| 2 | 3.22 | 3.70 | 2.801 | 1.674 | 0.094 | 0.870 |
| 3 | 3.58 | 4.44 | 2.882 | 1.698 | 0.090 | 0.806 |
| 4 | 3.58 | 4.42 | 2.896 | 1.702 | 0.089 | 0.810 |
| 5 | 3.58 | 4.56 | 2.802 | 1.674 | 0.094 | 0.784 |

As anticipated, the multiplier effect of $qr/o$ being introduced by the complementary log-log approach (Methods 3 to 5) yields higher values of $Z$ compared to those based on the log-odds ratio (Methods 1 and 2). Comparing the values of $V$, the largest amount of information is given by Method 5, in which a hypergeometric distribution is assumed. Under the null hypothesis of no treatment effect, $Z^2/V$ follows the chi-squared distribution on 1 degree of freedom. The 2 sided p-values are shown in Table 3.5, indicating positive outcomes for this exploratory trial where a significance level of 10% level (two-sided) was being sought. On comparing Methods 2 and 5 which are both based on conditional likelihood, their p-values, and logrank test statistics are almost the same. Assuming $\theta_1 = \theta_2 = \theta_3 = \theta$, then the treatment advantage can be estimated by $\hat{\theta} = Z/V$. It is observed that Method 5 gives the smallest estimate of treatment advantage compared to other methods. Nevertheless, this investigation gives a good indication as to the applicability of all five methods which are to be further compared on a bigger scale in Section 3.4.

### 3.3. Power Specification of a Clinical Trial

With reference to Section 1.7 earlier, equation (1.19) can be applied in this study design. It is to be noted that in this chapter, $r$ is used to denote sample size which is the same as the number at risk of an event at time 0. From equation (3.16) based on the logit link, Fisher's information, $V = r_E r_C o(r-o)/r^3$. In cases of equal sample size on each treatment arm, $r_E = r_C = r/2$ and $o/r = p$, where $p$ is the average probability of occurrence; therefore

$$V = \frac{o(r-o)}{4r} = \frac{p(1-p)r}{4}. \tag{3.26}$$

Using equation (3.24) based on the complementary log-log link, $V = q^2(r-o)r_E r_C/or$, and with similar treatment allocation,

$$V = \frac{q^2(r-o)r}{4o} = \frac{\{-\log(1-p)\}^2(1-p)r}{4p}. \tag{3.27}$$

For the calculation of sample size, equation (1.19) can be used to find $V$, and then equation (3.27) to convert to $r$, the required sample size, given by $4Vp/\left[\{-\log(1-p)\}^2(1-p)\right]$. Both approaches can be used to calculate the required sample size, but only the one based on the complementary log-log link is used in this design.

In this study, the determination of the sample size is based on commonly used error probabilities for clinical trials. The probability of type I error, $\alpha$ is chosen as 0.025 (one-sided) and the probability of type II error, $\beta$ is targeted at 0.10, hence the power of the test $(1-\beta)$ is 0.90. The probability of an event occurring under the control treatment, $p_C$ is fixed at 0.60, while its experimental counterpart, $p_E$ under the alternative is varied from 0.35 to 0.55 in 0.05 increments. Based on the

complementary log-log approach, the reference improvement is given by $\theta_R = -\log\{-\log(1-p_E)\} + \log\{-\log(1-p_C)\}$. Taking the individual probabilities for control and experimental groups, an average probability of occurrence is estimated by $p = (p_C + p_E)/2$ and the required sample size $r$ is calculated as displayed in Table 3.6.

Table 3.6: Sample sizes, $r$, determined from the power calculation approach, for each of the five methods.

| Method | $p_C$ | $p_E$ | $p$ | $\theta_R$ | $V^*$ | $r$ |
|---|---|---|---|---|---|---|
| 1 | 0.600 | 0.350 | 0.475 | 0.755 | 18.45 | 162 |
| 2 | 0.600 | 0.400 | 0.500 | 0.584 | 30.78 | 256 |
| 3 | 0.600 | 0.450 | 0.525 | 0.427 | 57.64 | 460 |
| 4 | 0.600 | 0.500 | 0.550 | 0.279 | 134.93 | 1036 |
| 5 | 0.600 | 0.550 | 0.575 | 0.138 | 555.18 | 4104 |

As anticipated, a bigger sample size $r$ is needed to show significant results when the treatment advantage, $\theta_R$ is smaller. The values in the above table are used in the simulation study described next.

### 3.4. Simulation and Results

To investigate the accuracy of the values derived in Table 3.5, a simulation study is conducted to verify the type I error rate and the power of the test to detect treatment advantage at the 2.5% (one-sided) significance level. The data sets of specified sample sizes (Table 3.6) are generated from a binary distribution and randomized to control and experimental groups based on equal treatment allocation. Under the null hypothesis of no treatment effect, the same failure probabilities are applied for both treatment groups, that is $p_C = p_E$, while those values shown in Table 3.6 ($p_E < p_C$) are used for the case under the alternative hypothesis. A scenario of half the treatment advantage is also simulated to evaluate the resulting effect on the power of the test.

Investigation begins with data for the individual fixed sample sizes for the targeted power to detect the desired significance level and simulated under both the null and alternative hypotheses. For each data set, simulations of 20,000 replicates are conducted to verify the type I error and the power of the test, with results shown in Table 3.7 below. All methods are applied to the same data sets for their comparison purposes. Under the null, the average p-value (denoted by $p\_v_0$) should be very close to 0.5 as the mean for a distribution ~ U (0, 1) equals ½, while the proportion of p-value $\leq 0.025$ (denoted by $ind_0$) is the type I error. Under the alternative hypothesis, the proportion of p-value $\leq 0.025$ (denoted by $ind_1$) illustrates the power of the test: the degree of certainty that the treatment difference, $\theta$, if present will be detected.

Table 3.7: Results for the average p-value and type I error using each of the five methods, simulated under the null for each sample size $r$.

| $r$ | $\theta$ | Results | Method 1 | Method 2 | Method 3 | Method 4 | Method 5 |
|---|---|---|---|---|---|---|---|
| 162 | 0 | $p\_v_0$ | 0.499 | 0.499 | 0.499 | 0.499 | 0.499 |
| | | $ind_0$ | 0.023 | 0.023 | 0.023 | 0.023 | 0.023 |
| 256 | 0 | $p\_v_0$ | 0.500 | 0.500 | 0.500 | 0.500 | 0.500 |
| | | $ind_0$ | 0.026 | 0.026 | 0.026 | 0.026 | 0.026 |
| 460 | 0 | $p\_v_0$ | 0.500 | 0.500 | 0.500 | 0.500 | 0.500 |
| | | $ind_0$ | 0.025 | 0.025 | 0.025 | 0.025 | 0.025 |
| 1036 | 0 | $p\_v_0$ | 0.501 | 0.501 | 0.501 | 0.501 | 0.501 |
| | | $ind_0$ | **0.029** | **0.029** | **0.029** | **0.029** | **0.029** |
| 4104 | 0 | $p\_v_0$ | 0.501 | 0.501 | 0.501 | 0.501 | 0.501 |
| | | $ind_0$ | 0.026 | 0.026 | 0.026 | 0.026 | 0.026 |

*N.B. Text in bold highlights an out of limit situation.*

As illustrated in Table 3.7, under the null hypothesis, type I error rates are found to be within the 95% probability interval of (0.022, 0.028), except for those of sample size 1036 when they slightly exceed the upper limit (0.029 in bold). Note that

the normal approximation of 0.025 in reality is not perfectly exact. All five methods give exactly the same average for p-values as well as for type I error rates at three decimal points significance. However, further checking reveals that not all the values of $Z$ and $V$ were identical; coincidentally the proportion of p-values $\leq 0.025$ happened to be exactly the same to 3 decimal places.

Table 3.8: Results for the average p-value and power using each of the five methods, simulated under the alternative for each sample size $r$.

| $r$ | $\theta = \theta_R$ | Results | Method 1 | Method 2 | Method 3 | Method 4 | Method 5 |
|---|---|---|---|---|---|---|---|
| 162 | 0.755 | $p\_v_1$ | 0.012 | 0.012 | 0.012 | 0.012 | 0.012 |
| | | $ind_1$ | 0.896 | 0.896 | 0.896 | 0.896 | 0.896 |
| 256 | 0.584 | $p\_v_1$ | 0.011 | 0.011 | 0.011 | 0.011 | 0.011 |
| | | $ind_1$ | 0.901 | 0.901 | 0.901 | 0.901 | 0.901 |
| 460 | 0.427 | $p\_v_1$ | 0.011 | 0.011 | 0.011 | 0.011 | 0.011 |
| | | $ind_1$ | **0.899** | **0.895** | **0. 905** | **0. 899** | **0.895** |
| 1036 | 0.279 | $p\_v_1$ | 0.010 | 0.010 | 0.010 | 0.010 | 0.010 |
| | | $ind_1$ | 0.902 | 0.902 | 0.902 | 0.902 | 0.902 |
| 4104 | 0.138 | $p\_v_1$ | 0.011 | 0.011 | 0.011 | 0.011 | 0.011 |
| | | $ind_1$ | 0.900 | 0.900 | 0.900 | 0.900 | 0.900 |

*N.B. Texts in bold highlight different values observed.*

Under the alternative, where $\theta = \theta_R$ (Table 3.8) for all sample sizes p-values $\leq$ 0.025 lie within the 95% probability interval of (0.894, 0.906) based on the power target of 0.90 specified for the complementary-log-log approach. Apparently, the probability intervals seem to be applicable also to the log-odds ratio counterpart (Methods 1 and 2). All approximation methods give the same values except for those highlighted in bold text, where Method 3 yields slightly higher power than Methods 1 and 4, then followed by Methods 2 and 5. In the case of observing only half of the treatment advantage, the results are shown in Table 3.9.

Table 3.9: Results for the average p-value and power using each of the five methods simulated under the alternative (half the treatment advantage), for each sample size, $r$.

| $r$ | $\theta = \frac{1}{2}\theta_R$ | Results | Method 1 | Method 2 | Method 3 | Method 4 | Method 5 |
|---|---|---|---|---|---|---|---|
| 162 | 0.377 | $p\_v_{0.5}$ | **0.121** | **0.121** | **0.120** | **0.121** | **0.121** |
| | | $ind_{0.5}$ | 0.381 | 0.381 | 0.381 | 0.381 | 0.381 |
| 256 | 0.292 | $p\_v_{0.5}$ | **0.126** | **0.127** | **0.126** | **0.126** | **0.127** |
| | | $ind_{0.5}$ | 0.368 | 0.368 | 0.368 | 0.368 | 0.368 |
| 460 | 0.214 | $p\_v_{0.5}$ | **0.110** | **0.111** | **0.110** | **0.110** | **0.111** |
| | | $ind_{0.5}$ | **0.421** | **0.419** | **0.423** | **0.421** | **0.419** |
| 1036 | 0.14 | $p\_v_{0.5}$ | **0.123** | **0.124** | **0.123** | **0.123** | **0.124** |
| | | $ind_{0.5}$ | 0.361 | 0.361 | 0.361 | 0.361 | 0.361 |
| 4104 | 0.069 | $p\_v_{0.5}$ | 0.084 | 0.084 | 0.084 | 0.084 | 0.084 |
| | | $ind_{0.5}$ | 0.494 | 0.494 | 0.494 | 0.494 | 0.494 |

*N.B. Texts in bold highlight different values observed.*

The results of all approximation methods achieve exactly the same power values, except for the case of the sample size 460 where values were slightly lower than those of Method 3. Halving the treatment advantage appears to reduce the power of the test to as low as 36% ($r$ = 1036). While earlier results at $\theta = 0$ and $\theta = \theta_R$ failed to show any differences among the methods, however at $\theta = \frac{1}{2}\theta_R$, it now appears that Method 1 always yields exactly the same result as for Method 4, and likewise for Methods 2 and 5. While Table 3.7 and Table 3.8 earlier show that all methods gave the same p-values, Table 3.9 reveals that the p-values actually do vary. Despite the fact of these methods being mathematically different, they tend to generate identical results for average p-value and type I error rate or the power of the test.

## 3.5. Discussion

In conclusion, the simulation results have shown no appreciable difference between the approximation methods, despite small differences observed using real data sets earlier. All five methods have achieved the intended type I error rates under the null hypothesis, and yielded accurate powers under the alternative. The efficient score using the complementary log-log approach shows a multiplier of magnitude $(qr/o)$ of that derived on the basis of the log-odds ratio. The Fisher's information using the former approach has a multiplier of $(qr/o)^2$ compared to that using the latter.

Essentially, interval-censoring can be regarded as a case of missing data. Independent or random censoring relates to the mechanism of missing completely at random (MCAR), while the most common assumption of non-informative censoring corresponds to missing at random (MAR). Meanwhile informative censoring is due to the non-ignorable mechanism of missingness. Coarsening at random (Heitjan and Rubin, 1991) is another topic that is closely related to interval censoring. Consequently, the imposed stratification allows different proportional hazards within intervals, thus making interval-censored data less rigid than the continuous data.

At present, methods of estimation for interval-censored data are readily available to handle cases where data are independent (Section 1.3). However, methods for correlated interval-censored data are not well developed. Statistical methods for the analysis of covariate effects on interval-censored discrete survival data are needed since these data frequently arise in clinical trials and other biomedical studies involving periodic monitoring of patients for multiple outcomes. This comparison of methods in the derivation of $Z$ and $V$ sets the foundation for further work in the estimation of the correlation between two score statistics for interval-censored survival data using the complementary log-log approach (Method 5).

# Chapter 4. The Correlation between Two Score Statistics

The main objectives of this chapter are to derive an estimator for the correlation between two score statistics arising from interval-censored survival data in the absence of covariates, and illustrate its applications to real data. The theories described in the earlier chapters which form the building blocks of this study, are now put together in this core chapter.

To begin with, a description of bivariate survival data relating to correlated outcomes is given in Section 4.1. A global test approach is applied to interval-censored survival data in Section 4.2. In Chapter 3, the efficient score and Fisher's information were derived on the basis of the complementary log-log link and were then applied to binary data. An adaptation to interval-censored data is now provided in Section 4.3, with a procedure for the derivation of an estimator for the covariance between two score statistics, which consequently gives an estimator for the correlation. A description of common bivariate survival data then follows.

The estimation of an overall treatment effect is provided in Section 4.4. The proposed method is then applied in Section 4.5 to non-recurrent real data sets from various clinical trials. Recurrent events are next described based on some key model components, followed by an application of the proposed method to a recurrent real data set, in Section 4.6. An overall discussion is given in Section 4.7.

## 4.1. Bivariate Survival Data

Clinical trials are often conducted to compare two or more treatment groups with regard to their efficacy. Efficacy, for the purpose of this comparison, is often measured by more than one patient response, thus leading to multivariate data. Existing univariate methods can be employed for assessing each variable characteristic, but often an overall objective measure to quantify the efficacy is still required. As described in Chapter 2, global tests are becoming the methods of choice in stroke studies, to detect differences between groups when multiple endpoints are concerned. Bivariate survival data involves two endpoints which cannot be assumed to be independent, and one of the main interests in the analysis of bivariate survival data is the measure of dependence or association of these two variables. The complexity of studies concerning such correlated times-to-event which may involve multiple endpoints on the same subject, requires methods to take into account the correlation between multiple endpoints. For such data, the correlation between two score statistics can be used to obtain an overall treatment efficacy, which is central to this study.

Outcomes are regarded as correlated if one occurrence is dependent on the others. An example of such correlated occurrences is provided by time to cytomegalovirus (CMV) shedding in blood and time to CMV shedding in urine observed in an AIDS clinical trial on HIV-infected individuals. Similarly, time to disease progression (a pathological condition characterized by identifiable symptoms) and time to death for a cancer patient are also correlated. Another situation is present when the event times of several individuals are somehow related; for example lifespans of twins, or married couples who are exposed to the same household conditions. Correlated outcomes also arise from similar organs of a person, such as blindness in left and right eyes or failure of both the left and right kidneys of a chronic

diabetic patient, cartilage loss of left and right knees or failure of both the left and right hip joint replacements for an osteoarthritis patient. Another common type of correlated outcome is recurrence, for example, event times of the first and second asthma exacerbations, or of tumour recurrences of a bladder cancer patient.

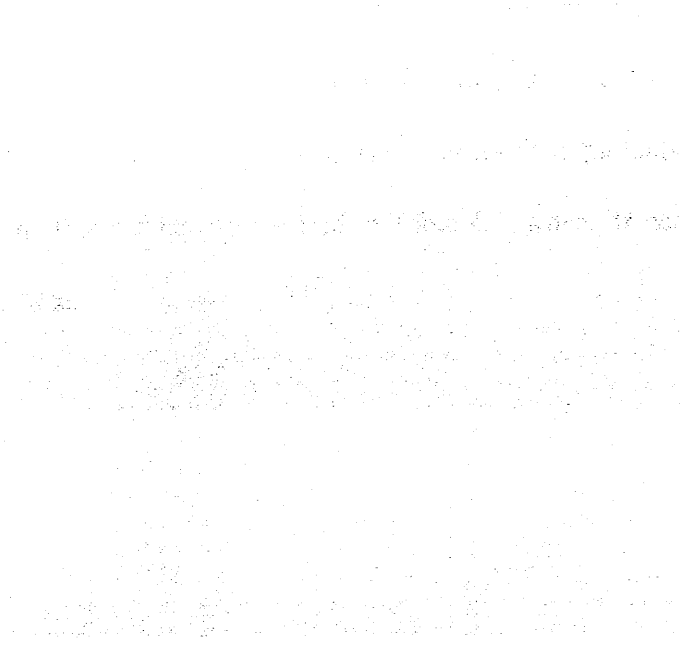
In dealing with correlated survival outcomes in cross-over trials, fixed effects models can be applied by fitting Cox's proportional hazards regression model stratified by subject. However, in parallel group trials where patients are randomized to experimental and control treatments, as considered in this study, these methods fail. To overcome this difficulty, recourse can be made to one of two methods, namely marginal and frailty modelling. Marginal modelling involves fitting data to Cox's regression model without any assumption of correlation, and then adjusting the estimated variance of the coefficients. This type of model, which relates to our approach, is later described in Section 6.1.1. A frailty model is a random effects model for event time data where subject effects are modelled as random variables; a good description is given by Hougaard (2000).

## 4.2. Adaptation of the Global Test to Interval-censored Survival Data

In Chapter 2, global test methodology was demonstrated in the context of multiple binary endpoints involving stroke data. In this section, the methodology is applied to bivariate interval-censored survival data. An instance of univariate interval-censored survival data was illustrated in Section 3.1.2 earlier, with consideration of multiple intervals. Referring to the 2 x 2 tables given in that section (Table 3.1 and Table 3.2), similarly the measure of treatment advantage given by the log hazard ratio, $\theta = -\log\{h_E(t)/h_C(t)\}$ is assumed to be constant over all times $t$. This method of

combining the 2 x 2 tables for univariate interval-censored survival data is now extended to the bivariate case, invoking the proportional hazards assumption.

This study considers three types of real bivariate survival data. The first type involves failure of similar physically related parts or paired organs: for example, right/left hip of an individual. The second concerns time to related events or indicators: for example, time to disease progression and time to death of a patient. The third involves recurrent events, where the same event can happen several times for an individual: for example, tumour recurrences. The first two types do not involve distinct ordering and hence are simpler than the third. These cases are termed paired, general and recurrent events respectively. Complete (uncensored) bivariate data are also examined for comparison purposes. Further descriptions of these data types and their classification are given in the subsequent sections, within the context of correlation between two score statistics.



### 4.2.1. **Bivariate Interval-censored Survival Data**

Suppose now that there are two endpoints of interest, giving two failure times, say $T_1$ and $T_2$, with each patient's two outcomes recorded for each of a series of prespecified time intervals. To describe this scenario fully, the following notation is necessary, where

$r_{1i}$ = no. of patients at risk of event 1 at the end of interval $i$,

$o_{1i}$ = no. of patients who had event 1 at the end of interval $i$,

$r_{2j}$ = no. of patients at risk of event 2 at the end of interval $j$, and

$o_{2j}$ = no. of patients who had event 2 at the end of interval $j$.

The combined outcomes relating to any pair of intervals $(t_{i-1}, t_i)$ and $(t_{j-1}, t_j)$ can be presented in Table 4.1. Note that this table is similar to Table 3.2, except that the counts of outcomes are now for combined variables. The abbreviations, *FF* and *SF* respectively refer to the number of patients who failed for both events and those succeeded for event 1 but failed for event 2. A similar convention applies for *FS* and *SS*. In reality, most survival data naturally involve censoring. Say the event times, $T_1$ and $T_2$ are subject to their own censoring with variables $C_1$ and $C_2$ accordingly. The consequences of such a censoring mechanism may result in patients missing $T_1$, or $T_2$ or even both, and this can be summarized in Table 4.1, where *M* denotes the censored or missing outcome.

Table 4.1: A 2 x 2 contingency table of censored bivariate data for patients on control for each time interval.

| For control | | $T_1$ occurence in $(t_{i-1}, t_i)$ | | | |
|---|---|---|---|---|---|
| | | Failure | Success | Missing | Total |
| $T_2$ occurence in $(t_{j-1}, t_j)$ | Failure | $FF$ | $SF$ | $MF$ | $o_{2jC}$ |
| | Success | $FS$ | $SS$ | $MS$ | $r_{2jC}$-$o_{2jC}$ |
| | Missing | $FM$ | $SM$ | $MM$ | $m_{2jC}$ |
| | Total | $o_{1iC}$ | $r_{1iC}$-$o_{1iC}$ | $m_{1iC}$ | ? |

The question mark at the bottom right corner of Table 4.1 is intentionally inserted to emphasize that the total number of patients at risk of event 1, event 2and of both, may no longer be the same when survival data are censored. Focussing only on the failures and successes in the paired $(ij)$ intervals, further notation follows:

$o_{(12),(ij)}$ = no. of patients who had event 1 during interval $i$ and also had event 2 during interval $j$,

$r_{(12),(ij)}$ = no. of patients at risk of event1 at the end of interval $i$ and also at risk of event 2 at the end of interval $j$,

$o_{(1\bullet),(ij)}$ = no. of patients who had event 1 during interval $i$ and also at risk of event 2 (but did not have event 2) during interval $j$, and

$o_{(\bullet 2),(ij)}$ = no. of patients who had event 2 during interval $j$ and also at risk of event 1 (but did not have event 1) during interval $i$.

From Table 4.1 (for control), the number of patients who had both events $o_{(12),(ij)C}$ is given by $FF$ and the number at risk for both events, $r_{(12),(ij)C}$ is given by the sum of $FF$, $FS$, $SF$ and $SS$. Within the same risk set, $o_{(1\bullet),(ij)C} = FF + FS$, and $o_{(\bullet 2),(ij)C} = FF + SF$. When the events are related, the marginal contribution of individual event 1 during interval $i$, and event 2 during interval $j$, and the combined contribution of both

events from the paired *ij* intervals are required to derive the correlation between two score statistics.

## 4.3. The Correlation between Two Score Statistics

Correlation is a measure of association estimated by the correlation coefficient, often itself abbreviated to 'correlation', and this terminology is used throughout this thesis. As with all types of data, techniques are required for analyses performed on correlated survival data. Such a technique is not as simple as those established over the past few decades for continuous and binary data. Our proposed method offers a straightforward account of such correlation. This section describes the theories and procedures involved in deriving various estimators of the parameters of interest; namely the covariance and the correlation between two score statistics, $Z_1$ and $Z_2$.

The proposed method is a marginal modelling, an approach where the effect of explanatory variables is estimated on the basis of the marginal distributions. In survival applications, the related events are usually assumed to be independent or to have some other imposed structure and existing models are often fitted to Cox's model described in Section 3.1.1. The subsequent section demonstrates our direct approach to the estimation of the covariance and correlation between two score statistics.

### 4.3.1. **Derivation of Estimators for Covariance and Correlation**

Assuming independent events for bivariate interval-censored survival data, the values of $Z$ and $V$ can be obtained for each event by using Method 5 (Table 3.5). As described in Section 2.2, the covariance between two score statistics is denoted by $\text{cov}(Z_u, Z_v)$, where $u$, $v$ = 1, 2 now denote two survival endpoints. The covariance between the score statistic with respect to event 1 during interval $i$ and the score statistic with respect to event 2 during interval $j$, is given by $\text{cov}(Z_{1i}, Z_{2j}) = E(Z_{1i}\, Z_{2j}) - E(Z_{1i})E(Z_{2j})$. Upon substituting the expressions for $Z$ as per equation (3.22), we get

$$\begin{aligned}
&\text{cov}(Z_{1i}, Z_{2j} \mid o_{1i}, o_{2j}) = E\left[\frac{q_{1i}}{o_{1i}}\left(r_{1iE}o_{1iC} - r_{1iC}o_{1iE}\right)\frac{q_{2j}}{o_{2j}}\left(r_{2jE}o_{2jC} - r_{2jC}o_{2jE}\right) \mid o_{1i}, o_{2j}\right] \\
&-E\left[\frac{q_{1i}}{o_{1i}}\left(r_{1iE}o_{1iC} - r_{1iC}o_{1iE}\right) \mid o_{1i}\right] E\left[\frac{q_{2j}}{o_{2j}}\left(r_{2jE}o_{2jC} - r_{2jC}o_{2jE}\right) \mid o_{2j}\right] \qquad (4.1) \\
&= \frac{q_{1i}q_{2j}}{o_{1i}o_{2j}}\left\{r_{1iE}r_{2jE}\text{cov}(o_{1iC}, o_{2jC} \mid o_{1i}, o_{2j}) + r_{1iC}r_{2jC}\text{cov}(o_{1iE}, o_{2jE} \mid o_{1i}, o_{2j})\right\},
\end{aligned}$$

where the covariances between the numbers of failures for event 1, and those for event 2, on experimental and control, are each conditioned on the total numbers of events, $o_{1i}$ and $o_{2j}$ respectively. Such conditioning on the margins for bivariate data is based on the familiar construction of a 2 x 2 table depicted in Table 4.1 earlier. For bivariate survival data, the covariance between the total numbers of failures for event 1, and event 2, respectively for the subjects on control is given by the covariance between individual failures summed over their risk sets:

$$\begin{aligned}
&\text{cov}(o_{1iC}, o_{2jC} \mid o_{1i}, o_{2j}) = \text{cov}(\delta_{1igC}X_{1igC} + .... + \delta_{1r_{1iC}gC}X_{1r_{1iC}gC}, \\
&\delta_{2jhC}X_{2jhC} + .... + \delta_{2r_{2jC}jC}X_{2r_{2jC}jC} \mid o_{1i}, o_{2j}) \\
&= \text{cov}\left(\sum_{i=1}^{r_{1iC}}\delta_{1igC}X_{1igC}, \sum_{j=1}^{r_{2jC}}\delta_{2jhC}X_{2jhC} \mid o_{1i}, o_{2j}\right) \\
&= \sum_{i=1}^{r_{1iC}}\sum_{j=1}^{r_{2jC}}\delta_{1igC}\delta_{2jhC}\text{cov}(X_{1igC}, X_{2jhC} \mid o_{1i}, o_{2j}), \qquad (4.2)
\end{aligned}$$

where $X_{1igC}$ = 1 if event 1 occurred in $i^{th}$ interval to $g^{th}$ patient on $C$,

0 otherwise,

$\delta_{1igC}$ = 1 if $g^{th}$ patient on $C$ is at risk of event 1 in $i^{th}$ interval ,

0 otherwise,

and $r_{1iC}$ = no. of patients on $C$ who are at risk of event 1 in $i^{th}$ interval.

A similar convention is applicable for patients in the experimental group and also for event 2, as denoted by the subscripts $E$ and 2 respectively. The individual outcome is a random variable $X_{1igC}$, which follows a Bernoulli distribution with parameter $p_{1iC}$, that is the probability of a patient in the control group for whom event 1 had occurred in the $i^{th}$ interval. $X_{1igC} \sim Bern(p_{1iC})$, and therefore, $E(X_{1igC}) = p_{1iC}$, while $E(X_{2jhC}) = p_{2jC}$ accordingly. Note that $\text{cov}(X_{1igC}, X_{2jhC}) = 0$ if $g \neq h$, for different individuals. Otherwise, for the same patient for whom both event 1 and event 2 had occurred,

$$\text{cov}(X_{1igC}, X_{2jgC}) = E(X_{1igC}X_{2jgC}) - E\left(X_{1igC}\right)E(X_{2jgC}), \tag{4.3}$$

where

$(X_{1igC}X_{2jgC}) = 1$ if the $g^{th}$ patient on $C$ has experienced both event 1 and event 2,

0 otherwise.

Similar to the case where the $g^{th}$ patient on $C$ only experienced either event 1 or event 2, the random variable $X_{1igC}X_{2jgC}$ also follows a Bernoulli distribution with parameter $p_{12(ij)C}$, which is the probability that for the $g^{th}$ patient on $C$ both event 1 and event 2 had occurred. $(X_{1igC}X_{2jgC}) \sim Bern(p_{12(ij)C})$, and thus the expected value is $E(X_{1igC}X_{2jgC}) = p_{12C}$. The covariance of the outcomes for both events for the same

individual on control can be approximated by $\text{cov}(X_{1igC}, X_{2jgC}) = p_{12(ij)C} - p_{1iC}p_{2jC}$ and similarly on experimental, $\text{cov}(X_{1igE}, X_{2jgE}) = p_{12(ij)E} - p_{1iE}p_{2jE}$. From equations (4.2) and (4.3), the covariance is given by

$$\text{cov}(o_{1iC}, o_{2jC} \mid o_{1i}, o_{2j}) = r_{12,(ij)C}(p_{12,(ij)C} - p_{1iC}p_{2jC}), \tag{4.4}$$

for sample sizes of $r_{1iC}$, , $r_{2jC}$, and, $r_{12,(ij)C}$ for control and similarly for experimental, as already defined. Under the null hypothesis, the probabilities of occurrences of both events on $C$ and $E$ are equal: $p_{12,(ij)C} = p_{12,(ij)E} \approx \hat{p}_{12,(ij)}$, where $\hat{p}_{12,(ij)}$ is the estimated average probability of occurrence of both events during $i^{\text{th}}$ and $j^{\text{th}}$ intervals. Similarly for the probabilities of occurrence of individual events, $p_{1iC} = p_{1iE} \approx \hat{p}_{1i}$, and $p_{2jC} = p_{2jE} \approx \hat{p}_{2j}$ where $\hat{p}_{1i}$ and $\hat{p}_{2j}$ are the estimated average probabilities of occurrence of event 1 during the $i^{\text{th}}$ interval and event 2 during $j^{\text{th}}$ interval respectively. Substituting these estimates into equations (4.4) and (4.1) consecutively, the covariance between the two score statistics is estimated by

$$\text{cov}(Z_{1i}, Z_{2j}) = \frac{q_{1i}q_{2j}}{o_{1i}o_{2j}}\left\{(r_{1iE}r_{2jE}r_{(12),(ij)C} + r_{1iC}r_{2jC}r_{(12),(ij)E})(\hat{p}_{(12),(ij)} - \hat{p}_{(1\bullet),(ij)}\hat{p}_{(\bullet 2),(ij)})\right\}. \tag{4.5}$$

Further, the probability of occurrence of both events can be approximated by the proportion of the number of occurrences from the number at risk: $\hat{p}_{(12),(ij)} = o_{(12),(ij)} / r_{(12),(ij)}$, and similarly for individual events, $\hat{p}_{(1\bullet),(ij)} = o_{(1\bullet),(ij)} / r_{(12),(ij)}$ and $\hat{p}_{(\bullet 2),(ij)} = o_{(\bullet 2),(ij)} / r_{(12),(ij)}$. Plugging these estimates into equation (4.5), the covariance estimator, denoted by $C_{12(ij)}$ is written as

$$C_{12(ij)} = \frac{q_{1i}q_{2j}}{o_{1i}o_{2j}r_{(12),(ij)}^{2}}\left\{(r_{1iE}r_{2jE}r_{(12),(ij)C} + r_{1iC}r_{2jC}r_{(12),(ij)E})(r_{(12),(ij)}o_{(12),(ij)} - o_{(1\bullet),(ij)}o_{(\bullet 2),(ij)})\right\}. \tag{4.6}$$

The covariance between two score statistics can be directly calculated from the observed events and the numbers at risk for individual $i$ and $j$ intervals as well as the paired $ij$ intervals. Equation (4.6) is of the utmost importance in this research. Conditioning on the successive risk sets, the covariance estimator between two score statistics, $C_{12}$ is obtained by the summation of the covariances from each pair of intervals, denoted by $C_{12(ij)}$ as in equation (4.6). In the case of interval-censored data, the covariance of each pair of intervals is summed to give the total covariance:

$$C_{12} = \sum_{i=1}^{u} \sum_{j=1}^{v} C_{12(ij)} .$$

It can be shown that for very large samples $n \rightarrow \infty$, $V_1 \rightarrow \text{var}(Z_1)$, $V_2 \rightarrow \text{var}(Z_2)$, and the estimate $C_{12} \rightarrow \text{cov}(Z_1, Z_2)$. In this study, our prime interest is in the estimator for $\text{cov}(Z_1, Z_2)$, which subsequently gives us the estimator for the correlation between the two score statistics. Fundamentally, the correlation between the two score statistics is given by the division of their covariance by the square root of the product of their variances. As in Section 1.4, the variance of $Z$ can be approximated by Fisher's information $V$; hence the correlation is expressed as

$$\rho = \text{cov}(Z_1, Z_2) / \sqrt{V_1 V_2} . \tag{4.7}$$

Therefore, the correlation $\rho$ between these two score statistics, can be estimated by

$$\hat{\rho} = C_{12} / \sqrt{V_1 V_2} . \tag{4.8}$$

In equation (4.6), the covariance formulation comprises two distinct parts. The first concerns marginal variables of $q_i$, $o_i$ and $r_i$, while the second involves the combined or paired intervals with subscripts 12. As the number of marginal failures, $o_i$, approaches very small values relative to the number at risk, $r_i$, the quantity $q_i$= -log $(1-o_i/r_i) \approx o_i/r_i$. Consequently, the marginal failures cancel out, leaving only the combined failures and the various risk sets. This implies that with heavy censoring,

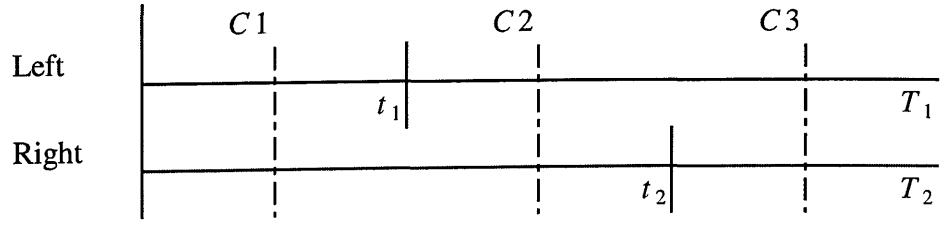
the contribution of the marginal failures diminishes and the estimator relies largely on the risk sets and the combined failures.

To illustrate the techniques developed, three real bivariate survival data sets are showcased in Sections 4.5.1, 4.5.2 and 4.6.2. First, a description of unordered (non-recurrent) events is presented in the context of randomized clinical trials.

### 4.3.2. **Paired Organs**

The human anatomy includes many paired parts or organs and such pairs are inevitably associated to some extent. The case of paired organs concerns two event times for the same individual, and thus only one censoring variable applies, if censoring is by death. Consider a study of times to failure of non-simultaneous joint replacement of the left and right hips of an osteoarthritis patient. Suppose $T_1$ is the time to failure of the left hip, and $T_2$ is the time to failure of the right hip. Figure 4.1 shows the possible scenarios for the paired case on a total time scale, where $C1$ to $C3$ represent possible censoring scenarios.

Figure 4.1: Various censorings for the paired case, an example for $T_1$ and $T_2$ of left and right hips respectively.




For scenario $C1$, the patient is censored for both events and $t_1 = t_2 = c_1$, while for $C2$, where $t_1 < c_2 < t_2$, the left hip failure occurred at $t_1$ but the right hip was censored at $t_2 = c_2$. Meanwhile, in scenario $C3$, both events occurred when censoring time exceeds both $T_1$ and $T_2$: $t_1 < t_2 < c_3$ as shown above. An example of a study

involving the survival of paired organs is provided by a case of non-simultaneous bilateral hip fractures, which reportedly occur, on average, five years apart from each other (Gaumetou, Zilber and Hernigou, 2011). For a hip fracture, an artificial hip joint replacement is the common treatment, and the time from hip replacement to its failure in each case, is also of interest. An example of such a study is illustrated in Section 4.5.1. In general, five of the major paired organs often studied are the breasts, lungs, kidneys, testes and ovaries (Roychoudhuri, Putcha and Moller, 2006). The importance of paired organ analysis justifies its inclusion as a case to be examined in this thesis.

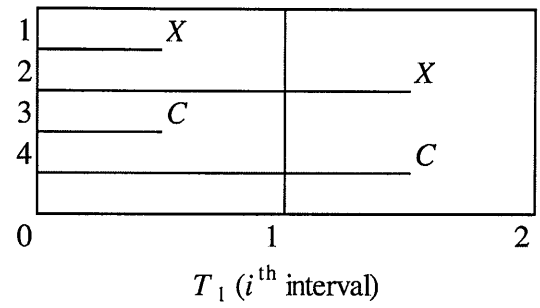
### 4.3.3. **Generally Related Events or Indicators**

In clinical trials, often there are different measurements or indicators used to assess the effect of treatment on patients. For example, a cancer patient may be assessed on time to disease progression, $T_1$, or time to death itself, $T_2$. In this thesis, such a case is termed generally as having related events or indicators. Another common case involves time to a certain event (for example death) for subjects who are related, such as twins or married couples or co-workers, who have been exposed to a similar environment. Say $T_1$ is the lifetime of person 1 and $T_2$ is the lifetime of person 2. In a study of the lifetimes of married couples, a clear choice for the assignment of $T_1$ and $T_2$ is often determined by gender. However, in the case of twins, $T_1$ may be artificially assigned to the elder sibling and $T_2$ to the younger; the choice can also be based on other factors deemed more appropriate. The basis for selection has its importance in the comparison of analyses and thus should be explicitly described by researchers.

Such cases therefore involve independent censoring variables; say $C_1$ and $C_2$ which give different numbers for the subjects at risk of individual failures, as well as for those at risk of both failures: $r_{1C} \neq r_{2C} \neq r_{12C}$, for control treatment. Similar

expressions are applicable to the experimental group. An example of this general censoring for the first failure for interval-censored data is depicted in Figure 4.2 where $X$ and $C$ indicate the occurrence of an event and a censoring respectively.

Figure 4.2: Example of possible outcomes for the first failure, as measured by $T_1$ for generally related indicators.




For each subject at each $i^{th}$ interval, there are four possible outcomes as indicated by the numbers on the left. As shown above, for outcome 1, the patient failed during interval 1, outcome 2: survived interval 1 and failed during interval 2, outcome 3: censored during interval 1, and outcome 4: survived interval 1 and was censored during interval 2. Note that only subjects with outcomes 1, 2, and 4 are considered to be at risk during interval 1 since subjects with outcome 3 have been censored before completing the $1^{st}$ interval. The above diagram only shows outcomes for event 1 relating to $T_1$, and for each of these outcomes, its corresponding outcome for event 2 relating to $T_2$ can also take any of the four possibilities. Note that for the $1^{st}$ interval, outcomes 2 and 4 are considered as successes. There are, in all, nine possible outcomes which contain both outcomes relating to $T_1$ and $T_2$ at any $i^{th}$ interval of $T_1$, and $j^{th}$ interval of $T_2$, as displayed in Table 4.1 earlier.

### 4.3.4. **Progression-Free Survival**

In recent years, progression-free survival (PFS) has been increasingly accepted as the surrogate endpoint for overall survival (OS) in oncology trials. OS is defined as the time from randomization until death from any cause, while PFS refers to the time from randomization until tumour progression (TTP) or death from any cause, whichever occurs first. TTP considers the time from initial therapy to first evidence of tumour progression (either objective radiographic documentation or clinical deterioration). For more precise definitions of PFS and TTP, the reader is referred to the guidelines for criteria such as RECIST (Therasse et al., 2000, and Eisenhauer et al., 2009) for solid tumours.

An example of a PFS study is available in Lim et al. (2011) where the primary objective was the assessment of the PFS rate after 12 weeks of treatment with Cetuximab plus irinotecan in pretreated metastatic colorectal cancer patients. Secondary objectives included further evaluation of PFS, time to treatment failure (TTF) and overall survival time. PFS was defined as the time from the first study medication to the first observation of radiologically confirmed disease progression, symptomatic deterioration leading to discontinuation of the study treatment (unless imaging confirmed absence of progressive disease), or death due to any cause.

Increasing interest in PFS analysis is evident from the number of papers published within the past decade. A search in the Web of Science for "progression-free survival" in the title shows that between years 1990 and 2000 only 13 of 20 articles published were in the subject area of Oncology. As of late March 2011, a similar search between years 2001-2011 reveals a dramatic ten-fold increase with 128 of 188 articles published in Oncology. The prominence of PFS analysis justifies its consideration in this research study.

## 4.4. Estimation of an Overall Treatment Effect

As mentioned in Chapter 2, a key question in a clinical trial involving multiple endpoints concerns the magnitude of the overall treatment advantage. Suppose the parameter of interest corresponding to treatment is the log hazard ratio denoted by $\theta$. Under the null hypothesis that the two treatment groups have identical survival experience, the experimental has zero treatment effect and the proportional hazards assumption is true. There exists a common treatment advantage, $\theta_1 = \theta_2 = \theta$ and under $H_0$: $\theta = 0$; hence p-values are always valid. However, under the alternative, $H_1$: $\theta_1 = \theta_2 = \theta$, but $\theta \neq 0$. The logrank test is efficient in detecting such a proportional hazards alternative. When the assumption of equal treatment effect is met, the complex multivariate problem of analyzing the multiple endpoints is simplified to the univariate problem of comparing the common effect across the treatments. Even if the equality assumption is not met, the power should be good if the spread of $\theta$ is reasonably small.

The estimated common treatment advantage is given by $\hat{\theta} = w_1\hat{\theta}_1 + w_2\hat{\theta}_2$ where $w$ is some weighting with subscripts 1 and 2 for the 1st and 2nd events respectively, and $w_1 + w_2 = 1$. Two ways can be employed to estimate a common treatment advantage; the simplest being to take the ratio of the global score statistic to its variance: $\hat{\theta}^* = Z^*/V^*$. As for its marginal counterparts, the variance of $\hat{\theta}^*$ is simply given by $1/V^*$. From the expressions for $Z^*$ and $V^*$ in Section 2.1.1, $Z^* = Z^+V^+/\text{var}(Z^+)$ and $V^* = V^{+2}/\text{var}(Z^+)$. The effect of var($Z^+$) which contains the covariance cancels out, leaving only the ratio of a sum of $Z$s to a sum of $V$s, and hence $\hat{\theta}^* = (Z_1 + Z_2)/(V_1 + V_2)$. This estimate is equal to $w_1(Z_1/V_1) + w_2(Z_2/V_2)$ for the weights

$$(w_1, w_2)_* = \left( \frac{V_1}{V_1 + V_2}, \frac{V_2}{V_1 + V_2} \right). \tag{4.9}$$

Notice that the weighting in equation (4.9) is a direct form of the proportion of the variances and is ideal for independent endpoints.

Alternatively, when endpoints are not independent, it is possible to derive an optimal weighting (Wei and Johnson, 1985), which yields the smallest variance out of all weighted averages of $\theta_1$ and $\theta_2$. From the common formulation: $\hat{\theta} = w_1\hat{\theta}_1 + w_2\hat{\theta}_2$, the variance is given by

$$\text{var}(\hat{\theta}) = \frac{w_1^2}{V_1} + \frac{w_2^2}{V_2} + \frac{2w_1 w_2 C_{12}}{V_1 V_2},$$

where $V_1 = \text{var}(Z_1)$, $V_2 = \text{var}(Z_2)$ and $C_{12} = \text{cov}(Z_1, Z_2)$ as described in Section 4.3.1. Let the variance of $\hat{\theta}$ be a function of $w_1$:

$$f(w_1) = \text{var}\,\hat{\theta} = \frac{w_1^2}{V_1} + \frac{(1-w_1)^2}{V_2} + \frac{2w_1(1-w_1)C_{12}}{V_1 V_2}.$$

Upon simple differentiation with respect to $w_1$, and setting its derivative to zero, the optimal weighting is given by

$$(w_1, w_2) = \left( \frac{V_1 - C_{12}}{V_1 + V_2 - 2C_{12}}, \frac{V_2 - C_{12}}{V_1 + V_2 - 2C_{12}} \right). \tag{4.10}$$

Using the weighting in equation (4.10), the variance of $\hat{\theta}$ is given by

$$\text{var}(\hat{\theta}) = \frac{V_1 V_2 - C_{12}^2}{V_1 V_2 (V_1 + V_2 - 2C_{12})}. \tag{4.11}$$

This variance is the ratio of unity minus the square covariance to the sum of variance-covariance. It is to be noticed that equation (4.9) is identical to equation (4.10) when the covariance is zero. For comparison purposes, both standard and optimal estimates of the overall treatment advantage, $\hat{\theta}*$ and $\hat{\theta}$ respectively, are reported in the data analyses performed in the subsequent sections. It is to be recalled from equation (1.16) that $\theta = -\log\{-\log S_E(t)\} + \log\{-\log S_C(t)\}$ and a positive value indicates superiority of treatment $E$ over treatment $C$. In this thesis, $\theta = -\beta$ where $\beta$ is the coefficient of the regression, given by the SAS PROC PHREG procedure, when $E$ is given a bigger coding integer than $C$. For example, treatment $E$ is coded as 1, while $C$ is coded as 0, as is commonly practiced. An example is shown later in Chapter 6.

### 4.5. Application to Real Data: Non-recurrent Events

The proposed method is illustrated by application to two data sets from previous clinical trials: hip revision and cancer. Prior to any processing, each raw data set is examined using the existing standard survival analysis tools which are described in the next section. Upon gaining more knowledge about the data set itself, the actual procedure of the proposed method is carried out. The bivariate survival endpoints are first categorized into multiple intervals; as if they were interval-censored data and the marginal estimates of the treatment effect can be obtained from the expression $\theta \approx Z/V$. The covariance is then obtained directly from the two survival endpoints, say $T_1$ and $T_2$, using equation (4.6) as derived in Section 4.3.1. The following analyses have been conducted using specially written SAS programs. The programs were written primarily to enable calculation of $Z$, $V$, $C_{12}$ and consequently the correlation estimate

$\hat{\rho}$. For application to the real data, the significance test level is arbitrarily set at 2.5% (one-sided) for convenience.

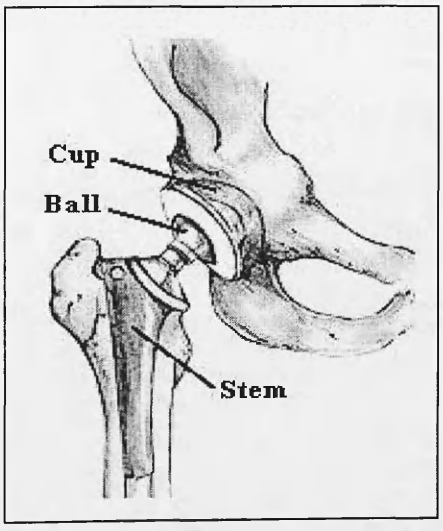
The selection of specific cut-off points to define the time intervals is aspect important in ensuring minimal loss of information while achieving a reasonable covariance estimate, $C_{12}$. Coarse intervals may result in loss of information, and overestimation of the treatment effect, as $\hat{\theta} = Z / V$. Finer intervals may contain more information, but may lead to less reliable parameter estimation if the number of failures in each interval is too small. How do we strike the right balance? In this study, the interval setting is based on equal distribution of failures for each event type. To determine an appropriate number of intervals for analysing survival data using our method, an exploration is undertaken in the subsequent sections. In the meantime, for the purpose of illustration, a convenient choice of intervals is made for each data set. A step-by-step explanation is given for the first data set, while others are described more concisely. A complete set of SAS codes for the analysis of real data, using the proposed method, is given in Appendix A.

### 4.5.1. **Paired Organs: Hip Replacement Revision**

A hip replacement data set from Young Patient study, which is currently on-going at the Centre for Hip Surgery, Wrightington Hospital, is used to illustrate the method for paired organs. The data have been collected over 40 years from patients who have had revision of artificial hip joints, known as Charnley low-friction arthroplasty, LFA (Wroblewski and Siney, 1992). As with a natural hip, the artificial joint is subject to wear and tear; hence requires a revision. Revision is defined as exchange or removal of one or both components consisting of a plastic cup and a metal stem.

In this analysis, only 342 bilateral patients who had revisions on both hips are considered. Since the data were not taken from a randomized trial, the factor of interest will be taken to be the positioning of the cup in relation to the acetabulum: a "socket" formed by the cavity in the pelvic bone. We consider treatment 1 for when it is located in the acetabulum (Medial) and treatment 2 for when the cup is located either on the rim of the acetabulum (Rim) or when a part of the cup is not supported by the bone (Uncovered). There should be no treatment difference in this setting as the two "treatments" are generally considered to be equivalent. A schematic diagram showing the basic components of Charnley's LFA, namely the stem, ball and cup, is given in Figure 4.3.

Figure 4.3: A schematic diagram of the basic components of Charnley's LFA.




The hip data contain the following variables: ID = patient's identification, Hip = hip number (1 = 1st hip, 2 = 2nd hip), Opdate = time of the hip replacement surgery for Hip 1 or 2, Revdate = time of the hip revision, or last follow-up time if the hip revision does not occur, Status = event status (1 = revision, 0 = censored) and Cup positions (1 = Medial, 2= Rim/Uncovered) are taken as treatments. It is noted that the

hips were not revised simultaneously and the ordering is by time of hip revision: the time to revision of the first hip is considered as $T_1$, while that of the other hip is $T_2$.

Prior to further preparation, the raw data were first examined for survival functions, median survival times, hazard functions and censoring proportion by using the PROC LIFETEST procedure and fitting to Cox's model, via the PROC PHREG procedure. The Kaplan-Meier plots of survival distributions for $T_1$ and $T_2$ are given in Figure 4. 4, and Figure 4.5 respectively.

Figure 4.4: Survival distribution for time to left hip revision, $T_1$ in years, stratified by the cup position (cup_pos).

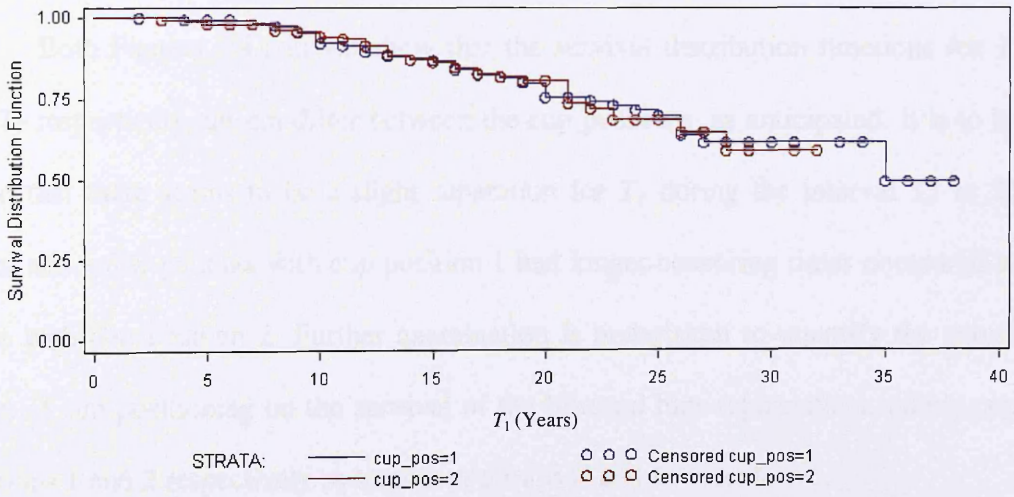
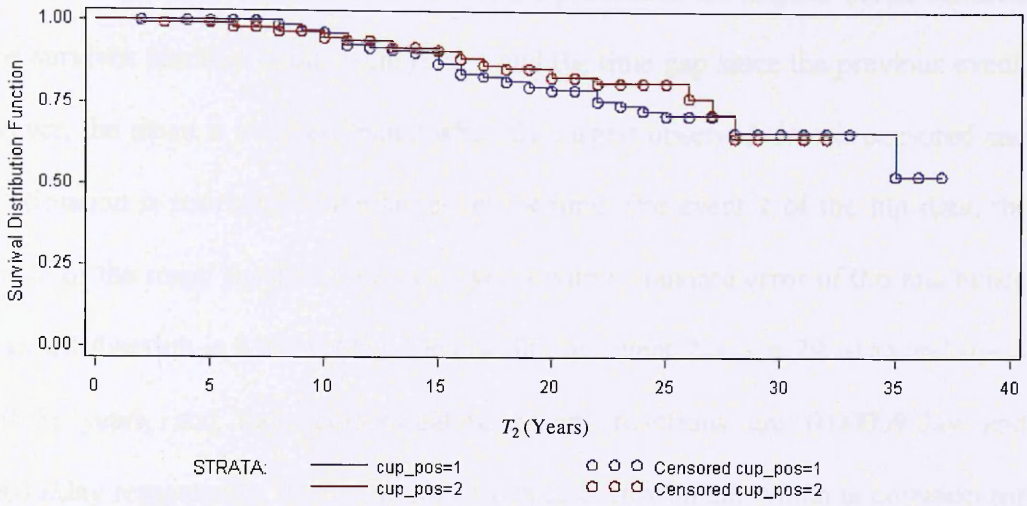


Figure 4.5: Survival distribution for time to right hip revision, $T_2$ in years, stratified by cup position (cup_pos).




Both Figures 4.4 and 4.5 show that the survival distribution functions for $T_1$ and $T_2$ respectively, do not differ between the cup positions, as anticipated. It is to be noted that there seems to be a slight separation for $T_2$ during the interval 15 to 25 years, and a few patients with cup position 1 had longer censoring times compared to those with cup position 2. Further examination is undertaken to quantify the actual effect of cup positioning on the survival of the bilateral hips replacement, taking cup positions 1 and 2 respectively as treatment groups $C$ and $E$, hereafter.

Stratifying by the treatment group for individual event, the median survival times, $m_{Gm}$ and estimated mean survival times, $\mu_{Gm}$ are obtained from the former procedure, $G = E$, $C$ and $m = 1, 2$. By definition, half of the patients have event times longer than the median. Assuming an exponential distribution, the hazard function of patients on $C$ for event 1, $\lambda_{C1} = (log\ 2)/m_{C1}$ and similarly for those on $E$. In this case, the median survival for patients on $C$ is 35 years, and hence the hazard function is 0.00005/day. Alternatively, or when a median does not exist (due to heavy censoring),

such as for the case of hip patients on $E$ in Figure 4.4, the estimate of the mean can be used to find the hazard function where $\lambda_{E1} = 1/\mu_{E1}$.

In SAS, the mean is a summation of the product of the Kaplan-Meier estimate of the survivor function at the event time $t_i$ and the time gap since the previous event. However, the mean is underestimated when the largest observed time is censored and the estimation is restricted to the largest event time. For event 1 of the hip data, the estimate of the mean survival on $E$ is 24 years with a standard error of 0.6 and hence the hazard function is 0.00011/day. Meanwhile, for event 2, $\mu_{C2} = 29$ (0.8) and $\mu_{E2} =$ 25 (0.5) years, and their corresponding hazard functions are 0.00009/day and 0.00011/day respectively. These hazards are indeed very small, which is common for such a low risk event as that of hip replacement failure. A summary of the censoring proportions stratified by event type (hip) is presented in Table 4.2.

Table 4.2: Summary for hip replacement revision data of 342 bilateral patients, stratified by hip (left and right).

| Stratum | Observations | Events | Censored (Percent) |
|---|---|---|---|
| 1 | 342 | 80 | 262 (77%) |
| 2 | 342 | 64 | 278 (81%) |
| Total | 684 | 144 | 540 (79%) |

As shown in Table 4.2, a large proportion of the observations are censored (about 80%). A detailed description is now provided for the proposed method. First, an equal distribution of failures for each event is obtained by ranking each failure time and dividing equally into $k$ intervals. For simplicity, $k = 2$ is chosen in this detailed illustration. This is achieved by using the PROC UNIVARIATE procedure specifying the percentile points (50 and 100 for $k = 2$) for each $T_1$ and $T_2$. For the hip data, coincidentally the cut-off points are at 15 and 35 years for both $T_1$ and $T_2$. Therefore, the two intervals of equal failures are fixed at (0, 15] and (15, 35] for each $T_1$ and $T_2$ accordingly. Next, the failures and successes are counted for each pair of intervals, which can be presented by multiple 2 x 2 tables, similar to Table 4.1. The summary of failures and number at risk for individual events and paired intervals is given in Table 4.3.

Table 4.3: Count of failures and numbers at risk for both patients on $E$ and $C$ for each pair of intervals $ij$ relating to both $T_1$ and $T_2$, and those relating to individual $i$ and $j$ intervals, for the hip replacement revision data.

| Double failure (No. at risk): $o_{(12),(ij)}(r_{(12),(ij)})$ | | $T_1$ | | | |
|---|---|---|---|---|---|
| | | (0, 15] | | (15, 35] | |
| Treatment | | $E$ | $C$ | $E$ | $C$ |
| $T_2$ | (0, 15] | 10(120) | 7 (137) | 2 (15) | 3 (26) |
| | (15, 35] | 2(12) | 2 (22) | 5 (5) | 7 (9) |
| | | | | | |
| $T_1$ | (0, 15] | (15, 35] | $T_2$ | (0, 15] | (15, 35] |
| $o_{1iE}$ | 20 | 16 | $o_{2jE}$ | 13 | 12 |
| $o_{1iC}$ | 21 | 23 | $o_{2jC}$ | 20 | 19 |
| $o_{1i}$ | 41 | 39 | $o_{2j}$ | 33 | 31 |
| $r_{1iE}$ | 122 | 16 | $r_{2jE}$ | 120 | 12 |
| $r_{1iC}$ | 141 | 26 | $r_{2jC}$ | 138 | 22 |
| $r_{1i}$ | 263 | 42 | $r_{2j}$ | 258 | 34 |
| $q_{1i}$ | 0.169 | 2.639 | $q_{2j}$ | 0.137 | 2.428 |

The top part of Table 4.3 contains the components of combined events, with double failure and number at risk for that specific pair of intervals. For example, in the control group, 2 patients had event 1 failure within the first 15 years, and had event 2 failure within the following 20 years. The other 20 patients in the control group, who were also at risk of failure for the same pair of intervals, did not fail. The lower part of the table lists the marginal failures and numbers at risk for the individual events. For the control group, during the 1st interval (0, 15], only 21 patients failed out of 141 patients who were at risk of event 1 during this interval. At the rightmost column, 19 of 22 patients at risk of event 2 during the 2nd interval (15, 35], had failed. The values of $q_i$ given by - $\log(1-o_i/r_i)$ for each event and interval are also computed accordingly.

Putting the required values from Table 4.3 above into equation (4.6), the estimate of covariance between $Z_{1i}$ and $Z_{2j}$ is directly computed for each pair of intervals, while $Z$ and $V$ are calculated for each event. The results are summarized in Table 4.4.

Table 4.4: Calculated values for covariance, score statistic, and Fisher's information for the hip replacement revision data, using $k = 2$ intervals.

| $C_{12(ij)}$ | | $T_1$ | | $Z_{2j}$ | $V_{2j}$ | | |
|---|---|---|---|---|---|---|---|
| | | (0, 15] | (15, 35] | | | | |
| $T_2$ | (0, 15] | 3.624 | -0.398 | 2.513 | 8.229 | | |
| | (15, 35] | -0.369 | 2.977 | -2.819 | 4.563 | | |
| | $Z_{1i}$ | -1.066 | -3.248 | | | $Z_1 = \sum Z_{1i}$ | -4.315 |
| | $V_{1i}$ | 10.211 | 5.436 | | | $V_1 = \sum V_{1i}$ | 15.647 |
| | | | | $Z_2 = \sum Z_{2j}$ | $V_2 = \sum V_{2j}$ | $C_{12} = \sum C_{12(ij)}$ | |
| | | | | -0.306 | 12.792 | | 5.834 |

The covariances for each pair of $ij$ intervals are listed in Table 4.4: for example, the covariance for incidence of event 1 within the first 15 months and that of event 2 within the next 20 months is -0.369. Summing up the covariances for each of the four pairs of intervals, gives us the estimated covariance, $C_{12}$ at the bottom right corner, 5.834. The score statistics, $Z_1$ and $Z_2$ as well as the Fisher's information $V_1$ and $V_2$ are also determined from the sum for each interval as shown above. Using equation (4.8), the correlation can be estimated accordingly. For the hip revision data analyzed using two intervals, the estimated correlation, $\hat{\rho} = 0.412$. This example demonstrates the capability of the proposed method to provide an estimator for the correlation between two score statistics, arising from interval-censored survival data. The estimates of treatment advantages, $\hat{\theta}^*$ and $\hat{\theta}$ are also obtained via the methods described in Section 4.4, but only the p-values relating to the optimal $\hat{\theta}$ are reported

in this thesis. The same procedure is repeated using five and ten intervals by specifying the percentile points accordingly. The results for analyses using $T_1$ and $T_2$ intervals of two, five and ten are summarized in Table 4.5.

Table 4.5: Results for the hip replacement revision data, using different number of intervals ($k$ = 2, 5 and 10).

| Parameter (s.e.) | @ 2 intervals | @ 5 intervals | @ 10 intervals |
|---|---|---|---|
| $Z_1$ | -4.31 | -3.12 | -2.70 |
| $Z_2$ | -0.31 | 0.91 | 2.23 |
| $Z^*$ | -3.28 | -1.63 | -0.35 |
| $V_1$ | 15.65 | 18.77 | 18.94 |
| $V_2$ | 12.79 | 14.75 | 15.52 |
| $V^*$ | 20.17 | 24.74 | 25.30 |
| $\hat{\theta}_1$ | -0.2758 (0.2528) | -0.1662 (0.2308) | -0.0969 (0.2298) |
| $\hat{\theta}_2$ | -0.0239 (0.2796) | 0.0617 (0.2603) | 0.2625 (0.2538) |
| $\hat{\theta}^*$ | -0.1625 (0.2227) | -0.0659 (0.2010) | 0.0618 (0.1988) |
| $\hat{\theta}$ | -0.1713 (0.2225) | -0.0734 (0.2008) | 0.0492(0.1987) |
| p-value | 0.779 | 0.448 | 0.544 |
| $C_{12}$ | 5.834 | 5.948 | 6.235 |
| $\hat{\rho}$ | 0.412 | 0.357 | 0.364 |

All the estimates of treatment effect are small relative to their standard errors, indicating zero treatment effect. As anticipated, the result clearly shows a lack of overall treatment effect and the p-values are in agreement for all three different numbers of intervals used. The sums of the score statistics derived from this interval-censored data set of 10 intervals are closest to the logrank statistics and their corresponding Fisher's information, as shown in parentheses; $Z_1$ = -2.70 (-0.31), $Z_2$ = 2.23 (3.25), $V_1$ = 18.94 (19.08) and $V_2$ = 15.52 (15.16). The logrank test statistics, given by $Z/\sqrt{V}$, from the standard analysis for the 1st and 2nd events, are 0.94 and 0.32 respectively, while their corresponding values using our method for 10 intervals are

0.38 and 0.40. The values for the standard analysis are obtained via PROC LIFETEST in SAS for illustration purposes. The information $V$ and covariance are quite consistent for all of the three different intervals used, giving the correlation $\hat{\rho} = 0.4$. At this juncture, the choice of $k = 10$ intervals seems to be favourable since it gives the smallest standard error of estimates, the largest information and the closest values of $Z$ and $V$ when compared to the logrank. Nevertheless, the subsequent investigations continue with 2, 5 and 10 intervals until sufficient evidence is obtained to determine the best choice of $k$.

It is to be noted that the hip data are not 'parallel' in that the patients were recruited sequentially, over a period of 27 years (1963 - 1990) and moreover, primary bilateral prostheses may not be implanted on the same day. For example, a patient had the first hip replaced in 1978 and the second hip replaced in 1988, yielding a time-gap of 10 years. Such an effect of non-parallel data may reduce the correlation between two endpoints. However, such an extreme situation (time-gap equal to or greater than 10 years) accounts only for less than 10% of the hip patients and hence the effect may be minimal.

Another feature of the data which may impact in terms of the treatment comparison is that the patients may undergo different procedures with regard to the cup position (treatment group). For the hip data, 78 of 342 patients (23%) had both treatment types, the majority of which were from those with wider time-gap between replacements. The presence of these patients with both treatment types implies that the data are not ideal for the implementation of the method. Nevertheless, for illustration purposes, the treatment type for the first hip was considered for both hips of the same

patient. Alternatively and more appropriately, these 78 patients should either be omitted or a more elaborate analysis should be performed on them.

### 4.5.2. **Related Indicators and PFS: Cancer Data**

To illustrate the proposed method for the cases of related indicators and PFS, a data set from a cancer study in a pharmaceutical company is used. Unfortunately, no further detail is available for disclosure. The trial consists of 330 patients each with two endpoints recorded, $T_1$ and $T_2$, relating to disease progression and death respectively. Plots of survival distribution functions for these endpoints are given in Figure 4.6 and Figure 4.7.

Figure 4.6: Survival distribution function for time to disease progression, $T_1$ in days, for the cancer data, stratified by treatment group (trt = 1, 2).

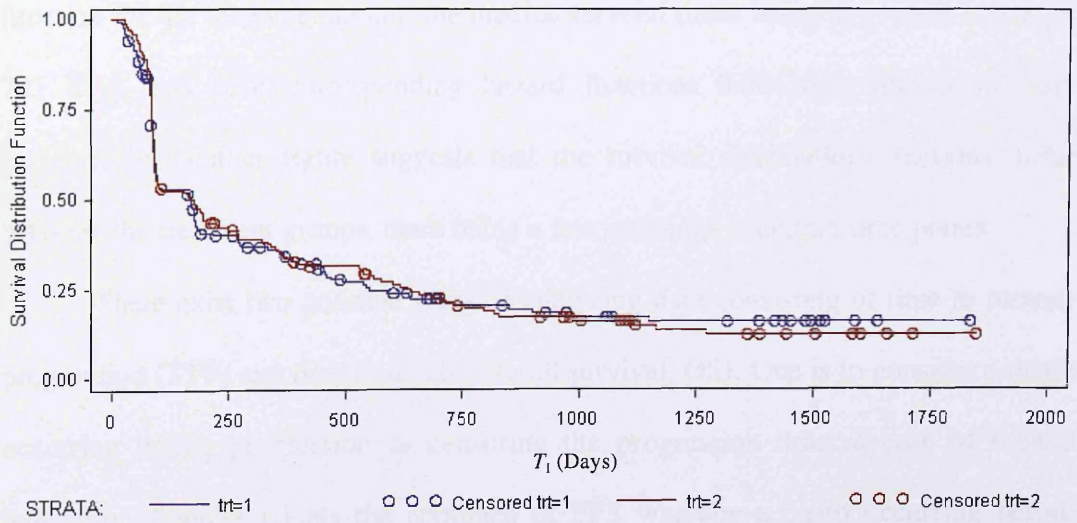


Figure 4.7: Survival distribution function for time to death, $T_2$ in days, for the cancer data, stratified by treatment group.

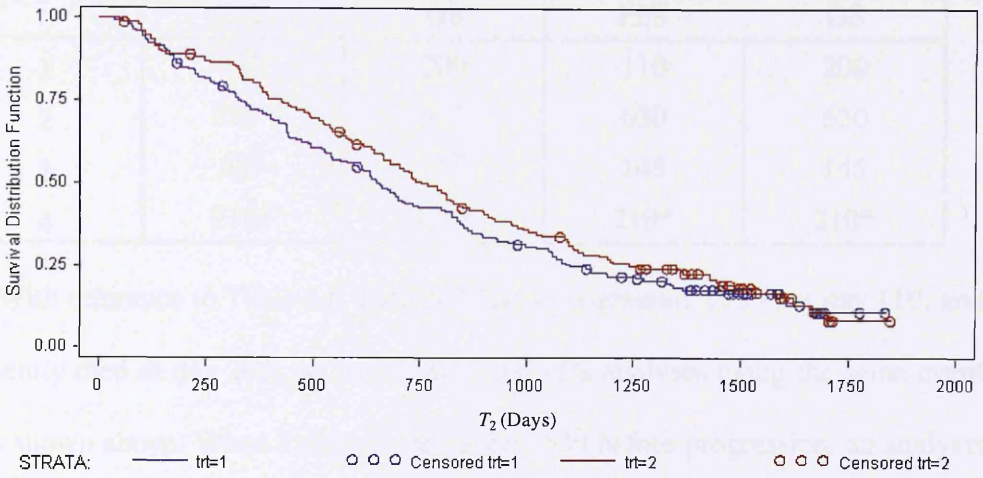



Figure 4.6 shows a steep slope for $T_1$, and the median survival times for patients on $C$ and $E$ respectively are 169 and 170 days. These translate into similar hazard functions, $\lambda_{C1} = \lambda_{E1} = 0.0041$/day, which are rather common for an aggressive type of cancer. Meanwhile, Figure 4.7 depicts a gradual decline in the survival function for the second endpoint, the median survival times being $m_{C2} = 653$ and $m_{E2} = 751$ days, and their corresponding hazard functions 0.0011/day and 0.0009/day respectively. Neither figure suggests that the survival distribution functions differ between the treatment groups, there being a few crossings at certain time points.

There exist two possible ways of analyzing data consisting of time to tumour progression (TTP) and time to death (overall survival, OS). One is to consider a death occurring before progression as censoring the progression time: a case of related indicators. Another adopts the approach of PFS whereby a death occurring before progression will be taken to be a progression event. This implies that PFS = minimum (TTP, OS) which requires some data preparation as illustrated in Table 4.6.

Table 4.6: Example of data sets for related indicators and PFS. *(* indicates censoring)*

| Patient ID | Indicators | | PFS | |
|---|---|---|---|---|
| | TTP | OS | PFS | OS |
| 1 | 110 | 200 | 110 | 200 |
| 2 | 630* | 630 | 630 | 630 |
| 3 | 50* | 145 | 145 | 145 |
| 4 | 210* | 150* | 210* | 210* |

With reference to Table 4.6, Patient 1 had a progression (TTP) at day 110, and subsequently died at day 200, both indicators and PFS analyses using the same event times as shown above. When Patient 2 died at day 630 before progression, an analysis for related indicators considers censoring for TTP at the same time. However, PFS analysis will consider an event for both PFS and OS at day 630. In the case of Patient 3, censoring for TTP occurred at day 50, before death at day 145. The data remain the same for indicators analysis, but PFS considers an event for both PFS and OS at day 145. For Patient 4, when both TTP and OS are censored, a PFS analysis takes the later censoring to be applicable for both PFS and OS, in this case at day 210. In both analyses of indicators and PFS, event 2 is taken as death (OS), while the former and the latter consider TTP and PFS respectively as event 1. The impact of such event dependence imposed by PFS analysis on the cancer data is reflected Table 4.7.

Table 4.7: Summary of the cancer data for the indicators and PFS analyses.

| Stratum | Observations | Indicators | | PFS | |
|---|---|---|---|---|---|
| | | Events | Censored | Events | Censored |
| 1 | 330 | 261 | 69 (21%) | 312 | 18 (5%) |
| 2 | 330 | 271 | 59 (18%) | 271 | 59 (18%) |
| Total | 660 | 532 | 128 (19%) | 583 | 77 (12%) |

Table 4.7 shows that for event 1, the censoring percentage is reduced from 21% to 5% as an event of death (OS) before progression is taken to be an event of progression or death (PFS), too, when considering a PFS analysis. Technically these two types of analysis are expected to yield different results as presented in Table 4.8.

Table 4.8: Results for the cancer data using 2, 5 and 10 intervals based on the analysis of related indicators.

| Parameter (s.e.) | @ 2 intervals (Indicators) | @ 5 intervals (Indicators) | @ 10 intervals (Indicators) |
|---|---|---|---|
| $Z_1$ | -0.30 | -0.91 | 1.51 |
| $Z_2$ | 8.11 | 11.80 | 10.96 |
| $Z^*$ | 5.70 | 7.60 | 8.84 |
| $V_1$ | 55.95 | 63.69 | 65.09 |
| $V_2$ | 47.66 | 62.42 | 66.21 |
| $V^*$ | 75.52 | 88.02 | 93.13 |
| $\hat{\theta}_1$ | -0.0053 (0.1337) | -0.0143 (0.1253) | 0.0232 (0.1239) |
| $\hat{\theta}_2$ | 0.1702 (0.1448) | 0.1890 (0.1266) | 0.1655 (0.1229) |
| $\hat{\theta}^*$ | 0.0755 (0.1151) | 0.0863 (0.1066) | 0.0949 (0.1036) |
| $\hat{\theta}$ | 0.0713 (0.1150) | 0.0855 (0.1066) | 0.0954 (0.1036) |
| p-value | 0.268 | 0.211 | 0.179 |
| $C_{12}$ | 19.267 | 27.294 | 26.905 |
| $\hat{\rho}$ | 0.373 | 0.433 | 0.410 |

From the analysis of the indicators data it is seen that all the estimates of treatment effect for progression, $\hat{\theta}_1$ are smaller than those for death, $\hat{\theta}_2$. The estimates of treatment effect are small relative to their standard errors, with p-values indicating no significance for all choices of intervals. The estimates of covariance between the score statistics are very close (except for 2 intervals), and the correlation estimates are 0.4 (moderate) for all interval settings. The values of $Z$ and $V$ for the cases of 5 and 10 intervals are consistently larger than those for the 2 intervals case. The $Z$ and $V$ for 10 intervals case are the closest to those obtained from the standard logrank test shown in

parentheses: $Z_1$ = 1.51 (1.45), $Z_2$ = 10.96 (9.02), $V_1$ = 65.09 (64.16) and $V_2$ = 66.21 (67.56). This finding, that values from the proposed method matched closely those from the logrank test, is consistent with that for the hip data in Section 4.5.1.

Table 4.9: Results for the cancer data using 2, 5 and 10 intervals based on PFS.

| Parameter (s.e.) | @ 2 intervals (PFS) | @ 5 intervals (PFS) | @ 10 intervals (PFS) |
|---|---|---|---|
| $Z_1$ | 4.28 | 15.14 | 13.24 |
| $Z_2$ | 8.11 | 11.80 | 10.96 |
| $Z^*$ | 8.13 | 17.61 | 15.92 |
| $V_1$ | 50.19 | 69.04 | 74.54 |
| $V_2$ | 47.66 | 62.42 | 66.21 |
| $V^*$ | 64.19 | 85.95 | 92.62 |
| $\hat{\theta}_1$ | 0.0853 (0.1412) | 0.2193 (0.1204) | 0.1776 (0.1158) |
| $\hat{\theta}_2$ | 0.1702 (0.1448) | 0.1890 (0.1266) | 0.1655 (0.1229) |
| $\hat{\theta}^*$ | 0.1267 (0.1248) | 0.2049(0.1079) | 0.1719 (0.1039) |
| $\hat{\theta}$ | 0.1255 (0.1248) | 0.2058 (0.1078) | 0.1723 (0.1038) |
| p-value | 0.157 | 0.028 | 0.049 |
| $C_{12}$ | 25.650 | 34.802 | 36.574 |
| $\hat{\rho}$ | 0.524 | 0.530 | 0.521 |

For PFS, a small estimate of $\theta_1$ is obtained when using 2 intervals with a p-value larger than when using 5 and 10 intervals. Nevertheless, analyses using all the interval setttings yield p-values indicating no significant evidence (at the 2.5% level) to reject the null hypothesis. Moderate correlations of 0.5 are obtained in all three scenarios ($k$ = 2, 5, 10). A bigger $V_1$ compared to $V_2$ across all settings of intervals suggests that there is more information relating to event 1, due to the imposed event 2, as described earlier. The value of $V_1$ when using 10 intervals is closest to that given by the standard logrank analysis as shown in parentheses: $V_1$ = 74.54 (76.71). However, $Z_1$ from the 2 intervals case, 4.28 is closest to the logrank statistic, 6.99.

## 4.6. Recurrent Events

Recurrent events data arise in many diverse fields: numerous examples from medicine, manufacturing and the social sciences are given by Nelson (2003). The methodology for survival analysis of recurrent events has been applied in biostatistics (Genser & Wernecke, 2005), marketing (Bijwaard et al., 2006), sports (Gutierrez et al., 2011) and even in political science (Box-Steffensmeier & Zorn, 2002). Consequently, interest in recurrent events has grown over the recent decades. A quick scan on the Web of Science (search by recurrent in the title) shows more than 13,000 articles from 2001 until March 2011, which is equivalent to the number of articles for the prior two decades. More interesting is the rapid growth in the subject area of Oncology, where in the past decade, over 2,000 articles have been published, double the number in the 90s. Oncology has also overtaken Surgery as the subject area with the most publications on recurrent events in the past decade. This could be driven by the overwhelming public health concerns regarding the widespread increase of cancer around the globe.

There are a few special features distinguishing recurrent events data from the rest. Firstly, the events are ordered within each individual: the second event can only occur after the first event. Censoring applies only to the last observed time for each individual and if the first observation is censored, then the second observation is completely missing. An individual is at risk for one event at any particular time, since these events occur in sequence. Naturally, event times for the same individual are correlated; hence the study of within-subject correlation is of interest. Methods depend on the choice of risk intervals, baseline hazards and techniques used to allow for the subject effect, as described in the next section.

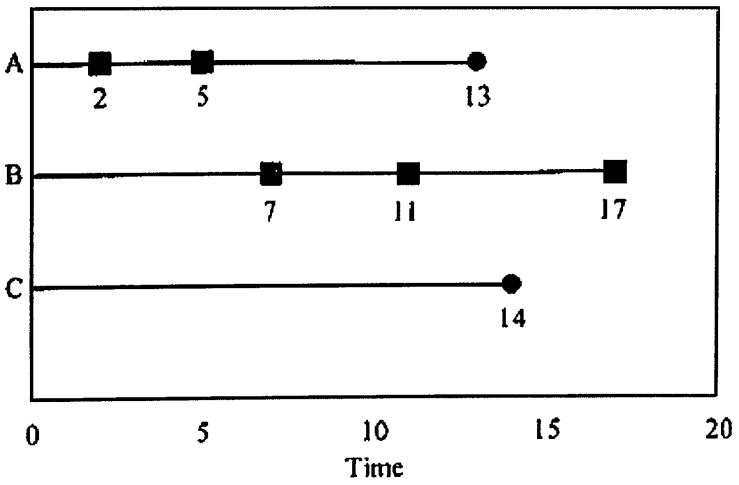
Among the methods for recurrent events are those by Cook & Lawless (1997) and Wei, Lin & Weissfeld (1989). Counting process models provide a powerful framework for the analysis of such event history data when subjects are under continuous observation (Anderson et al., 1993). Random effects models (Abulibdeh, Turnbull and Clark, 1990) and marginal models (Ng and Cook, 1999) are also convenient for adoption in such a setting. Numerous examples exist in other settings when the events of interest are detected only by periodic intensive examination: interval-censored recurrent events. For example, in urology, different types of recurrent superficial bladder cancer tumours may be detected via periodic examination, and in studies of osteoporosis, different types of skeletal changes may be of interest and detectable only by periodic radiographic examination. Prior to describing such data structure in detail, it is important to review the key components that make up the models for recurrent events.

### 4.6.1. **Key Model Components**

In a study, the choice of model is highly dependent upon the questions to be answered. Some of the common questions for recurrent data are as follows. How can the influence of a factor such as treatment be measured? What is the treatment effect for each event? What is the average effect over all events? How can all of the data be used to test for the effect of a factor? How to allow for subject effect? Systematic identification as to how the models differ might assist in understanding and hence, assist in achieving an appropriate model selection. Kelly & Lim (2000) developed key components for a Cox-based recurrent events model. The components concerning the risk intervals, baseline hazards and allowance for within-subject correlation are first described here and are referred to again in later sections.

Risk intervals define when a subject is at risk of the $m^{th}$ event for a given time $T = t_m$. There are three types of risk intervals namely total time, gap time, and counting process. Total time is measured from a specified time, for example the time from the subject's entry into the study or the time from randomization. Gap time concerns the time to an event since the prior event. The term 'total time' is applicable to any form of survival data, while gap time and counting process are specific to recurrent events. However, the counting process is not covered in this thesis. The reader can be referred to texts such as Therneau & Grambsch (2000). To describe the total time and gap time risk intervals, consider a hypothetical example of patients A, B and C, each with recorded recurrences or last follow-up times at the months as shown below. Figure 4.8 to Figure 4.10 are taken from Kelly & Lim (2000) for the purpose of illustration.

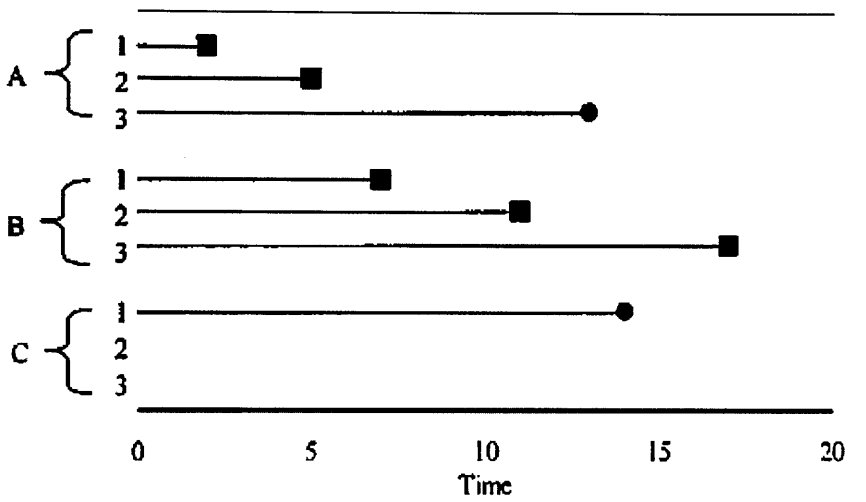
Figure 4.8: Example of three patients with recurrent events.




Patient A has two recurrences at months 2 and 5, before being censored at month 13, while patient B has three recurrences at months 7, 11 and 17. However, patient C has no recurrence until being censored at month 14. The following figures

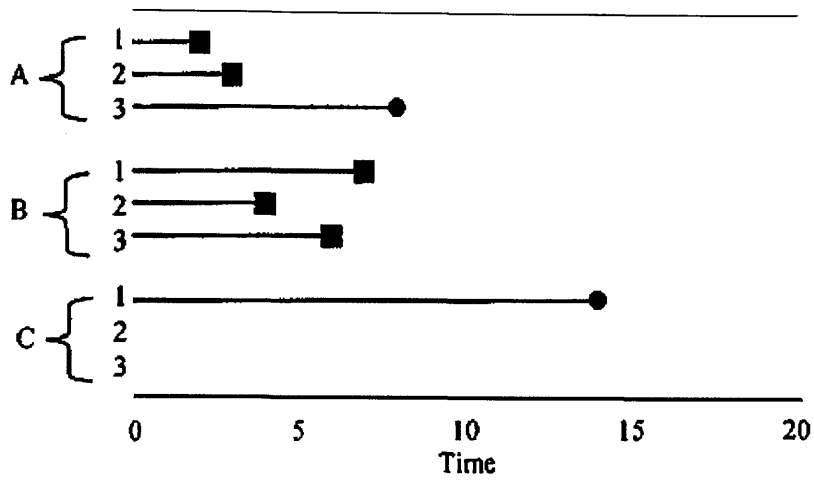
depict the two different risk intervals for this scenario, where the numbers on the left hand side indicate the recurrences.

Figure 4.9: Total time risk intervals for the scenario depicted in Figure 4.8.




A total time model considers each recurrence from the beginning of the study or randomization, therefore an $m^{th}$ event time is always greater than or equal to the $(m-1)^{th}$ event time, as shown above. By definition, it involves cumulative time to event, which is completely different from gap time. It is to be noted that the total time approach is also applicable to non-recurrent events such as those for paired organs and related indicator data, since the ordering of events is not actually used. The recurrence time can also be measured as gap time, as is described next.

Figure 4.10: Gap time risk intervals for the scenario depicted in Figure 4.8.




Gap time is defined as the time since the previous event; for example, patient A has a gap time of 3 months for the 2nd recurrence. In terms of the total time convention used earlier, patient A has a total time of 5 months for the 2nd recurrence as measured from the time of randomization. For gap time, the time scale is no longer time-since-randomization because the clock restarts after each occurrence, as depicted in Figure 4.10. Therefore, for gap time, it is possible for $T_2$ to be smaller than $T_1$, whereas $T_2 \geq T_1$ for total time.

For interval-censored recurrent events, there are only six ways of combining outcomes of which the subjects are considered at risk of the 1st event during the $i^{th}$ interval: *FF*, *FS*, *FM*, *SS*, *SF*, and *SM* (as shown in Table 4.1). When a subject is censored for the 1st recurrence, then the 2nd recurrence is no longer possible. Therefore, there are in that case, only four combined outcomes possible for $T_2$: *FF*, *FS*, *SS*, and *SF*. This means a substantial loss of information about $T_2$ especially if there is a large proportion of censored $T_1$.

Depending on data preparation, both total time and gap time can be fitted by the same algorithm; the choice of model should answer the question of interest. For example, total time is the logical choice if the interest is in modelling the full time

course of the recurrent events, such as evaluating whether treatment is effective for the $m^{th}$ event since the start of treatment. Meanwhile, the gap time model is appropriate when the goal of the analysis is to determine whether treatment is effective for the $m^{th}$ event since the time of the prior event. An example is to be found in assessing whether treatment is effective in delaying the first infection (say, after a surgery) but not for subsequent recurrences.

Another key component of a recurrent events model is baseline hazards which relate to stratification of data. As briefly decribed in Section 1.6 earlier, stratification implies that patients in each stratum have a different baseline hazard function, but all other explanatory variables satisfy the proportional hazards assumption within each stratum. A common baseline hazard is assumed if no stratification is applied to the data modelling. In the case of recurrent events, when data are stratified by event type, for example when the $1^{st}$ recurrence is in stratum 1 and the $2^{nd}$ recurrence is in stratum 2, the baseline hazards are event-specific. In other words, each stratum is fitted by a model separately, thus giving a marginal estimate of the parameter of interest. Further description and examples of such models are given in Section 6.1.1.

The last key component concerns the subject effect which is an intrinsic factor. Apart from the observed explanatory variables and factors that influence the outcomes for a subject with regard to treatment given, naturally there also exist random chances as well as unobserved characteristics of a subject. When there are repeated observations from a subject, such as recurrent events, the unobserved characteristics of the subject may be taken into account. Allowance for subject effect which represents such characteristics can be achieved by using a marginal model (as with the approach taken here), frailty model or copula. The first is described in Section 6.1.1 while the other two are out of scope.

### 4.6.2. **Application to Bladder Cancer Data**

To illustrate the analysis of a recurrent events case, a bladder cancer data set based on a study conducted by the Veterans Administration Cooperative Urological Research Group, is used. The full data set is listed in Wei, Lin, and Weissfeld (1989). The study comprises 86 patients with superficial bladder tumours, which were removed transurethrally when the patients entered the study; 48 were randomized into the placebo group (control), and 38 were randomized into the thiotepa group (experimental). The majority of patients experienced multiple recurrences of tumours during the study, and new tumours were removed at each visit.

The original data set contains the first four recurrences of the tumour for each patient, and each recurrence time was measured from the patient's entry time into the study. However, our analysis setting is limited to the first and second recurrences, with only one covariate, that is the treatment group. The bladder cancer data consist of the following variables: ID = patient ID; Trt = treatment group (1 = placebo and 2 = thiotepa); $T_1$ = time to first recurrence; $T_2$ = time to second recurrence (total time); $T_{2G}$ = time to second recurrence (gap time); cens1 is the censoring indicator for $T_1$ (1 = event and 0 = censored) and similarly, cens2 for that of $T_2$. The bladder data as used in this study are presented in Table 4.10.

Table 4.10: Tumour recurrence data extracted from Wei, Lin & Weissfeld (1989), presented in total time ($T_1$, $T_2$) and gap time ($T_{2G}$).

| ID | Trt | $T_1$ | cens1 | $T_2$ | $T_{2G}$ | cens2 | ID | Trt | $T_1$ | cens1 | $T_2$ | $T_{2G}$ | cens2 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 0 | 0 | 0 | 0 | 0 | 44 | 1 | 3 | 1 | 15 | 12 | 1 |
| 2 | 1 | 1 | 0 | 1 | 0 | 0 | 45 | 1 | 59 | 0 | 59 | 0 | 0 |
| 3 | 1 | 4 | 0 | 4 | 0 | 0 | 46 | 1 | 2 | 1 | 15 | 13 | 1 |
| 4 | 1 | 7 | 0 | 7 | 0 | 0 | 47 | 1 | 5 | 1 | 14 | 9 | 1 |
| 5 | 1 | 10 | 0 | 10 | 0 | 0 | 48 | 1 | 2 | 1 | 8 | 6 | 1 |
| 6 | 1 | 6 | 1 | 10 | 4 | 0 | 49 | 2 | 1 | 0 | 1 | 0 | 0 |
| 7 | 1 | 14 | 0 | 14 | 0 | 0 | 50 | 2 | 1 | 0 | 1 | 0 | 0 |
| 8 | 1 | 18 | 0 | 18 | 0 | 0 | 51 | 2 | 5 | 1 | 5 | 0 | 0 |
| 9 | 1 | 5 | 1 | 18 | 13 | 0 | 52 | 2 | 9 | 0 | 9 | 0 | 0 |
| 10 | 1 | 12 | 1 | 16 | 4 | 1 | 53 | 2 | 10 | 0 | 10 | 0 | 0 |
| 11 | 1 | 23 | 0 | 23 | 0 | 0 | 54 | 2 | 13 | 0 | 13 | 0 | 0 |
| 12 | 1 | 10 | 1 | 15 | 5 | 1 | 55 | 2 | 3 | 1 | 14 | 11 | 0 |
| 13 | 1 | 3 | 1 | 16 | 13 | 1 | 56 | 2 | 1 | 1 | 3 | 2 | 1 |
| 14 | 1 | 3 | 1 | 9 | 6 | 1 | 57 | 2 | 18 | 0 | 18 | 0 | 0 |
| 15 | 1 | 7 | 1 | 10 | 3 | 1 | 58 | 2 | 17 | 1 | 18 | 1 | 0 |
| 16 | 1 | 3 | 1 | 15 | 12 | 1 | 59 | 2 | 2 | 1 | 19 | 17 | 0 |
| 17 | 1 | 26 | 0 | 26 | 0 | 0 | 60 | 2 | 17 | 1 | 19 | 2 | 1 |
| 18 | 1 | 1 | 1 | 26 | 25 | 0 | 61 | 2 | 22 | 0 | 22 | 0 | 0 |
| 19 | 1 | 2 | 1 | 26 | 24 | 1 | 62 | 2 | 25 | 0 | 25 | 0 | 0 |
| 20 | 1 | 25 | 1 | 28 | 3 | 0 | 63 | 2 | 25 | 0 | 25 | 0 | 0 |
| 21 | 1 | 29 | 0 | 29 | 0 | 0 | 64 | 2 | 25 | 0 | 25 | 0 | 0 |
| 22 | 1 | 29 | 0 | 29 | 0 | 0 | 65 | 2 | 6 | 1 | 12 | 6 | 1 |
| 23 | 1 | 29 | 0 | 29 | 0 | 0 | 66 | 2 | 6 | 1 | 27 | 21 | 0 |
| 24 | 1 | 28 | 1 | 30 | 2 | 1 | 67 | 2 | 2 | 1 | 29 | 27 | 0 |
| 25 | 1 | 2 | 1 | 17 | 15 | 1 | 68 | 2 | 26 | 1 | 35 | 9 | 1 |
| 26 | 1 | 3 | 1 | 6 | 3 | 1 | 69 | 2 | 38 | 0 | 38 | 0 | 0 |
| 27 | 1 | 12 | 1 | 15 | 3 | 1 | 70 | 2 | 22 | 1 | 23 | 1 | 1 |
| 28 | 1 | 32 | 0 | 32 | 0 | 0 | 71 | 2 | 4 | 1 | 16 | 12 | 1 |
| 29 | 1 | 34 | 0 | 34 | 0 | 0 | 72 | 2 | 24 | 1 | 26 | 2 | 1 |
| 30 | 1 | 36 | 0 | 36 | 0 | 0 | 73 | 2 | 41 | 0 | 41 | 0 | 0 |
| 31 | 1 | 29 | 1 | 36 | 7 | 0 | 74 | 2 | 41 | 0 | 41 | 0 | 0 |
| 32 | 1 | 37 | 0 | 37 | 0 | 0 | 75 | 2 | 1 | 1 | 27 | 26 | 1 |
| 33 | 1 | 9 | 1 | 17 | 8 | 1 | 76 | 2 | 44 | 0 | 44 | 0 | 0 |
| 34 | 1 | 16 | 1 | 19 | 3 | 1 | 77 | 2 | 2 | 1 | 20 | 18 | 1 |
| 35 | 1 | 41 | 0 | 41 | 0 | 0 | 78 | 2 | 45 | 0 | 45 | 0 | 0 |
| 36 | 1 | 3 | 1 | 43 | 40 | 0 | 79 | 2 | 2 | 1 | 46 | 44 | 0 |
| 37 | 1 | 6 | 1 | 43 | 37 | 0 | 80 | 2 | 46 | 0 | 46 | 0 | 0 |
| 38 | 1 | 3 | 1 | 6 | 3 | 1 | 81 | 2 | 49 | 0 | 49 | 0 | 0 |
| 39 | 1 | 9 | 1 | 11 | 2 | 1 | 82 | 2 | 50 | 0 | 50 | 0 | 0 |
| 40 | 1 | 18 | 1 | 48 | 30 | 0 | 83 | 2 | 4 | 1 | 24 | 20 | 1 |
| 41 | 1 | 49 | 0 | 49 | 0 | 0 | 84 | 2 | 54 | 0 | 54 | 0 | 0 |
| 42 | 1 | 35 | 1 | 51 | 16 | 0 | 85 | 2 | 38 | 1 | 54 | 16 | 0 |
| 43 | 1 | 17 | 1 | 53 | 36 | 0 | 86 | 2 | 59 | 0 | 59 | 0 | 0 |

It is to be noted that, for recurrent events, a patient with only one recurrence is censored for the 2nd recurrence relating to $T_2$. For example, in Table 4.10, patient ID 6 had the 1st recurrence six months after tumour removal and was followed up for the next four months, therefore cens1 =1, $T_1 = 6$, $T_2 = 10$, $T_{2G} = 4$ and cens2 = 0. The plots of survival distribution function against survival time in months, for the bladder cancer data, are depicted in Figures 4.11 to 4.13.

Figure 4.11: Survival distribution for time to first recurrence, $T_1$ in months, for the bladder cancer data, stratified by treatment group.

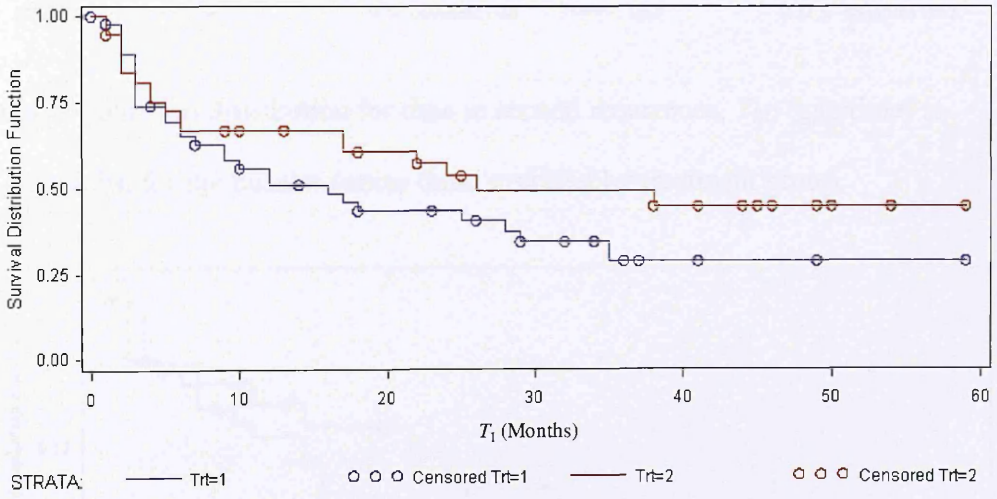


Figure 4.12: Survival distribution for time to second recurrence, $T_2$ (total time) in months, for the bladder cancer data, stratified by treatment group.

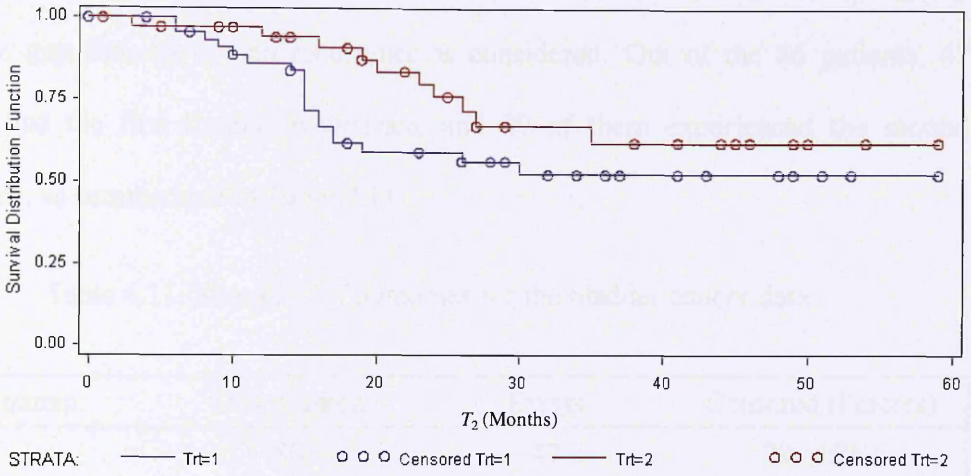
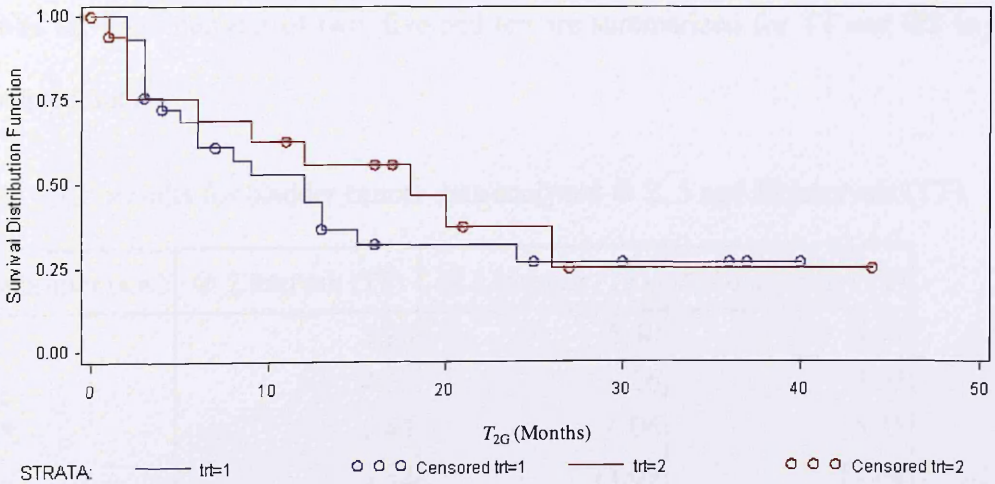



Figure 4.13: Survival distribution for time to second recurrence, $T_{2G}$ (gap time) in months, for the bladder cancer data, stratified by treatment group.




Figures 4.11 and 4.12 for the survival distribution functions using total time, show similar trends indicating some treatment effect on the experimental group (trt = 2). For the 1st recurrence, the median survival times are 16 and 26 months for patients on control and experimental respectively; hence the hazard function for patients on $E$ is 0.0010/day, slightly lower than 0.0014/day for patients on $C$. For the 2nd recurrence,

no median exists, and the means are $\mu_{C2}$ = 23 (1.4) and $\mu_{E2}$= 30 (1.6) years; their corresponding hazard functions are 0.0014/day and 0.0011/day respectively. Figure 4.13 does not indicate a consistent treatment advantage for the experimental group when the gap time to second recurrence is considered. Out of the 86 patients, 47 patients had the first tumour recurrence, and 29 of them experienced the second recurrence, as summarized in Table 4.11.

Table 4.11: Summary of outcomes for the bladder cancer data.

| Stratum | Observations | Events | Censored (Percent) |
|---|---|---|---|
| 1 | 86 | 47 | 39 (45%) |
| 2 | 86 | 29 | 57 (66%) |
| Total | 172 | 76 | 96 (56)% |

Table 4.11 shows heavy censoring for the bladder cancer data for each event. The results for time intervals of two, five and ten are summarized for TT and GT in the subsequent tables.

Table 4.12: Results for bladder cancer data analyzed @ 2, 5 and 10 intervals (TT).

| Parameter (s.e.) | @ 2 intervals (TT) | @ 5 intervals (TT) | @ 10 intervals (TT) |
|---|---|---|---|
| $Z_1$ | 5.96 | 5.40 | 4.47 |
| $Z_2$ | 4.33 | 4.26 | 3.95 |
| $Z^*$ | 6.45 | 6.06 | 5.25 |
| $V_1$ | 11.32 | 11.69 | 11.82 |
| $V_2$ | 7.22 | 7.34 | 7.35 |
| $V^*$ | 11.63 | 11.93 | 11.96 |
| $\hat{\theta}_1$ | 0.5262 (0.2972) | 0.4618 (0.2925) | 0.3782 (0.2909) |
| $\hat{\theta}_2$ | 0.5999 (0.3721) | 0.5806 (0.3691) | 0.5369 (0.3689) |
| $\hat{\theta}^*$ | 0.5549 (0.2932) | 0.5076 (0.2895) | 0.4391 (0.2891) |
| $\hat{\theta}$ | 0.5430 (0.2891) | 0.4877 (0.2851) | 0.4110 (0.2844) |
| p-value | 0.030 | 0.044 | 0.074 |
| $C_{12}$ | 5.510 | 5.659 | 5.771 |
| $\hat{\rho}$ | 0.609 | 0.611 | 0.619 |

The estimates of treatment effect seem to decrease with the number of intervals used, while the p-values consistently indicate non-significance at the 2.5% level. The estimator of covariance yields consistent values, regardless of how many intervals are used, and the correlation is accurate at 0.6 (highly correlated). It can be concluded that a patient with a long time to the 1st recurrence is likely to experience a long time to the 2nd recurrence from the start of study. This is anticipated as $T_2$ contains $T_1$ in TT risk interval.

The analysis using 10 intervals yields the values of $Z$ and $V$ closest to those from the logrank test, as shown in parentheses; $Z_1 = 4.47$ (4.09), $Z_2 = 3.95$ (3.83), $V_1 = 11.82$ (10.99) and $V_2 = 7.35$ (7.06). Unlike the earlier data sets, the bladder cancer data set concerns recurrences, whereby a censored 1st recurrence results in a 2nd recurrence being missed. The impact of such dependence is evident from the values of $V_2$ which are smaller than $V_1$ in all settings: the information contained in $T_2$ being less due to the censoring imposed by $T_1$. Meanwhile, the resulting outputs when applying GT risk interval to the bladder cancer data, are summarized in Table 4.13.

Table 4.13: Results for bladder cancer data analyzed @ 2, 5 and 10 intervals (GT).

| Parameter (s.e.) | @ 2 intervals (GT) | @ 5 intervals (GT) | @ 10 intervals (GT) |
|---|---|---|---|
| $Z_1$ | 5.96 | 5.40 | 4.47 |
| $Z_2$ | 0.84 | 1.14 | 1.14 |
| $Z^*$ | 7.26 | 5.84 | 5.92 |
| $V_1$ | 11.32 | 11.69 | 11.82 |
| $V_2$ | 7.26 | 7.11 | 7.11 |
| $V^*$ | 19.16 | 19.69 | 19.96 |
| $\hat{\theta}_1$ | 0.5262 (0.2972) | 0.4618 (0.2925) | 0.3782 (0.2909) |
| $\hat{\theta}_2$ | 0.1265 (0.3886) | 0.1930 (0.3758) | 0.1608 (0.3751) |
| $\hat{\theta}^*$ | 0.3787 (0.2284) | 0.3604 (0.2192) | 0.2966 (0.2239) |
| $\hat{\theta}$ | 0.3756 (0.2284) | 0.3575 (0.2192) | 0.2952 (0.2238) |
| p-value | 0.050 | 0.051 | 0.093 |
| $C_{12}$ | -0.572 | -0.919 | -0.488 |
| $\hat{\rho}$ | -0.066 | -0.101 | -0.053 |

The estimates of treatment effect $\hat{\theta}_2$ for GT are appreciably smaller than those for TT since there is no carry-over effect from $T_1$; hence a similar trend for $\hat{\theta}$. There is no evidence to suggest any advantage from the experimental treatment, as is evident from the p-values at all interval settings. It is to be noted that the values of $Z_2$ and $V_2$ when using 10 intervals are close to those given by the standard logrank analysis as shown in parentheses: $Z_2$ =1.14 (1.15) and $V_2$ = 7.11 (6.39). This shows that the interval-censored approach works well for the marginal analysis.

The main difference between TT and GT in this case is that GT yields negative covariance between the two score statistics. Mathematically, this is owing to the situation whereby the summation of covariance for each pair of intervals is close to zero. Upon checking for each of the 100 pairs of intervals ($k$ = 10), 17 have positive covariance, 19 negative, 31 zeroes while 33 do not contribute (zero failures within the pair of intervals). This small total value leads to zero correlation, which suggests that

the cumulative treatment effects for the 1st recurrence and time to the next recurrence are not correlated. It can be inferred that a patient with a long time to the 1st recurrence of a tumour is not unlikely to experience a short time to the 2nd recurrence. In short, the time to the 1st recurrence does not predict the time to the next one. This finding is consistent with that reported by Yan et al. (2002) whereby time to bladder tumour recurrence becomes shorter as the number of recurrences increases. Therefore, it is logical that recurrent GT analysis tends to yield negative or close to zero correlation between the score statistics as shown above. It is also uncommon for recurrence events analysis to be limited to a certain $m$ number of events: for example, $m = 4$ in Wei et al. (1989), since events tend to be fewer for surviving subjects.

## 4.7. Discussion

The most important part of this thesis, consists in deriving the estimated correlation coefficient for the complementary log-log model, incorporating interval-censored data structure, has been covered. Straightforward computation to yield the covariance between two score statistics for each pair of intervals was demonstrated with real data sets. Although the scope is only for bivariate survival data, the method can be easily extended for other multivariate data.

The results show that the proposed method works well for the real data sets comprising paired organs, related indicators, progression-free survival and recurrent events. Regardless of the number of intervals, our proposed method provides consistent estimates of the correlation, but treatment effects seem to vary: using fewer intervals tending to overestimate the treatment effects. It is evident that the number of intervals selected has an impact on the standard error of the parameter estimate: finer intervals give smaller standard errors.

As anticipated, the coarsening of the data through reduction in the number of intervals used in the analysis, does impact on the resulting output. Thall & Lachin (1988), too, have also commented that test result could depend on the selection of the number of intervals and the intervals themselves. Basing results from the real data sets, using only two intervals is ruled out since it gives the biggest standard error and seems to overestimate the overall treatment effect, as is to be expected from the use of such coarse intervals. The true measure of the accuracy and suitability of this method as a survival analysis tool is demonstrated in an extensive simulation study in the next chapter.

# Chapter 5. Simulation Study

In Chapter 4, the proposed method for combining bivariate survival endpoints was successfully applied to real data. For convenience, it is now called ZW (Zain & Whitehead) throughout this thesis. The estimator of covariance and the global score test approach derived in Section 4.3.1 worked well for the interval-censored survival data. This chapter now investigates the properties of the estimator and evaluates the accuracy of bivariate tests using simulation.

Design details for the simulation study are developed in Section 5.1, followed by descriptions of key performance measures and combined hypothesis tests. Investigation of the properties of the estimator and evaluation of its accuracy are reported in Section 5.2, on the basis of the simulations performed. Results for each of the six cases investigated in Sections 5.2.1 to 5.2.6, are first given in turn, and followed by a summary of correlation ratios and overall results. A discussion concludes this chapter.

## 5.1. Design of Simulation Study

A full account of the method used to generate simulated data sets, is now given. As earlier described in Section 1.7, the amount of information, $V = \{(u_\alpha + u_\beta)/\theta_R\}^2$ for a one-sided test. The relationship between $V$ and $n$ can be quantified by a constant value $b$: $V = bn$, which will then lead to the sample size needed for the trial. Tang et al. (1989) showed the advantage in setting the sample size based on multiple endpoints, which requires a smaller sample size when compared to a design using only one endpoint. Using the same principle, for a fixed sample size, the treatment advantage, $\theta_R$, at which a given power is achieved based on multiple endpoints, is smaller compared to that using a single endpoint.

In this simulation study, the target type I error rate is 2.5% level (one-sided) and the target power is 90%; $(u_\alpha + u_\beta)^2 = 10.51$. Subscripts 1 and 2 respectively indicate that a marginal parameter relates to the individual event based on $T_1$ and $T_2$, while an asterisk or a subscript 12 indicates a global parameter based on both events. Note that the subscript $m$ is used also to denote event, $m = 1, 2$ and $G$ to denote treatment group, $G = C, E$. The six cases of bivariate survival data, which have been described in the previous chapter, are considered: complete (or uncensored), paired organs, related indicators, progression-free survival (PFS), recurrent events total time (TT) and recurrent events gap time (GT).

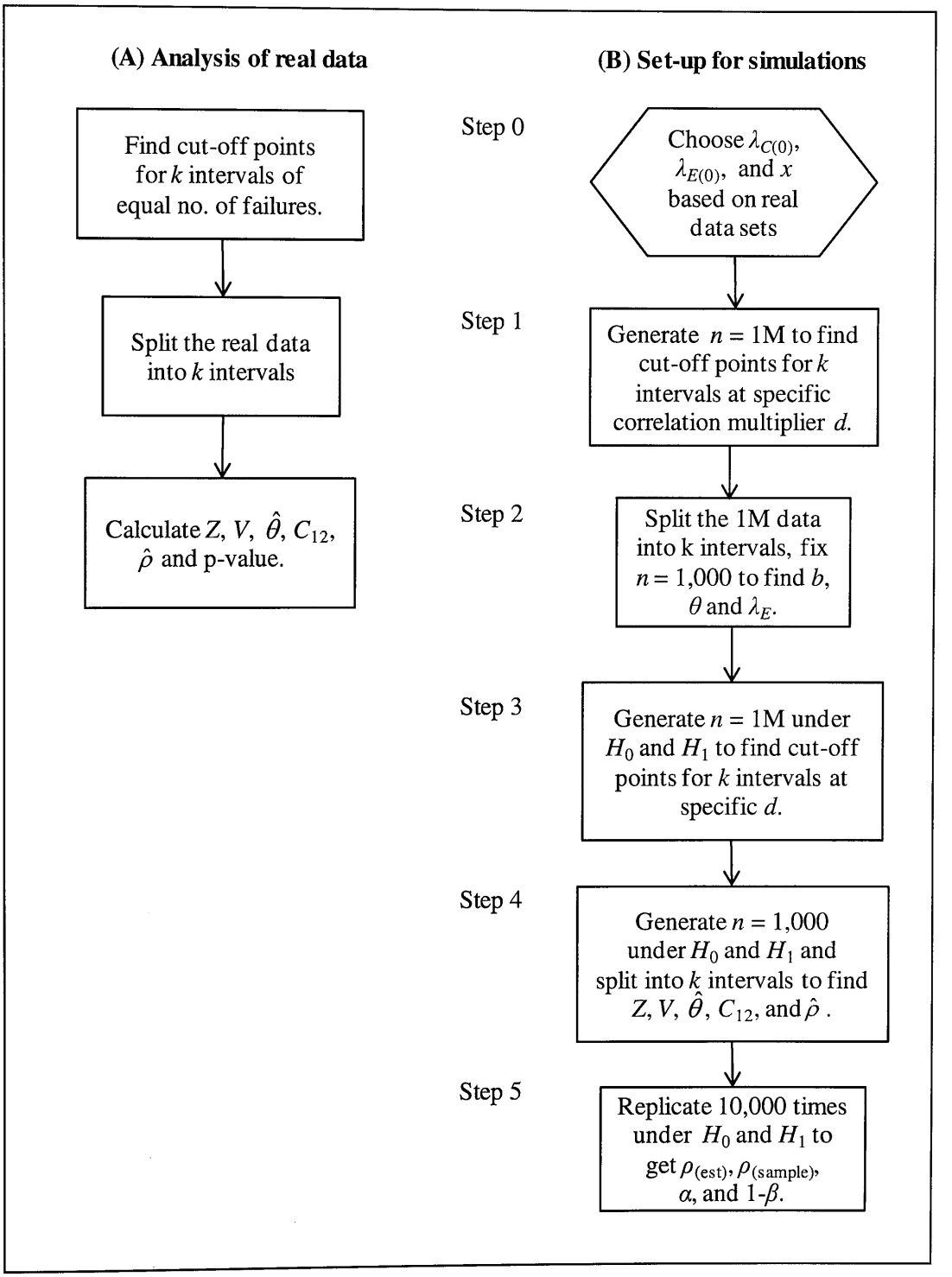
To determine sensible values for the hazard function for simulation purposes, the three real data sets from the previous chapter: hip revision, cancer and bladder cancer, are considered. The case of cancer with the highest hazard of disease progression of 0.004 (Section 4.5.2), is chosen. For simplicity in generating a big data set to estimate the cut-off points and $b$, $\lambda_{E(0)}$ is chosen to be 0.004, while $\lambda_{C(0)}$ is fixed at 0.006 and $\theta_{(0)} = 0.4$. This simulation model could represent an aggressive cancer study with time unit in days, whereby the median survival times for patients on $C$ and $E$ respectively are 115 days (3.8 months) and 173 days (5.8 months). Alternatively, it could be a longer term condition with time unit in months and the median survival times for patients on $C$ and $E$ respectively, are 115 months (9.5 years) and 173 months (14.4 years).

With the exception of the complete case, all the other five cases (listed above) are subjected to their specific censoring rules and censoring variables for simulation. The censoring proportions are set based on the three real data sets used in Sections 4.5.1, 4.5.2 and 4.6.2. However, for the paired data, an alternative censoring proportion of 50% is used since the analysis with 80% censoring ( as observed in the

real data set), was found to be less reliable (inflated type I error). Let the censoring proportion be $x$, so that $x = 0.50$ for the paired case. The percentage values of the censoring proportion ($T_1$, $T_2$) for each simulated case are: paired (50, 50), indicators (20, 20), PFS (8*, 20) and recurrent TT/GT (40, 60*). The censoring proportion without an asterisk is entered directly as $x$ in the simulation program for each $T_1$ and $T_2$. Meanwhile, those with asterisks are further subjected to specific censoring rules, as discussed in their respective sections. The resulting censoring proportions are then verified to match those from the real data sets. For example, in the case of PFS, $T_1$ is amended according to the censoring rule illustrated in Table 4.6. For the recurrent events, the censoring rules applied to the simulated data are described in Section 4.6.2. In other words, each simulation model is designed to simulate each case the closest possible to its corresponding real data set. It is to be noted that a wide range of censoring proportion (20% to 60%) is examined, for thoroughness.

In Chapter 4, the settings of five and ten intervals gave comparable results, hence $k = 5$, 10 are selected for the simulation study. From the simulation results, the most suitable number of intervals will be selected for subsequent analysis and comparison in later chapters. It is to be recalled from Chapter 4 that the procedure for ZW involves splitting the data set into $k$ intervals of equal failures for each event and calculating the quantities of interest for each set of intervals. The procedure for analysis of real data is presented in a flow chart for comparison with the simulation procedure in this section. It is to be noted that the complete simulation procedure is repeated for specific case, hypothesis, and values of $d$ and $k$, as will be decribed next.

Figure 5.1: Process flow charts for analysis of real data and simulation run



The hexagonal box in Figure 5.1 (B) represents the initial preparation, Step 0, which has been described earlier. Steps 1 to 3 are necessary to determine the set-up values for the actual simulation run as shown in Steps 4 and 5. Inputting the hypothetical values of $\lambda_{E(0)} = 0.004$ and $\lambda_{C(0)} = 0.006$, the survival times, $T_1$ and $T_2$ (days) are generated from an exponential distribution, $T_{im} \sim EXP(\lambda_G \exp(s_i))$, where $s_i$ is a subject effect for patient $i$, following a normal distribution $N(0, \sigma^2)$. The standard deviation $\sigma$ of the subject effect is set to be $d(\log \lambda_C - \log \lambda_E)$, where $d$ is a constant multiplier chosen to impose varying degrees of correlation: setting $d = 1$, 5 and 10, creates low, medium and high correlations, respectively.

The Cox's PH requires non-informative censoring such that the censoring is independent of the survival times. Therefore, assuming equal hazards, $\lambda_C = \lambda_E = \lambda$, an overall censoring variable, $C_i \sim EXP(2\lambda y)$ is applied to the whole data set (both patients on $C$ and $E$), where $y = x/\{2(1-x)\}$ and $x$ is the censoring proportion. It is to be recalled that, in general, patient $i$ is censored for an event when $C_i < T_i$ for that event. Further censoring rules, specific to the simulated case will then be applied to the data sets accordingly.

To find the interval cut-off points, a very big sample size, ranging from 10,000 to 10 million, is generated and simulated as the bivariate interval-censored case of interest. A big sample size is necessary to ensure large $V$, which is the amount of information on $\theta$ contained in $Z$ and consequently give the constant $b$, as will be shown next. As described in Section 4.5.1, these cut-off points are obtained by specifying the percentile points. The values of the cut-off points converge satisfactorily when $n = 1$ million, hence only this value of $n$ is shown in Step 1 of the simulation flow chart in Figure 5.1.

Step 2 involves finding specific values for the parameters required for the two hypotheses, including $\theta_R$ under the alternative, while $\lambda_C$ is fixed at 0.006. Based on the useful expression, $V = bn$, and also $V = \{(u_\alpha + u_\beta)/ \theta_R \}^2$, we can find the $\theta_R$ at which a given power is achieved by fixing the sample size $n$, provided we know the constant $b$. Using the output for the interval cut-off points from Step 1, the same data sets, of 10,000 to 10 million patients are then split into the specified $k$ intervals. This procedure is considered to yield the "truth" about the information, $V$, when the value of $b$ converges satisfactorily. The fixed sample size is chosen to be 1,000 to calculate the value of the constant $b$ to mimic a real study. The output of this procedure, when using $k = 5$ intervals, for the complete data is given as an example in Table 5.1.

Table 5.1: Average values for $V^*$ and $b$ computed at various sample sizes for the complete data ($d$ =10, $k$ = 5,).

| Parameter | 10,000 | 100,000 | 1,000,000 | 10,000,000 |
|---|---|---|---|---|
| $V^*$ | 2151 | 21540 | 215260 | 2153136 |
| $b$ (@ $n$ = 1,000) | 0.21508 | 0.21540 | 0.21526 | 0.21531 |
| Duration (minutes) | 0.13 | 0.60 | 6.08 | 58.27 |

As displayed above, the value of $b$ seems to converge at 0.2153 for $n$ = 1 million with acceptable computing time (6 minutes). The corresponding values of $b$ for $d$ = 1, and 5 are 0.3460 and 0.2399 respectively. Once the estimate of $b$ (for $n$ =1,000) is obtained, the $\theta_R$ and $\lambda_E$ can be calculated accordingly, where $\lambda_E = \lambda_C(\exp(-\theta))$. It is to be noted that now $\theta_R = \sqrt{(10.51/1000b)}$ as defined earlier. Plugging in these values for the settings under the null and alternative, the interval cut-off points are determined from a sample size of 1 million as shown in Step 3 of the simulation flow chart. The cut-off points for $T_1$ when $d$ = 10 in this scenario are 4, 40, 343 and 4134 days. These cut-off points translate to intervals such as (0, 4], which means that patients with $0 < T_1 \leq 4$ are grouped in the first interval and those with $T_1 > 4134$ in the

last interval. It is to be noticed that Steps 1 to 3 in Figure 5.1 only show $n$ = 1 million, as it is shown to be the most practical choice of sample size, to give reliable estimates.

It is to be clarified that these process steps involving big data set, is needed only for simulation runs, whereas in the actual analysis of real data set, the process steps are simply as demonstrated in Chapter 4 and shown in the flow chart (Figure 5.1). In Step 4, a sample size of 1,000 patients is simulated under both hypotheses to yield the estimates of $Z$, $V$, $C_{12}$ and $\rho$. Finally, Step 5 involves 10,000 times replication of Step 4 under each hypothesis to obtain the estimates for $\hat{\theta}$, $\rho_{(est)}$, $\rho_{(sample)}$, type I error $\alpha$, and power 1-$\beta$. The best estimates are taken to be the average values from the replications: each of these quantities will be further described in the next section.

The whole process from Step 1 to Step 5 is repeated for each scenario as specific cut-off points are applied to specific values of $d$, $k$, $\theta_R$ and $\lambda_E$ for each case. This translates into 72 different scenarios for the whole simulation study (6 cases, 2 hypotheses, $k$ = 5, 10 and $d$ = 1, 5, 10). It is to be noted that a mandatory censoring of patients in the last interval is imposed to emulate the end of study in a real clinical trial. An example of the input values for all cases is depicted in Table 5.2.

Table 5.2: Example of simulation setting for each case at $d$ = 10 and $k$ = 5 intervals.

| Case | % censored for $T_1, T_2$ | $b$ ($d$ = 10, @ 5 intervals) | $\theta_R$ | $\lambda_E$ |
|---|---|---|---|---|
| Uncensored | N/A | 0.2153 | 0.221 | 0.00481 |
| Paired | 50, 50 | 0.1127 | 0.305 | 0.00442 |
| Indicators | 20, 20 | 0.1394 | 0.275 | 0.00456 |
| PFS | 8*, 20 | 0.1387 | 0.275 | 0.00456 |
| Recurrent (TT) | 40,60* | 0.1043 | 0.317 | 0.00437 |
| Recurrent (GT) | 40,60* | 0.1259 | 0.289 | 0.00449 |

The complete or uncensored case has the biggest $b$, resulting in the smallest $\theta_R$ when $n$ is fixed: as anticipated since no censoring is involved. With a heavy censoring for the paired case, it is to be noticed that $b$ is almost half that for the complete case and hence $\theta_R$ is bigger. This implies that for the same sample size, a bigger treatment effect is needed for highly censored data to give the test a power similar to that of the uncensored. As already noted, the values for $b$, $\theta_R$ and $\lambda_E$ when $d = 10$ and $k = 5$ vary for each case. Therefore, a similar set of values is also needed to simulate data for $d =$ 1, 5 and $k = 10$ accordingly.

Under the null, the hazards are indeed proportional with a ratio equal to 1 (zero treatment advantage); hence it is easy to verify the type I error rate. However, under the alternative, the assumption of proportional hazards may no longer hold, as the common $\theta$ assumed may actually vary. Under $H_1$: $\theta = \theta_R$ when $\theta_R$ corresponds to $n =$ 1000. When $n$ is fixed, $\theta$ is adjusted to account for the ***within-subject*** correlation, to give the desired power. Therefore, the $\theta$ input to the program is actually the log hazard ratio ***within subject***, say $\theta_W$. This is a notional quantity as each subject receives only one treatment. Initially, $\theta$ was set with a view to fixing power ($1\text{-}\beta_{12} = 0.90$), but in fact the power is determined by $\theta_B$, the log hazard ratio ***between subjects***. This means that, the simulation output $\hat{\theta}$ indeed gives the estimated treatment effect associated with the log hazard ratio ***between subjects*** on $E$ and $C$.

Suppose $p_E$ and $p_C$ are the probabilities of failure of an event for patients on $E$ and $C$ respectively. It is to be recalled from Section 3.1.4, that the log hazard ratio is expressed in terms of survivor functions: $\theta = -\log\{-\log S_E(t)\} + \log\{-\log S_C(t)\}$. However, when there is a random subject effect, a problem arises. Substituting the survivor functions with these probabilities into the equation, we get

$\theta = -\log\{-\log(1-p_E)\} + \log\{-\log(1-p_C)\}$. Conditional on the subject effect $s_i$, a random effect model is expressed as $\log\{-\log(1-p)\} = \alpha + \theta z + s_i$ and $(1-p) = \mathrm{E}\left[\exp\{-\exp(\alpha + \theta z + s_i)\}\right]$. For simplicity, setting $\alpha$ = 0, and $z$ = 1 if the patient is on $E$, 0 otherwise, the treatment effect based on the log hazard ratio between subjects is given by

$$\theta_B = -\log\left\{-\log(\mathrm{E}\left[\exp\{-\exp(\theta + s_i)\}\right])\right\} + \log\left\{-\log(\mathrm{E}\left[\exp\{-\exp(s_i)\}\right])\right\}.$$

Comparing the expressions for $\theta$ and $\theta_B$ above, it is evident that $\theta_B < \theta$, and as the standard deviation of $s_i$ approaches zero: $\sigma \rightarrow 0$, $\theta_B \rightarrow \theta$. This means that the power is achieved only at low correlations. Therefore, it is anticipated that the program output will give $\hat{\theta}$ which is the estimate of $\theta_B$, itself smaller than the input $\theta$: low power at high correlation. A similar relationship between $\theta$ and $\theta_B$ for a proportional odds model of binary data has been reported by Bolland (2003), which also relates to methodology presented by Neuhaus et al. (1991).

In the simulation results (Section 5.2), the directly estimated power is hence compared with its corresponding theoretical power, computed at the estimated $\theta_B$, which is $\hat{\theta}$. In reality, the assumption of a common treatment effect under $H_1$ may be violated. The best we can do is to identify the proportional hazards model that is closest to the actual case we have. We generate a big data set (for example in Table 5.1), and choose the sample size when the required parameter converges to a constant value, as decribed earlier.

### 5.1.1. **Key Performance Measures**

To quantify the accuracy of the proposed method, as well as to justify the choice of intervals ($k$ = 5 or 10?), there are five specific key measures which can be employed. The two most prominent performance measures are the type I error rate and the power of the test respectively, under the null and alternative hypotheses. Additionally, the correlation ratio, the coverage probability and the bias of $\theta$ can also be taken into account. However, the coverage probability and the bias can only be determined if indeed the actual $\theta$ is known. As already explained in the Section 5.1, in the computer code generating the simulated data we are able to fix $\theta_W$ (***within subject***), while from the resulting survival times we estimate $\theta_B$ (***between subjects***). Therefore, only the three aspects of accuracy assessment (type I, power and correlation ratio) are investigated and reported in Section 5.2.

The first two measures have been described in Section 1.7, and therefore are discussed only briefly here. As described in the previous section, some power loss is anticipated with increased correlation. The power will instead be compared to a theoretical power denoted by TP, which appropriately compares the estimated $\hat{\theta}$ to the input $\theta$ adjusted based on the same log hazard ratio between subjects. As already explained in Section 2.3, the advantage of a global test lies in between the two extreme cases of completely independent endpoints and totally dependent endpoints (as if there was only a single endpoint).

As to the correlation ratio, an ideal situation arises when the estimated correlation using the proposed method, denoted by $\rho_{(est)}$, is exactly the same as the "true" correlation. It is to be recalled that each simulation is replicated 10,000 times ($N$ = 10,000) under each hypothesis. The estimate of the covariance between two score statistics, $C_{12}$, is calculated from each replicate simulation and similarly for the

correlation estimate $\hat{\rho}$. The average value of $\hat{\rho}$ from the 10,000 replicates, gives the best estimate, $\rho_{(est)}$. Since the "true" correlation is unknown, it is assumed that the correlation observed from its own samples of $N$, denoted by $\rho_{(sample)}$ gives the true correlation asymptotically. With $N$ replicate simulations of the same study, the sample covariance, $\text{cov}(Z_1,Z_2)$ can be obtained from the expression $\text{cov}(Z_1,Z_2) = (\sum Z_1 Z_2 - ((\sum Z_1 \sum Z_2)/N)/(N-1))$. The correlation derived from the sample covariance is given by $\rho_{(sample)} = \text{cov}(Z_1,Z_2)/\sqrt{\text{var}(Z_1)\text{var}Z_2}$. Therefore, the correlation ratio of both estimates, $\rho_{(est)}/\rho_{(sample)}$, will be compared in investigating the properties of the correlation estimator and in evaluating the accuracy of this method.

Although the coverage probability and the bias are not considered in the assessment of accuracy in this study, they are briefly described here and are later commented in Section 5.2.8. The coverage probability of a confidence interval is estimated by the proportion of replicates for which the interval contains the true value of interest (Dodge, 2003), in this case the treatment effect. Meanwhile, the bias of the estimate of $\theta$ is given by $\hat{\theta} - \theta$ and unbiasedness is indeed a desired property of a good estimator. In order to evaluate the bias and the coverage probability, it is necessary to know the true value of $\theta$, and as explained earlier, the simulation method used does not allow this, except for simulations under the null hypothesis where $\theta = 0$. The three key performance measures (type I error, power and correlation ratio) are reported in the results for individual case (Sections 5.2.1 to 5.2.7) and later summarized in the overall results (Section 5.2.8).

### 5.1.2. **Combined Hypothesis Tests**

In Section 2.1.1, global tests have been described in the context of binary data. Here, they are extended to interval-censored survival data with $\theta_m$ representing log hazard ratio for $m^{th}$ event. Based on the normality assumption of $Z$, already described in Sections 1.4 and 2.1.1, the following hypothesis tests can be employed. In this section, only the score test and Wald test are considered, whilst a good description on test statistics which includes the likelihood ratio test is available from Azzalini (1996). It is to be recalled from Section 4.4 that there are two ways to estimate the overall treatment effect. Consequently, two options of hypothesis tests are available and are described here.

For the standard estimate of overall treatment effect, $\hat{\theta}^*$, under the null, $\theta = 0$; hence $Z^{*2}/V^*$ can be tested against the chi-squared distribution with one degree of freedom, $\chi_1^2$. Equivalently, $Z^*/\sqrt{V^*}$ can be compared with the critical value for a standard normal distribution, $N(0, 1)$. This is an extension of the logrank test for univariate case as described in Section 1.6, now adapted for bivariate case. From Section 4.4, the standard error of $\hat{\theta}^*$ is given by $1/\sqrt{V}$, and hence an approximate 95% confidence interval for $\hat{\theta}^*$ is expressed as $(Z^*/V^* \pm 1.96/\sqrt{V^*})$.

Alternatively, and more appropriately, using the optimal weighting in equation (4.10), the standard error of the optimal estimate of the treatment advantage, $s.e.(\hat{\theta})$, is given by the square root of equation (4.11). Similar to its standard counterpart described earlier, $\hat{\theta} / s.e(\hat{\theta})$ can be tested against $N(0, 1)$ or $(\hat{\theta}^2 / \text{var}(\hat{\theta})$ against $\chi_1^2$. Therefore, the p-value is given by $1-\Phi(\hat{\theta} / s.e.(\hat{\theta}))$ and this expression is used to compute all the p-values for the remainder of this thesis, unless noted otherwise.

## 5.2. Simulation and Results

With reference to Step 4 in Figure 5.1, the fixed number of patients, $n$ = 1,000 are generated and randomized equally to control and experimental. In this investigation, only $\theta$ corresponding to $V^*$ is used since the main interest is the overall treatment effect. The effect of increasing the standard deviation of the subject effect, $\sigma$ is also investigated by varying $d$. Each data set is generated based on the variables set for each value of $d$ and $k$, on each hypothesis: an example was earlier displayed in Table 5.2.

The score statistics, Fisher's information and covariance for each interval are computed to yield the global score statistics, variances and covariance estimator, $C_{12}$ and hence the correlation estimator, $\hat{\rho}$. It is to be recalled that this part of the procedure was illustrated in Section 4.5.1 earlier. All simulation runs are replicated 10,000 times under each hypothesis, the average values from which are taken to be the best estimates. The results for each of the six cases are summarized in tables and figures.

### 5.2.1. Complete or Uncensored Data

Complete data sets with no censoring are simulated as the first check point since the simple setting does not involve any censoring rules. In Section 2.1.1, the theory states that $Z \sim N(\theta V, V)$ for large $V$ and small $\theta$: the variance of $Z$ is approximately $V$. In this study involving 10,000 replicate simulations of a sample size of 1,000 each, the average values of $V$ and variance of $Z$, obtained under each hypothesis, are summarized in Table 5.3.

Table 5.3: Average values of $V$ and var $Z$ for the complete case under the null and alternative hypotheses.

| Complete | | d | $V_1$ | Var $Z_1$ | $V_2$ | Var $Z_2$ | V* | Var Z* |
|---|---|---|---|---|---|---|---|---|
| k = 5 | $H_0$ | 1 | 196.8 | 196.2 | 196.8 | 190.6 | 385.3 | 379.2 |
| | | 5 | 196.9 | 195.2 | 196.9 | 195.3 | 285.0 | 288.8 |
| | | 10 | 196.8 | 197.4 | 196.8 | 195.0 | 236.6 | 240.3 |
| | $H_1$ | 1 | 196.5 | 193.7 | 196.6 | 189.3 | 382.7 | 368.8 |
| | | 5 | 196.5 | 192.2 | 196.6 | 193.7 | 284.2 | 284.0 |
| | | 10 | 196.6 | 195.3 | 196.7 | 194.7 | 236.4 | 238.7 |
| k = 10 | $H_0$ | 1 | 223.2 | 225.3 | 223.2 | 217.9 | 436.8 | 434.4 |
| | | 5 | 223.2 | 224.7 | 223.1 | 222.0 | 326.0 | 337.3 |
| | | 10 | 223.2 | 224.8 | 223.2 | 223.6 | 271.4 | 284.4 |
| | $H_1$ | 1 | 222.3 | 220.1 | 222.2 | 214.9 | 432.6 | 417.5 |
| | | 5 | 222.5 | 221.9 | 222.5 | 218.1 | 324.8 | 331.4 |
| | | 10 | 222.9 | 224.1 | 222.9 | 222.3 | 271.0 | 282.9 |

It is clearly apparent in the table above that the variance of $Z$ is very close to information $V$ in each setting under both hypotheses. For example, the first row shows that $V_1$ (196.8) and var $Z_1$ (196.2). It is worth noting that a similar observation is found in all the cases simulated, where var $Z \approx V$ holds. Also, as expected, an increase in the number of intervals gives a bigger $V$, as more information about $\theta$ is contained in $Z$. These values are listed in this section only for illustration purposes, as they are not of prime interest in this study.

A standard table of results summarizing the three key measures is presented for each case studied under the null ($H_0$) and alternative ($H_1$), with varying degrees of correlation: $d$ = 1 (low), 5 (medium), 10 (high) and number of intervals ($k$ = 5, 10). In such a table as that below (Table 5.4), it is to be noted that the first two columns contain the values set for $d$ and $\theta$, while others relate to the estimates obtained. To

assess the validity of the assumption of a common treatment advantage, $\theta_1 = \theta_2 = \theta$, the estimates are tabulated alongside their corresponding type I errors and powers. The type I error rates, corresponding to the marginal tests based on $T_1$, $T_2$ and the global test based on both, are $\alpha_1$, $\alpha_2$ and $\alpha_{12}$ respectively. Similarly, their alternative counterparts are denoted by 1-$\beta_1$, 1-$\beta_2$ and 1-$\beta_{12}$. The theoretical power for $\hat{\theta}$ (Section 5.1) denoted by TP, serves as a baseline for comparison with the global power 1-$\beta_{12}$. Meanwhile, $\rho_{(est)}$ and $\rho_{(sample)}$ respectively are the correlation estimated using the proposed method, and the correlation derived from the samples. It is to be noted that in the tables of results for the remainder of this thesis, texts in bold highlight particular values which are worthy of remark.

Table 5.4: Summary of results for the complete case under the null and alternative, using 5 and 10 intervals.

| $H_0$ @ 5 intervals | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $d$ | $\theta$ | $\hat{\theta}_1$ | $\hat{\theta}_2$ | $\hat{\theta}$ | $\alpha_1$ | $\alpha_2$ | $\alpha_{12}$ | N/A | $\rho_{(est)}$ | $\rho_{(sample)}$ |
| 1 | 0.000 | 0.000 | 0.000 | 0.000 | 0.023 | 0.021 | 0.023 | N/A | 0.023 | 0.026 |
| 5 | 0.000 | -0.001 | -0.001 | -0.001 | 0.024 | 0.025 | 0.026 | N/A | 0.382 | 0.412 |
| 10 | 0.000 | -0.001 | -0.001 | -0.001 | 0.024 | 0.027 | 0.025 | N/A | 0.664 | 0.693 |
| $H_1$ @ 5 intervals | | | | | | | | | | |
| $d$ | $\theta$ | $\hat{\theta}_1$ | $\hat{\theta}_2$ | $\hat{\theta}$ | 1-$\beta_1$ | 1-$\beta_2$ | 1-$\beta_{12}$ | TP | $\rho_{(est)}$ | $\rho_{(sample)}$ |
| 1 | 0.174 | 0.171 | 0.171 | 0.171 | 0.67 | 0.67 | 0.92 | 0.89 | 0.028 | 0.027 |
| 5 | 0.209 | 0.139 | 0.139 | 0.139 | 0.50 | 0.49 | 0.65 | 0.58 | 0.384 | 0.412 |
| 10 | 0.221 | 0.089 | 0.088 | 0.088 | 0.24 | 0.23 | 0.28 | 0.25 | 0.664 | 0.694 |
| $H_0$ @ 10 intervals | | | | | | | | | | |
| $d$ | $\theta$ | $\hat{\theta}_1$ | $\hat{\theta}_2$ | $\hat{\theta}$ | $\alpha_1$ | $\alpha_2$ | $\alpha_{12}$ | N/A | $\rho_{(est)}$ | $\rho_{(sample)}$ |
| 1 | 0.000 | 0.000 | 0.000 | 0.000 | 0.024 | 0.022 | 0.024 | N/A | 0.023 | 0.026 |
| 5 | 0.000 | -0.001 | -0.001 | -0.001 | 0.024 | 0.026 | **0.028** | N/A | 0.370 | 0.415 |
| 10 | 0.000 | -0.001 | -0.001 | -0.001 | 0.025 | 0.026 | 0.027 | N/A | 0.645 | 0.715 |
| $H_1$ @ 10 intervals | | | | | | | | | | |
| $d$ | $\theta$ | $\hat{\theta}_1$ | $\hat{\theta}_2$ | $\hat{\theta}$ | 1-$\beta_1$ | 1-$\beta_2$ | 1-$\beta_{12}$ | TP | $\rho_{(est)}$ | $\rho_{(sample)}$ |
| 1 | 0.165 | 0.162 | 0.162 | 0.162 | 0.68 | 0.68 | 0.92 | 0.89 | 0.028 | 0.025 |
| 5 | 0.197 | 0.130 | 0.130 | 0.130 | 0.50 | 0.49 | 0.65 | 0.57 | 0.371 | 0.416 |
| 10 | 0.208 | 0.083 | 0.083 | 0.083 | 0.24 | 0.23 | 0.28 | 0.25 | 0.645 | 0.713 |

In each setting for the complete data, the type I error is well within the 95% probability interval (0.022, 0.028), the global power exceeds its theoretical power and the estimated correlation $\rho_{(est)}$ seems close to its sampled correlation $\rho_{(sample)}$. However, a slightly inflated type I error is observed at 10 intervals ($\alpha_{12}$ = 0.028). The estimates of the marginal and global $\theta$s are exactly of the same value in each setting under both hypotheses: a valid assumption of $\hat{\theta}_1 = \hat{\theta}_2 = \hat{\theta}$. The marginal powers $1-\beta_1$ and $1-\beta_2$, based on individual $T_1$ and $T_2$ respectively, are obviously smaller than the global power: similar findings which demonstrate the advantage of global test as seen in Chapter 2 earlier.

In Table 5.4 above, it is evident that upon increasing the correlation multiplier $d$, the required $\theta$ gets larger, but as anticipated the power is gradually lost with increasing $d$. As explained in Section 5.1, the estimate, $\theta_B$ is expected to be smaller than the value set $\theta = \theta_W$, as correlation increases. For example, when $d$ = 10 (5 intervals), $\hat{\theta}$ is only 0.088 compared to the setting of $\theta$ = 0.221, which seems to be about 60% underestimation. Consequently, the power of the test for the combined score, $1-\beta_{12}$ is only "accurate" compared to the 95% probability interval (0.89, 0.91) at low correlation ($d$ = 1) at both interval settings. Upon comparing the results for 5 and 10 intervals, the former seems to give better results in terms of type I error and correlation ratio ( $d$ = 10), while the power is exactly the same as that for the latter.

### 5.2.2. **Paired Organs**

To simulate the case of paired organs, $T_1$ and $T_2$ are generated in a similar way as to the complete case, except that a 50% censoring proportion is now applied to the data, as shown in Table 5.2 earlier. Only one censoring variable exists as $T_1$ and $T_2$ concern the same patient. The data are then subjected to the specific censoring rules, as already described in Section 4.3.2. The results are summarized in Table 5.5.

Table 5.5: Summary of results for the paired case under the null and alternative.

| $H_0$ @ 5 intervals | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $d$ | $\theta$ | $\hat{\theta}_1$ | $\hat{\theta}_2$ | $\hat{\theta}$ | $\alpha_1$ | $\alpha_2$ | $\alpha_{12}$ | N/A | $\rho_{(est)}$ | $\rho_{(sample)}$ |
| 1 | 0.000 | 0.000 | 0.000 | 0.000 | 0.025 | 0.023 | 0.026 | N/A | 0.035 | 0.027 |
| 5 | 0.000 | 0.001 | 0.000 | 0.000 | **0.028** | 0.025 | **0.028** | N/A | 0.421 | 0.437 |
| 10 | 0.000 | 0.001 | 0.002 | 0.001 | 0.027 | 0.025 | 0.026 | N/A | 0.674 | 0.685 |
| $H_1$ @ 5 intervals | | | | | | | | | | |
| $d$ | $\theta$ | $\hat{\theta}_1$ | $\hat{\theta}_2$ | $\hat{\theta}$ | $1\text{-}\beta_1$ | $1\text{-}\beta_2$ | $1\text{-}\beta_{12}$ | TP | $\rho_{(est)}$ | $\rho_{(sample)}$ |
| 1 | 0.241 | 0.235 | 0.236 | 0.235 | 0.65 | 0.65 | 0.90 | 0.89 | 0.041 | 0.022 |
| 5 | 0.287 | 0.190 | 0.190 | 0.190 | 0.47 | 0.48 | 0.61 | 0.57 | 0.422 | 0.437 |
| 10 | 0.305 | 0.120 | 0.120 | 0.120 | 0.23 | 0.22 | 0.25 | 0.25 | 0.692 | 0.704 |
| $H_0$ @ 10 intervals | | | | | | | | | | |
| $d$ | $\theta$ | $\hat{\theta}_1$ | $\hat{\theta}_2$ | $\hat{\theta}$ | $\alpha_1$ | $\alpha_2$ | $\alpha_{12}$ | N/A | $\rho_{(est)}$ | $\rho_{(sample)}$ |
| 1 | 0.000 | 0.000 | 0.001 | 0.001 | 0.025 | 0.025 | **0.028** | N/A | 0.031 | 0.026 |
| 5 | 0.000 | 0.000 | 0.001 | 0.001 | **0.028** | 0.024 | 0.025 | N/A | 0.412 | 0.428 |
| 10 | 0.000 | 0.001 | 0.002 | 0.002 | 0.026 | 0.025 | 0.026 | N/A | 0.679 | 0.706 |
| $H_1$ @ 10 intervals | | | | | | | | | | |
| $d$ | $\theta$ | $\hat{\theta}_1$ | $\hat{\theta}_2$ | $\hat{\theta}$ | $1\text{-}\beta_1$ | $1\text{-}\beta_2$ | $1\text{-}\beta_{12}$ | TP | $\rho_{(est)}$ | $\rho_{(sample)}$ |
| 1 | 0.228 | 0.222 | 0.224 | 0.223 | 0.65 | 0.66 | 0.91 | 0.89 | 0.037 | 0.023 |
| 5 | 0.271 | 0.180 | 0.181 | 0.180 | 0.48 | 0.48 | 0.62 | 0.58 | 0.413 | 0.429 |
| 10 | 0.288 | 0.115 | 0.116 | 0.115 | 0.22 | 0.23 | 0.26 | 0.25 | 0.679 | 0.707 |

For the paired data, slightly inflated type I error rates are observed, but they are still within the 95% PI (0.022, 0.028). The global power achieved is higher than the theoretical power and the correlation estimates $\rho_{(est)}$ and $\rho_{(sample)}$ are consistent for each scenario above. The global test advantage is clearly evident whereby the global power is much larger than that of the marginals. This is anticipated since the

assumption of equal treatment effect is satisfactorily met, as shown in the above table. Additionally, it seems that the 10 intervals setting yields a slightly higher power, by 1%, but a similar type I error compared to the 5 intervals.

### 5.2.3. **Related Indicators**

As mentioned in Section 5.1, the generally related indicators are simulated with mild censoring (20%). Unlike the paired case earlier, each $T_1$ and $T_2$ can be censored independently as described in Section 4.3.3 and the censoring rules illustrated in Section 4.5.2 are applied. The results for both interval settings are given in Table 5.6.

Table 5.6: Results for the indicators case under the null and alternative.

| $H_0$ @ 5 intervals | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $d$ | $\theta$ | $\hat{\theta}_1$ | $\hat{\theta}_2$ | $\hat{\theta}$ | $\alpha_1$ | $\alpha_2$ | $\alpha_{12}$ | N/A | $\rho_{(est)}$ | $\rho_{(sample)}$ |
| 1 | 0.000 | 0.000 | -0.001 | -0.001 | 0.022 | 0.023 | 0.023 | N/A | 0.023 | 0.032 |
| 5 | 0.000 | -0.001 | -0.001 | -0.001 | 0.022 | 0.026 | 0.027 | N/A | 0.387 | 0.404 |
| 10 | 0.000 | -0.001 | -0.001 | -0.001 | 0.025 | 0.025 | 0.026 | N/A | 0.670 | 0.688 |
| $H_1$ @ 5 intervals | | | | | | | | | | |
| $d$ | $\theta$ | $\hat{\theta}_1$ | $\hat{\theta}_2$ | $\hat{\theta}$ | $1-\beta_1$ | $1-\beta_2$ | $1-\beta_{12}$ | TP | $\rho_{(est)}$ | $\rho_{(sample)}$ |
| 1 | 0.194 | 0.189 | 0.189 | 0.189 | 0.66 | 0.66 | 0.91 | 0.89 | 0.028 | 0.034 |
| 5 | 0.245 | 0.161 | 0.161 | 0.161 | 0.50 | 0.49 | 0.65 | 0.57 | 0.389 | 0.406 |
| 10 | 0.275 | 0.107 | 0.107 | 0.107 | 0.24 | 0.23 | 0.26 | 0.24 | 0.671 | 0.685 |
| $H_0$ @ 10 intervals | | | | | | | | | | |
| $d$ | $\theta$ | $\hat{\theta}_1$ | $\hat{\theta}_2$ | $\hat{\theta}$ | $\alpha_1$ | $\alpha_2$ | $\alpha_{12}$ | N/A | $\rho_{(est)}$ | $\rho_{(sample)}$ |
| 1 | 0.000 | 0.000 | -0.001 | -0.001 | 0.021 | 0.023 | 0.022 | N/A | 0.023 | 0.033 |
| 5 | 0.000 | -0.001 | -0.001 | -0.001 | 0.024 | 0.027 | **0.028** | N/A | 0.374 | 0.404 |
| 10 | 0.000 | -0.001 | -0.001 | -0.001 | 0.025 | 0.025 | 0.027 | N/A | 0.649 | 0.686 |
| $H_1$ @ 10 intervals | | | | | | | | | | |
| $d$ | $\theta$ | $\hat{\theta}_1$ | $\hat{\theta}_2$ | $\hat{\theta}$ | $1-\beta_1$ | $1-\beta_2$ | $1-\beta_{12}$ | TP | $\rho_{(est)}$ | $\rho_{(sample)}$ |
| 1 | 0.218 | 0.180 | 0.179 | 0.179 | 0.67 | 0.66 | 0.92 | 0.89 | 0.028 | 0.032 |
| 5 | 0.2[illegible] | 0.151 | 0.151 | 0.151 | 0.49 | 0.50 | 0.66 | 0.57 | 0.375 | 0.402 |
| 10 | 0.258 | 0.101 | 0.101 | 0.101 | 0.23 | 0.23 | 0.28 | 0.25 | 0.649 | 0.686 |

For the related indicators, the type I error rates are within the 95% PI (0.022, 0.028) for all settings with 5 intervals showing a slight advantage over 10 intervals, with regard to the global $\alpha_{12}$. As with the earlier cases, the global power exceeds its theoretical power in each scenario. The setting of 10 intervals yields higher power (1% to 2%) compared to that of 5 intervals. However, the coarser intervals perform better than the finer in terms of the correlation estimates under both hypotheses.

### 5.2.4. **Progression-Free Survival**

A progression-free survival analysis dictates more events on $T_1$ (Section 4.3.4) compared to the earlier case of related indicators. Using the same initial data as generated in the latter case, adjustment is then made according to the censoring rules described in Section 4.5.2 earlier. The simulation study for PFS yields the following output as shown in Table 5.7.

Table 5.7: Summary of results for the PFS case under the null and alternative.

| $H_0$ @ 5 intervals | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $d$ | $\theta$ | $\hat{\theta}_1$ | $\hat{\theta}_2$ | $\hat{\theta}$ | $\alpha_1$ | $\alpha_2$ | $\alpha_{12}$ | N/A | $\rho_{(est)}$ | $\rho_{(sample)}$ |
| 1 | 0.000 | 0.000 | -0.001 | 0.000 | 0.025 | 0.023 | 0.024 | N/A | 0.589 | 0.601 |
| 5 | 0.000 | 0.000 | 0.000 | 0.000 | 0.026 | 0.025 | 0.026 | N/A | 0.755 | 0.779 |
| 10 | 0.000 | -0.001 | -0.001 | -0.001 | 0.024 | 0.023 | 0.025 | N/A | 0.869 | 0.882 |
| $H_1$ @ 5 intervals | | | | | | | | | | |
| $d$ | $\theta$ | $\hat{\theta}_1$ | $\hat{\theta}_2$ | $\hat{\theta}$ | $1-\beta_1$ | $1-\beta_2$ | $1-\beta_{12}$ | TP | $\rho_{(est)}$ | $\rho_{(sample)}$ |
| 1 | 0.223 | 0.214 | 0.225 | 0.218 | 0.83 | 0.81 | 0.90 | 0.89 | 0.592 | 0.601 |
| 5 | 0.252 | 0.156 | 0.167 | 0.159 | 0.55 | 0.52 | 0.59 | 0.53 | 0.755 | 0.779 |
| 10 | 0.275 | 0.102 | 0.108 | 0.103 | 0.24 | 0.23 | 0.25 | 0.23 | 0.869 | 0.882 |
| $H_0$ @ 10 intervals | | | | | | | | | | |
| $d$ | $\theta$ | $\hat{\theta}_1$ | $\hat{\theta}_2$ | $\hat{\theta}$ | $\alpha_1$ | $\alpha_2$ | $\alpha_{12}$ | N/A | $\rho_{(est)}$ | $\rho_{(sample)}$ |
| 1 | 0.000 | 0.000 | -0.001 | 0.000 | 0.026 | 0.021 | 0.024 | N/A | 0.608 | 0.629 |
| 5 | 0.000 | 0.000 | 0.000 | 0.000 | 0.026 | 0.025 | 0.026 | N/A | 0.734 | 0.774 |
| 10 | 0.000 | -0.001 | 0.000 | -0.001 | 0.024 | 0.026 | 0.025 | N/A | 0.848 | 0.876 |
| $H_1$ @ 10 intervals | | | | | | | | | | |
| $d$ | $\theta$ | $\hat{\theta}_1$ | $\hat{\theta}_2$ | $\hat{\theta}$ | $1-\beta_1$ | $1-\beta_2$ | $1-\beta_{12}$ | TP | $\rho_{(est)}$ | $\rho_{(sample)}$ |
| 1 | 0.212 | 0.201 | 0.214 | 0.206 | 0.84 | 0.81 | 0.90 | 0.88 | 0.610 | 0.628 |
| 5 | 0.238 | 0.146 | 0.157 | 0.149 | 0.55 | 0.53 | 0.59 | 0.53 | 0.757 | 0.798 |
| 10 | 0.259 | 0.096 | 0.102 | 0.097 | 0.25 | 0.24 | 0.26 | 0.23 | 0.861 | 0.891 |

As in the earlier cases, the type I error rate is within the 95% PI, and the global power exceeds its theoretical power (TP) in each setting. The correlation estimates are consistent, although slightly higher values are observed under the alternative, at 10 intervals ($d$ = 5, 10), compared to those under the null. It is notable that the estimates of marginal and global treatment effects are no longer equal, unlike the earlier cases. The impact of the invalid assumption of $\theta$ equality is reflected in the power of test: the marginal power is now closer to the global, compared to all the earlier cases. Upon comparing the performances of the 5 and 10 intervals, the former gives better results for the type I error and correlation estimate, while the latter is slightly more advantageous in terms of power.

### 5.2.5. **Recurrent Events (Total Time)**

For the recurrent events total time model, time to the second recurrence, $T_2$ contains time to the first recurrence, $T_1$ and therefore the data generation differs from the other cases. Two time variables, say $T_1$ and $T_{2G}$, are generated as described in Section 5.1, and their sum makes up $T_2$. An alternative method is also explored, whereby the smaller value of the two time variables is taken as $T_1$ and the larger as $T_2$: their results are displayed under the label "2nd option". The simulated data are then similarly subjected to the censoring rules relating to the features described in Section 4.6.1. For example, when $T_1$ is censored at $t_1 = c_1$, $T_2$ is also censored at the same time $c_1$. As already acknowledged, the recurrent TT case is expected to be highly correlated since the magnitude of $T_2$ includes $T_1$ as well. The simulation results are displayed in Table 5.8 and the results for the 2nd option are commented last.

Table 5.8: Results for the recurrent events TT case under the null and alternative.

| $H_0$ @ 5 intervals | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $d$ | $\theta$ | $\hat{\theta}_1$ | $\hat{\theta}_2$ | $\hat{\theta}$ | $\alpha_1$ | $\alpha_2$ | $\alpha_{12}$ | N/A | $\rho_{(est)}$ | $\rho_{(sample)}$ |
| 1 | 0.000 | 0.000 | 0.001 | 0.000 | 0.026 | 0.025 | 0.027 | N/A | 0.539 | 0.557 |
| 5 | 0.000 | 0.001 | 0.001 | 0.001 | **0.028** | 0.025 | **0.028** | N/A | 0.730 | 0.751 |
| 10 | 0.000 | 0.001 | 0.001 | 0.001 | 0.026 | 0.025 | 0.026 | N/A | 0.848 | 0.855 |
| $H_1$ @ 5 intervals | | | | | | | | | | |
| $d$ | $\theta$ | $\hat{\theta}_1$ | $\hat{\theta}_2$ | $\hat{\theta}$ | 1-$\beta_1$ | 1-$\beta_2$ | 1-$\beta_{12}$ | TP | $\rho_{(est)}$ | $\rho_{(sample)}$ |
| 1 | 0.293 | 0.282 | **0.430** | 0.317 | 0.86 | **0.96** | 0.95 | 0.94 | 0.543 | 0.555 |
| 5 | 0.311 | 0.192 | 0.238 | 0.200 | 0.53 | 0.57 | 0.58 | 0.55 | 0.730 | 0.749 |
| 10 | 0.317 | 0.118 | 0.131 | 0.120 | 0.24 | 0.24 | 0.24 | 0.23 | 0.848 | 0.856 |
| $H_0$ @ 10 intervals | | | | | | | | | | |
| $d$ | $\theta$ | $\hat{\theta}_1$ | $\hat{\theta}_2$ | $\hat{\theta}$ | $\alpha_1$ | $\alpha_2$ | $\alpha_{12}$ | N/A | $\rho_{(est)}$ | $\rho_{(sample)}$ |
| 1 | 0.000 | -0.001 | 0.001 | 0.000 | 0.026 | 0.025 | **0.028** | N/A | 0.551 | 0.588 |
| 5 | 0.000 | 0.000 | 0.000 | 0.000 | **0.028** | 0.025 | 0.027 | N/A | 0.723 | 0.751 |
| 10 | 0.000 | 0.001 | 0.001 | 0.001 | 0.026 | 0.023 | 0.026 | N/A | 0.839 | 0.855 |
| $H_1$ @ 10 intervals | | | | | | | | | | |
| $d$ | $\theta$ | $\hat{\theta}_1$ | $\hat{\theta}_2$ | $\hat{\theta}$ | 1-$\beta_1$ | 1-$\beta_2$ | 1-$\beta_{12}$ | TP | $\rho_{(est)}$ | $\rho_{(sample)}$ |
| 1 | 0.278 | 0.267 | 0.404 | 0.299 | 0.86 | 0.95 | 0.95 | 0.94 | 0.553 | 0.576 |
| 5 | 0.293 | 0.181 | 0.226 | 0.189 | 0.53 | 0.57 | 0.58 | 0.55 | 0.723 | 0.750 |
| 10 | 0.299 | 0.113 | 0.127 | 0.115 | 0.24 | 0.24 | 0.25 | 0.24 | 0.839 | 0.855 |
| 2nd option @ $H_1$ (5 intervals) | | | | | | | | | | |
| $d$ | $\theta$ | $\hat{\theta}_1$ | $\hat{\theta}_2$ | $\hat{\theta}$ | 1-$\beta_1$ | 1-$\beta_2$ | 1-$\beta_{12}$ | TP | $\rho_{(est)}$ | $\rho_{(sample)}$ |
| 1 | [illegible] | 0.247 | 0.360 | 0.280 | 0.85 | 0.93 | 0.96 | 0.95 | 0.414 | 0.418 |
| 5 | [illegible] | 0.177 | 0.223 | 0.187 | 0.54 | 0.57 | 0.60 | 0.57 | 0.655 | 0.669 |
| 10 | 0.301 | 0.113 | 0.128 | 0.115 | 0.24 | 0.23 | 0.25 | 0.23 | 0.803 | 0.812 |

The intended type I error rates are achieved, but they are close to the high limit for both interval settings. The global power exceeds the theoretical power and the values of estimated correlation, $\rho_{(est)}$ are close to their corresponding $\rho_{(sample)}$, in all settings. While the marginal powers for individual $T_1$ and $T_2$ were similar in the earlier cases, the recurrent case reveals a slight increase in power for the latter. In fact, for $d=1$ (5 intervals), a marginal power 1-$\beta_2$ of 0.96 exceeds the global power (0.95). It is to be noticed that the assumption of $\theta$ equality is violated, with the marginal estimate $\hat{\theta}_2$ much larger than the global estimate, $\hat{\theta}$. This accounts for the lack of

global power in this case. However, for the 2nd option the estimates of treatment effect are now closer to each other and consequently, the global power is always higher than the marginal power. This implies that the 2nd option yields recurrent TT data that is closer to the assumption of $\theta$ equality and hence resulting in an advantage of the global test methodology. Overall, for the recurrent TT case, the use of 10 intervals seems to yield improved performance in terms of type I error and power, but slightly lower correlation ratio compared to those for 5 intervals.

### 5.2.6. **Recurrent Events (Gap Time)**

To generate data for the recurrent events gap time model, $T_1$ and $T_{2G}$ are subjected to a common censoring variable, similar to the recurrent TT. However, in gap time convention, $T_2 = T_{2G}$, and when $T_1$ is censored, $T_2 = 0$ (gap time). In other words, when $T_1$ is censored, $T_2$ is completely missing. Since the magnitude of $T_1$ is no longer contained within $T_{2G}$ for GT, the outcomes for each of the paired intervals are quite different from those when using TT. Thus, it is only logical that the correlation between the events is much lower in GT compared to that for TT, as already observed with the bladder cancer data in Section 4.6.2. The simulation results for recurrent GT are summarized in Table 5.9.

Table 5.9: Results for the recurrent events GT case under the null and alternative.

| $H_0$ @ 5 intervals | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $d$ | $\theta$ | $\hat{\theta}_1$ | $\hat{\theta}_2$ | $\hat{\theta}$ | $\alpha_1$ | $\alpha_2$ | $\alpha_{12}$ | N/A | $\rho_{(est)}$ | $\rho_{(sample)}$ |
| 1 | 0.000 | -0.001 | 0.000 | 0.000 | 0.025 | 0.025 | 0.026 | N/A | 0.040 | 0.030 |
| 5 | 0.000 | 0.001 | 0.000 | 0.000 | 0.026 | 0.022 | 0.022 | N/A | **0.305** | 0.235 |
| 10 | 0.000 | 0.001 | 0.000 | 0.001 | 0.026 | 0.027 | **0.020** | N/A | **0.512** | 0.352 |
| $H_1$ @ 5 intervals | | | | | | | | | | |
| $d$ | $\theta$ | $\hat{\theta}_1$ | $\hat{\theta}_2$ | $\hat{\theta}$ | $1-\beta_1$ | $1-\beta_2$ | $1-\beta_{12}$ | TP | $\rho_{(est)}$ | $\rho_{(sample)}$ |
| 1 | 0.235 | 0.229 | 0.228 | 0.228 | 0.70 | 0.58 | 0.90 | 0.88 | 0.045 | 0.030 |
| 5 | 0.271 | 0.180 | 0.137 | 0.162 | 0.49 | 0.25 | 0.53 | 0.49 | **0.306** | 0.238 |
| 10 | 0.289 | **0.116** | 0.069 | 0.096 | **0.23** | 0.10 | 0.19 | 0.19 | **0.511** | 0.355 |
| $H_0$ @ 10 intervals | | | | | | | | | | |
| $d$ | $\theta$ | $\hat{\theta}_1$ | $\hat{\theta}_2$ | $\hat{\theta}$ | $\alpha_1$ | $\alpha_2$ | $\alpha_{12}$ | N/A | $\rho_{(est)}$ | $\rho_{(sample)}$ |
| 1 | 0.000 | -0.001 | 0.000 | 0.000 | 0.025 | 0.025 | 0.023 | N/A | 0.034 | 0.031 |
| 5 | 0.000 | 0.000 | 0.000 | 0.000 | 0.026 | 0.023 | **0.021** | N/A | **0.298** | 0.194 |
| 10 | 0.000 | 0.001 | 0.001 | 0.001 | 0.026 | 0.027 | **0.017** | N/A | **0.504** | 0.287 |
| $H_1$ @ 10 intervals | | | | | | | | | | |
| $d$ | $\theta$ | $\hat{\theta}_1$ | $\hat{\theta}_2$ | $\hat{\theta}$ | $1-\beta_1$ | $1-\beta_2$ | $1-\beta_{12}$ | TP | $\rho_{(est)}$ | $\rho_{(sample)}$ |
| 1 | 0.22 | 0.216 | 0.215 | 0.215 | 0.70 | 0.58 | 0.90 | 0.88 | 0.039 | 0.031 |
| 5 | 0.2 | 0.170 | 0.131 | 0.154 | 0.49 | 0.26 | 0.54 | 0.49 | **0.298** | 0.191 |
| 10 | 0.272 | 0.110 | 0.066 | 0.092 | **0.23** | 0.10 | **0.18** | 0.19 | **0.504** | 0.291 |

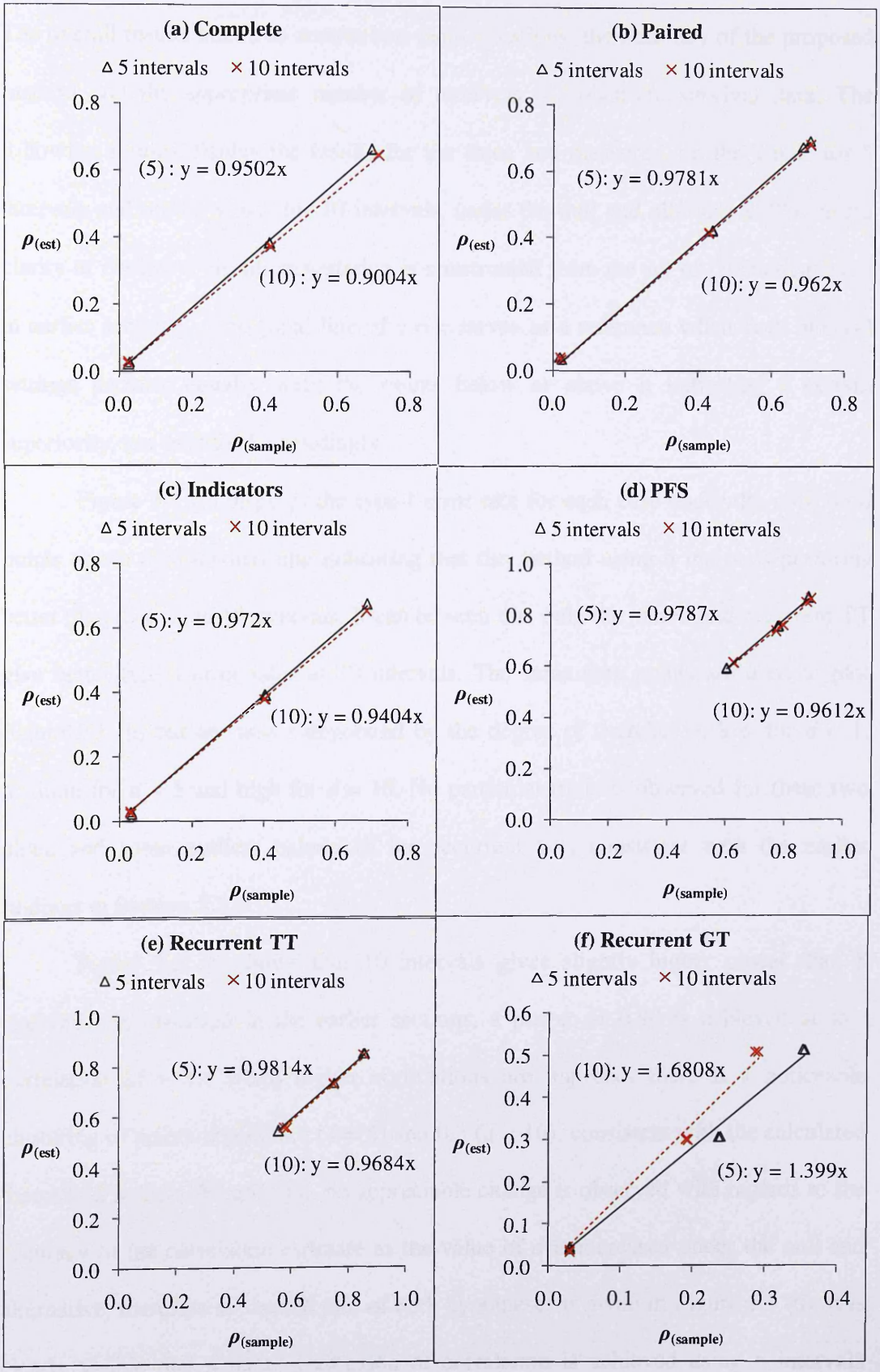
At high correlations ($d$ = 10), the global type I error rates are rather conservative and the global power is just equal to, or slightly lower than its theoretical power. The marginal type I error rates are within the 95% PI, and unlike the other cases, the marginal power $1-\beta_1$ is appreciably higher than $1-\beta_2$ and even exceeded the global power at $d$ = 10. Such trend is unsurprising for a recurrent events GT model. It is apparent that the estimated correlation, $\rho_{(est)}$, is substantially larger than $\rho_{(sample)}$ at higher correlations ($d$ = 5, 10). This overestimation suggests that the proposed method may not be suitable for the recurrent GT case when the correlation is high.

### 5.2.7. **Summary of Correlation Ratios**

A simple linear regression method is adopted in assessing the accuracy of the correlation estimate. The comparison between $\rho_{(est)}$ and $\rho_{(sample)}$ is given in a scatter plot with a trend line passing through the origin (0, 0) and the gradient from the linear equation gives the correlation ratio of $\rho_{(est)}/\rho_{(sample)}$: $y = x$ being the perfect estimation. The plot for each case under both hypotheses is given individually where (5) and (10) in front of the equation indicates the number of intervals.

As shown in Figure 5.2 (a), the linear fit equations for the complete case show gradients of 0.95 (5 intervals) and 0.90 (10 intervals): a 5% advantage for the coarser intervals. Similar trends are evident for all the other four cases shown in Figures 5.2 (b) to (e), with the advantage of 5 intervals ranging from 1% to 3% for correlation ratios of about 0.97 and 0.98. The first five charts show that the proposed method yields better correlation estimate using 5 intervals compared to 10 intervals. However, the recurrent GT shows some discrepancies in Figure 5.2 (f). There is a huge overestimation and the use of 10 intervals gives a higher correlation ratio (1.68) than does the 5 intervals option (1.40).

Figure 5.2: The correlation ratio for each case at 5 and 10 intervals.



### 5.2.8. **Overall Results: Five vs Ten Intervals**

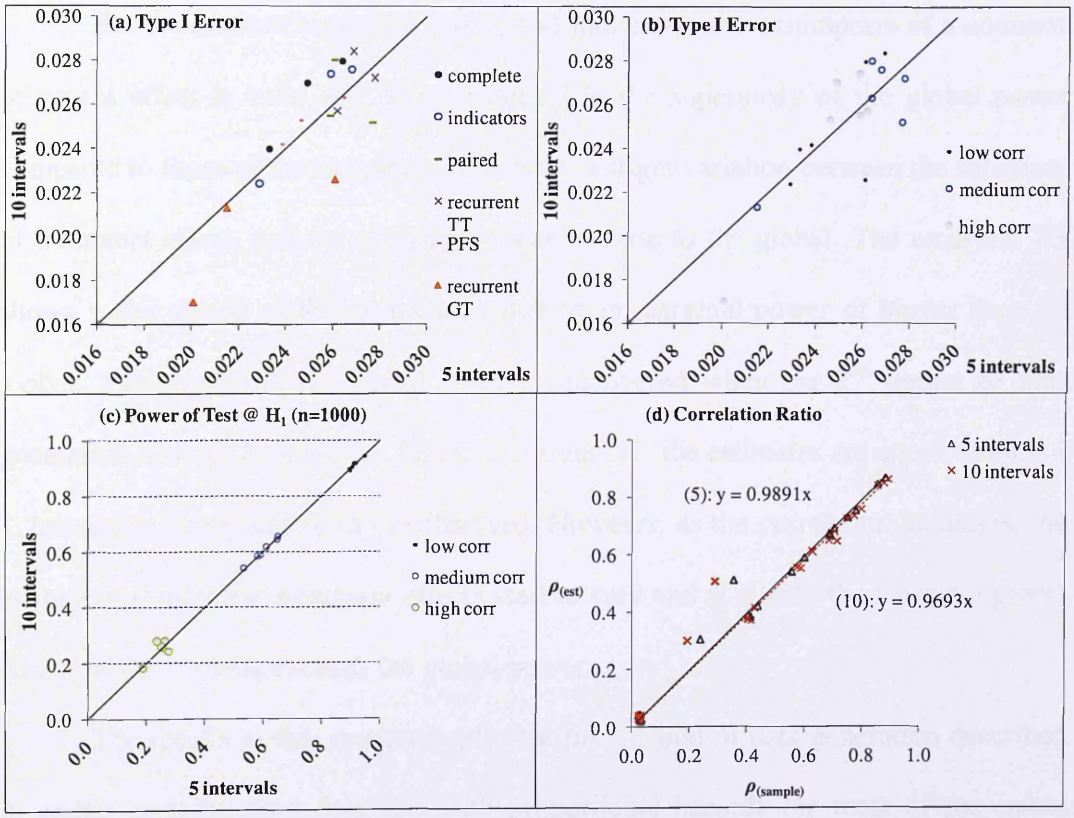
The overall results intend to answer two main questions: the accuracy of the proposed method and the appropriate number of intervals for bivariate survival data. The following figures display the results for the three key measures, on the $x$-axis for 5 intervals and on the $y$-axis for 10 intervals, under the null and alternative. To ensure clarity of choice, a visual presentation is constructed from the set of tables displayed in earlier sections. A diagonal line of $y = x$ serves as a reference when both interval settings perform equally well; the points below or above it indicating a certain superiority, are described accordingly.

Figure 5.3 (a) displays the type I error rate for each case under the null, with points above the diagonal line indicating that the method using 5 intervals performs better than that using 10 intervals. It can be seen that only the paired and recurrent TT give better type I error rates at 10 intervals. The same data points are used to plot Figure 5.3 (b) but are now categorized by the degree of correlation: low for $d = 1$, medium for $d = 5$ and high for $d = 10$. No particular trend is observed for these two plots, and some outliers belong to the recurrent GT, consistent with the earlier findings in Section 5.2.6.

Figure 5.3 (c) shows that 10 intervals gives slightly higher power than 5 intervals. As observed in the earlier sections, a power of 0.90 is achieved at low correlation ($d = 1$). When higher correlations are imposed, there is a noticeable clustering of points around 0.6 ($d = 5$) and 0.3 (d = 10), consistent with the calculated theoretical powers. Meanwhile, no appreciable change is observed with regards to the accuracy of the correlation estimate as the value of $d$ is increased under the null and alternative, therefore an overall plot of both hypotheses is given in Figure 5.3 (d). It is clearly visible that a better estimation of correlation is achieved using 5 intervals

compared to 10 intervals, with correlation ratios of 0.99 and 0.97 respectively. It is to be noted that the outliers above the lines are due to the recurrent GT, as already commented in the previous section. Excluding these points, the corresponding ratios for 5 and 10 intervals respectively are 0.98 and 0.95. On the basis of the overall results presented above, it can be concluded that the setting of 5 intervals yields higher accuracy than 10 intervals.

Figure 5.3: Type I error rates, power and correlation ratio using 5 and 10 intervals.




Meanwhile, it is worth commenting that under the null $H_0$: $\theta = 0$, bias $\theta$ was essentially zero, as $\theta_B \rightarrow \theta = 0$. Under the alternative, bias $\theta$ was also zero at low correlation ($d = 1$) as $\theta_B \rightarrow \theta = \theta_W$. Consequently, the coverage probabilities were essentially 0.95 under $H_0$ and when $d = 1$ under $H_1$. This topic has been discussed in Section 5.1. A general comparison is now given based on the individual case presented earlier. Unlike the other cases, the recurrent TT data is structured such that

$T_1$ is directly contained within $T_2$, hence embedding the treatment advantage from the 1st recurrent into the 2nd, resulting in larger treatment advantage and power. Whereas in a combined test, these effects are diluted as governed by the weighting scheme used (Section 4.4), leading to a smaller power compared to that for the marginal test on $T_2$. Also, notice that the values of $\theta$ for the paired, recurrent TT and recurrent GT cases (in the previous sections) are appreciably larger than those for other cases, to compensate for the heavy censoring imposed.

For the cases of complete, paired and indicators, the assumption of a common treatment effect is valid and hence reflected in the superiority of the global power compared to those of the marginal. PFS shows a slight variation between the estimates of treatment effect, and the marginal power is close to the global. The recurrent TT shows wider spread of the estimates, resulting in marginal power of bigger than the global. The superiority of global power is recovered when the 2nd option of data generation is used. Meanwhile, for the recurrent GT, the estimates are equal when $d$ = 1, hence the global superiority is observed. However, as the correlation increases, the estimates of marginal treatment effects start to vary and at $d$ =10, the marginal power based on the 1st event exceeds the global power.

The results in this section imply that the method of data generation described in earlier sections gave data sets with proportional hazards for most of the cases, except for the recurrent events, although the 2nd option was close. The relationship between adherence to this key assumption and the performance of the global test shown here is similar to that of the proportional odds assumption displayed in Table 2.12 on Page 50. Overall, this evidence further strengthens the theoretical advantage of the relationship between the Fisher's information and sample size, as well as the superiority of the global test methodology when the assumption is valid.

## 5.3. Discussion

The proposed method has achieved the intended type I error rate under the null hypothesis of zero treatment effect, and yielded accurate power relative to its theoretical power under the alternative hypothesis. The fundamental equations in Section 5.1, referencing to Section 1.7, simplify the procedure for deriving sample size and power for a clinical trial. This illustrates the benefit of the global test approach for the correlated bivariate survival data analysis. Meanwhile, the overall correlation ratio is good with a slight underestimation of 1%.

The comparison of the powers of marginal tests to that of the combined test in the previous section reveals an interesting finding with regard to the recurrent TT case. The higher power of the marginal test on $T_2$ indicates that combined analysis may not be beneficial for the case of recurrent TT. This may be one of the scenarios in which previous authors have cautioned about regarding the appropriateness of use of the global test approach (Section 2.1). Additionally, when the data were generated differently, so as to give better adherence to the assumption of proportional hazards, the global power won back. It is to be learned from this particular simulation that perhaps when the correlation between the endpoints is high (e.g. the $2^{nd}$ contains the $1^{st}$), combining their analyses into one may not be beneficial: one may as well use the simpler marginal analysis for each endpoint.

Since the assumption of a proportional hazards ratio is deployed in this study design, any deviation from it may degrade the power of the test. In reality, this assumption is not always true and perhaps using the global test is still the best choice in general. For example, in the case of recurrent TT (Section 5.2.5), we may lose a bit on the power of the global test over the marginal, but in most cases the former is clearly superior to the latter. The global test methodology is intended for use where

the correlation is positive, and it may be inappropriate in the case of negative correlation as might arise in situations such as duration of stay in ICU and time to kidney failure. As explained in Section 5.1, the diminishing power at high correlation is anticipated owing to the data generation method which yields $\theta_R$ associated with the log hazard ratio within subject, $\theta_W$. The remaining challenge is to devise a method that gives the appropriate $\theta_R$ which is based on the log hazard ratio between subjects, $\theta_B$.

In all of the analyses, the continuous data sets (real and simulated) were transformed into interval-censored data to enable the application of the proposed method. A problem inherent in our approach is that taking the intervals leads to a slight loss of power, as shown in the simulation results. Furthermore, appropriate choice of the number of intervals, as well as of the size of the interval, is important. Based on the exploration undertaken, setting up five intervals seemed to perform better than setting up ten in terms of the key performance measures defined. It is worth noting that the intervals used for the 10,000 simulation runs were fixed and thus may not have accurately divided the events from each run equally across each interval; although the impact should be minimal since the intervals were obtained using a very large sample size of 1 million.

As described in earlier sections, many real clinical studies deal with interval-censored survival data. An example of a naturally interval-censored survival data set is seen in a clinical trial comparing treatment advantage by analysing X-ray gradation which measures the loss of cartilage in the knee of arthritis patients (Whitehead & Thomas, 1997). As observed in the simulation study, application of interval censoring to continuous survival data should be carried out with caution as the number of intervals may affect the results appreciably.

To conclude, the encouraging results, evident for both real and simulated data, show that our method works for the bivariate interval-censored survival data: complete, paired, related indicators, progression-free survival and recurrent TT. However, its suitability as a correlation estimator for the recurrent gap time model is questionable and is further investigated in Chapter 6, upon comparing it with the existing method of Wei, Lin Weissfeld (1989).

# Chapter 6. Comparison with the Wei, Lin and Weissfeld Method

In this chapter, a well-known approach to the estimation of the correlation between two estimates of treatment advantage for multivariate survival data, namely the Wei, Lin and Weissfeld method (WLW), is compared with our present method (ZW), which was described and investigated respectively in Chapters 4 and 5.

To begin with, a number of existing methods for multivariate survival analysis are introduced to paint the background in Section 6.1. This introduction is followed in Section 6.1.1 by an overview of marginal models. A description of the influence of an individual observation on a parameter estimate, including the jackknife technique and the delta-beta approximation is given in the subsequent sections; these are related to WLW, as will be shown in later sections. The concepts of the WLW approach will be described at length in Section 6.2, and its use in estimating overall treatment effect in Section 6.3. The subsequent Section 6.4 focuses on a theoretical comparison of ZW and WLW.

Practical applications to real data sets are demonstrated in Section 6.5, followed by a short discussion to preview Section 6.6 which centres on a simulation study to evaluate and compare the accuracy of these contending methods. Simulation results for each of the seven scenarios are presented and the overall results are then summarized in Section 6.7. A discussion of the findings concludes this chapter.

## 6.1. Background

As clarified in earlier chapters, the scope of this thesis covers bivariate survival data, which provide the simplest case of multivariate data; an extension to other multivariate cases is also feasible. By definition, multivariate survival data arise when (i) each subject may experience several events or (ii) there exists some clustering of subjects, which induces dependence among event times within the same cluster. Extensions of methods of univariate survival analysis to the multivariate setting have proved to be rather difficult, resulting in many different approaches: three of these are known as marginal, frailty and copula.

In marginal models, the association structure is left unspecified. It is to be recalled that our proposed method (ZW) decribed in Chapter 4, was developed from a marginal model. In this chapter, other marginal methods are now reviewed: namely AG (Andersen and Gill, 1982), PWP (Prentice, Williams and Peterson, 1981), and LWA (Lee, Wei and Amato, 1992) as well as WLW (Wei, Lin and Weissfeld, 1989). The methods of AG, PWP and LWA will be covered briefly in the next section, while WLW is described in more detail throughout this chapter.

Another approach is a frailty model, whereby the distribution of a random effect is specified. Meanwhile, a copula model offers an alternative by combining marginal distributions via a copula function, which specifies the dependence structure. Both the frailty model and the copula model are not within the scope of this work, but good descriptions are made available by Hougaard (2000) and Sun (2006) respectively.

Over the past two decades, WLW has become a popular method for dealing with recurrent events. To date, the paper by Wei et al. (1989) has been cited over 600 times. It has attracted criticism from many quarters, and a complete summary and appraisal are given by Metcalfe & Thompson (2007). They also claimed that as of August 2003, application of WLW in the analysis of recurrent events was reported in 31 articles, while the use of the PWP method was reported in 24 publications. Several other authors have also studied WLW and assessed its suitability in many aspects; application to recurring and terminating events (Li and Lagakos, 1997), adaptation to the competing risks (Wei and Glidden, 1997), analysis of composite endpoints of longitudinal and survival data (Seville, Herring and Koch, 2010), and systematic characterization of methods for recurrent events using risk intervals, baseline hazard, risk set and correlation adjustment (Kelly and Lim, 2000). To describe the WLW method, we shall start with an overview of the existing marginal models, following this by a review of the jackknife technique as well as of both the exact delta beta technique and its approximation DFBETA, which are important in deriving the robust variance estimation in WLW.

### 6.1.1. **Marginal Models**

Marginal models are methods where the effects of explanatory variables are estimated on the basis of the marginal distributions. In the case of bivariate survival data involving two endpoints, this simply means that the set of data, say $T_1$ and $T_2$, for each endpoint is fitted to a standard Cox's proportional hazards (PH) regression model separately without assuming any correlation. The marginal hazard function for the $m^{\text{th}}$ type of event on the $i^{\text{th}}$ subject is given by $h_m(t;\boldsymbol{x}_{im}) = h_{0m}(t)\exp(\boldsymbol{\beta}'\boldsymbol{x}_{im})$ for a set of $p$ fixed covariates $\boldsymbol{x}_i$ with a vector of regression coefficients, $\boldsymbol{\beta}$, $m$ = 1, 2. It is to be

noticed that the baseline hazard is specific to the $m^{th}$ event, but can be considered as common across all $m$ events by omitting the subscript $m$ accordingly. The choice of stratification has been described in Sections 1.6 and 4.6.1.

As mentioned earlier, the 'working model' assumes that the observations are independent. A model-based variance estimate is obtained from the inverse of the information matrix. However, due to the correlation between event times, it may not be a consistent estimate of the asymptotic variance. Thus, the estimates of the regression coefficients are used with empirical adjustment of their estimated variance using a sandwich estimator (further described later). This is the means to account for within-subject correlation. In this section, the well-known marginal models due to Andersen & Gill (1982), Prentice, Williams & Peterson (1981), Lee, Wei & Amato (1992) and Wei, Lin & Weissfeld (1989) are reviewed. Only the general principles are considered here, while more technical details are provided in those references.

The intensity model, AG (Andersen and Gill, 1982) assumes that all events are independent and are of the same type, and hence provides estimates of the treatment effect on the rate of events as a whole. Meanwhile, the PWP model (Prentice et al., 1981) considers an extension of the stratified PH model with the $m$ events as strata: the hazard function is event-specific. AG and PWP methods originally used a model-based covariance estimate which is the default in the SAS PROC PHREG procedure, but the robust variance sandwich estimator has been applied to these models by other researchers (Finkelstein, Schoenfeld and Stamenovic, 1997), to adjust for correlation.

The LWA model (Lee et al., 1992) is unstratified and hence assumes a common baseline hazard. It allows a subject to be at risk of multiple events simultaneously and accounts for the within-subject correlation by using the sandwich estimator. The WLW model (Wei et al., 1989) is rather similar to LWA except that the

former is stratified by event type, giving event-specific hazard functions. For each model, the specific partial likelihood can be constructed accordingly in order to derive the parameter estimates and their variances. A good comparison of the hazard functions and partial likelihoods for these methods is given by Kelly & Lim (2000).

In normal distribution models, residuals are often evaluated for the purposes of checking the model assumptions and evaluating the variance estimates. Residuals for survival data can be similarly evaluated, although the precision depends on the nature of the data: observed event time or censored observation. The variance estimates derived in this manner are different from those usually obtained from the second derivative of the log likelihood function, which are the usual model-based estimators. The variance matrix of the regression coefficients is given by $\hat{Y} = \boldsymbol{I}^{-1}\boldsymbol{R}\boldsymbol{I}^{-1}$, where $\boldsymbol{I}$ is the matrix of second derivative of the log likelihood function and $\boldsymbol{R}$ is the variance matrix of a score vector. This sandwich-like form renders the name sandwich estimator, also often known as the robust covariance matrix estimator or the empirical covariance matrix estimator. A more detailed account of the derivation of residuals and influence diagnostics specifically for Wei, Lin & Weissfeld is given in the subsequent sections.

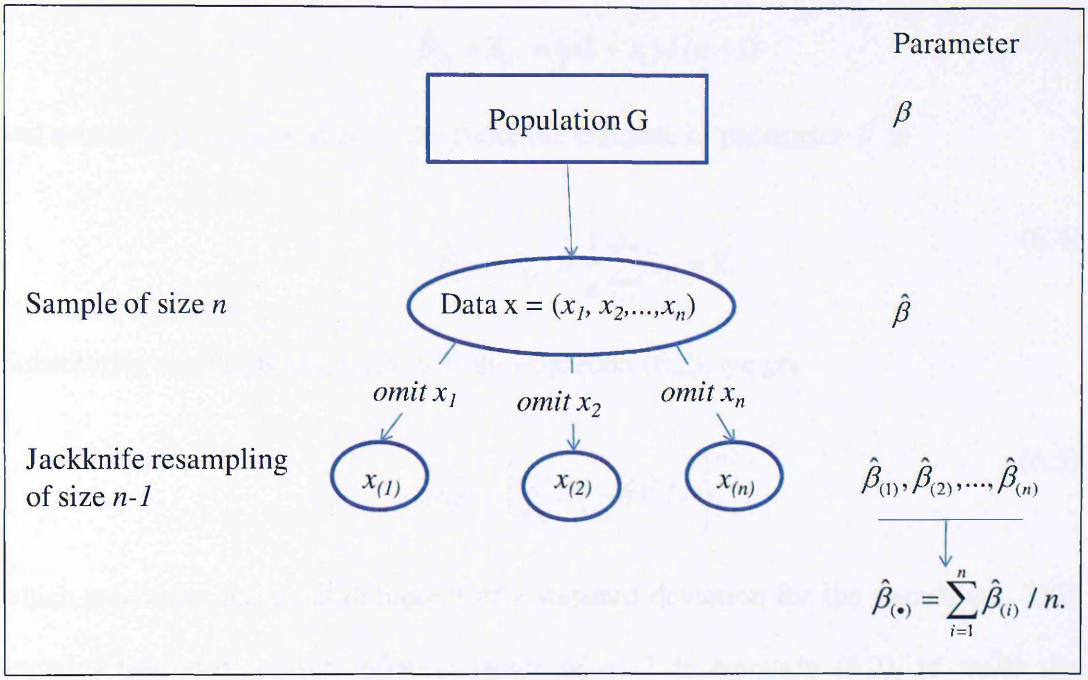
The sandwich estimator has achieved increasing use with the growing popularity of generalized estimating equations (GEE) in linear models; a similar version in survival analysis is provided by Wei et al. (1989). Traceable to Huber (1967) and White (1982), the method yields asymptotically consistent estimates of the covariance matrix for parameter estimates when a parametric model assumption is invalid, or is not even specified. Standard statistical softwares incorporate the various types of variance estimates into analyses such as SAS PROC PHREG, which can give both the model-based and sandwich-based statistics: illustrated in Section 6.2.

### 6.1.2. **Jackknife Method**

Jackknifing is a well-known computational technique for estimating the bias and standard error of an estimate; which is very similar to its famous successor, the bootstrap. The general idea was proposed by Quenouille (1949) for estimation of bias, and Tukey (1958) explored its potential for estimating standard errors: further development has been provided by others since then. As with any resampling technique, the jackknife does not make any specific distributional assumption. The procedure consists in taking repeated subsamples of the original sample of $n$ independent observations by omitting a single observation at a time: useful in evaluating subject or case influence. Its relationship to WLW is quite interesting and shall be demonstrated in the subsequent sections.

Suppose we are interested in a parameter $\beta$ on a basis of random sample $\boldsymbol{x} = (x_1, x_2,...,x_n)$ of an unknown population $G$. Taking this random sample of $n$ observations would give an estimate $\hat{\beta}$ which may be biased, and for which estimation of standard error is needed. The jackknife uses the sample by removing the first observation $x_1$, leaving a "jackknife data set" of resampled values, say $\boldsymbol{x}_{(1)}$. The statistical analysis is performed on the reduced sample size of $n$ - 1, giving another value of the parameter estimate, say $\hat{\beta}_{(1)}$. Then another resampling is done, this time throwing out the second observation $x_2$, and $\hat{\beta}_{(2)}$ is obtained from the subsample $\boldsymbol{x}_{(2)}$. The process is repeated for each observation in the sample resulting in a set of values $\hat{\beta}_{(1)}, \hat{\beta}_{(2)},..., \hat{\beta}_{(n)}$. Thus, each subsample $\boldsymbol{x}_{(i)}$ comprises $n - 1$ observations formed by deleting a different observation from the sample of size $n$. A schematic illustrating the jackknife algorithm for estimating parameter is provided in Figure 6.1.

Figure 6.1: Schematic of the jackknife algorithm for estimating parameter, inspired by the bootstrap algorithm (Efron & Tibshirani, 1993 p48).




The jackknife estimate $\hat{\beta}_{(\bullet)}$ and its standard error are then calculated from these reduced subsamples: $n$ subsamples each of size $n$ -1. The former is the mean of the jackknife estimates $\hat{\beta}_{(i)}$, given by

$$\hat{\beta}_{(\bullet)} = \sum_{i=1}^{n} \hat{\beta}_{(i)} / n \tag{6.1}$$

while the latter is defined as (Efron and Tibshirani, 1993)

$$\widehat{se}_{JK} = \left[ \frac{n-1}{n} \sum_{i=1}^{n} (\hat{\beta}_{(i)} - \hat{\beta}_{(\bullet)})^2 \right]^{1/2}. \tag{6.2}$$

Notice that the jackknife estimate of the standard error for $\hat{\beta}_{(\bullet)}$, involves an inflation factor of $n$-1 compared to the standard deviation for $\hat{\beta}_{(i)}$.

To illustrate the need for this inflation factor, Efron & Tibshirani (1993) consider a special case where $\beta = \mu$ and $\hat{\beta} = \bar{x}$. Then it can be shown that

$$\hat{\beta}_{(i)} = \bar{x}_{(i)} = (n\bar{x} - x_i)/(n-1) \tag{6.3}$$

and subsequently the average of the jackknife estimate of parameter $\beta$ is

$$\hat{\beta}_{(\bullet)} = \bar{x}_{(\bullet)} = \frac{1}{n}\sum_{i=1}^{n} \bar{x}_{(i)} = \bar{x}. \tag{6.4}$$

Substituting equations (6.3) and (6.4) into equation (6.2), we get

$$\widehat{se}_{JK} = \left\{ \sum_{i=1}^{n} (x_i - \bar{x})^2 / n \right\}^{1/2}, \tag{6.5}$$

which resembles the usual definition of a standard deviation for the population. This explains the need for the inflation factor of $n$ -1 in equation (6.2), to make the jackknife estimate of the standard error unbiased. The reason for the difference is that the jackknife sample means $\hat{\beta}_{(i)}$ are distributed $n$ -1 times closer to the mean $\hat{\beta}_{(\bullet)}$ than the original values, $\hat{\beta}_i$. Consequently, the jackknife estimate of variance for $\hat{\beta}$ is

$$\text{var}(\hat{\beta}) = \frac{n-1}{n}\sum_{i=1}^{n} (\hat{\beta}_{(i)} - \hat{\beta}_{(\bullet)})^2. \tag{6.6}$$

Although the inflation factor is justified by considering a special case in which $\hat{\beta} = \bar{x}$, it appears to be workable in general. Note that $\hat{\beta}$ is a scalar parameter in the above expression for univariate data. For ease of referencing, the term jackknife influence represents the quantity $\hat{\beta}_{(i)} - \hat{\beta}_{(\bullet)}$ hereafter in this thesis.

Suppose $n$ subjects are randomized between treatments $E$ and $C$. Equations (6.1) and (6.6) can be used to obtain the jackknife estimates of treatment advantage

and its variance respectively. In the case of bivariate survival data with two event times, $\boldsymbol{\beta}$ is now a vector parameter of treatment effects, and the corresponding estimate, $\hat{\boldsymbol{\beta}}' = (\hat{\beta}_1, \hat{\beta}_2)$. Their variances, $\text{var}(\hat{\beta}_1)$ and $\text{var}(\hat{\beta}_2)$ can be obtained using equation (6.6). The jackknife influence can be generalized by $\hat{\beta}_{(i)m} - \hat{\beta}_{(\bullet)m}$ where $m$ = 1, 2 for such bivariate case. To further simplify, put $J_{mi} = \hat{\beta}_{(i)m} - \hat{\beta}_{(\bullet)m}$ and equation (6.6) is rewritten as $\text{var}\,\hat{\beta}_m = \{(n-1)/n\}\sum_{i=1}^{n} J_{mi}^{\ 2}$.

As described in Section 4.1, to estimate an overall treatment advantage, $\hat{\beta}$, we need to find the covariance between the two estimates of the marginal treatment advantages, $\text{cov}(\hat{\beta}_1, \hat{\beta}_2)$. According to Efron (1982), the jackknife variance-covariance matrix between two parameter estimates, $\hat{\beta}_1$ and $\hat{\beta}_2$ is given by

$$\text{cov}(\hat{\beta}_1, \hat{\beta}_2) = \frac{n-1}{n}\sum_{i=1}^{n}(\hat{\beta}_{(i)1} - \hat{\beta}_{(\bullet)1})(\hat{\beta}_{(i)2} - \hat{\beta}_{(\bullet)2}) \tag{6.7}$$

The covariance is given by the summation of the product of the jackknife estimates for the 1st and 2nd events, with an inflation factor similar to that for the variance above. This can be rewritten as

$$\text{cov}(\hat{\beta}_1, \hat{\beta}_2) = \frac{n-1}{n}\sum_{i=1}^{n} J_{1i}J_{2i}. \tag{6.8}$$

Consequently, the jackknife estimate of correlation can be expressed as

$$\text{corr}_{JK}(\hat{\beta}_1, \hat{\beta}_2) = \frac{\sum_{i=1}^{n} J_{1i}J_{2i}}{\sqrt{\sum_{i=1}^{n} J_{1i}^{\ 2}\sum_{i=1}^{n} J_{2i}^{\ 2}}}. \tag{6.9}$$

In matrix notation, this can be further simplified by taking $\boldsymbol{J}_m' = [J_{m1},...,J_{mn}]$ being a column vector of the $n$ jackknife estimates for $m^{th}$ event, $m$ = 1, 2, and giving the covariance as

$$\text{cov}(\hat{\beta}_1, \hat{\beta}_2) = \frac{n-1}{n} \boldsymbol{J}_1' \boldsymbol{J}_2, \tag{6.10}$$

and the correlation

$$\text{corr}_{JK}(\hat{\beta}_1, \hat{\beta}_2) = \frac{\boldsymbol{J}_1' \boldsymbol{J}_2}{(\boldsymbol{J}_1' \boldsymbol{J}_1 \boldsymbol{J}_2' \boldsymbol{J}_2)^{1/2}}. \tag{6.11}$$

The correlation between the jackknife estimates of parameters $\hat{\beta}_1$ and $\hat{\beta}_2$ is used for comparison with alternative methods in the next section.

Efron & Tibshirani (1993) cover a good deal on the jackknife technique although specifically in its relation with the bootstrap method. Meanwhile, an account of the use of the jackknife in the derivation of a robust estimate of variance for the Cox's model is given by Therneau & Grambsch (2000). They claimed that such natural approximation $\text{cov}(\hat{\beta}_1, \hat{\beta}_2) \approx \boldsymbol{J}_1' \boldsymbol{J}_2$ by ignoring the inflation factor *(n - 1/n),* as mentioned earlier may be preferable. A study by Lipsitz et al. (1990) concluded that this approximation is indeed preferred even for a small sample size. It resembles a sandwich estimator which is common from robust variance estimation in parametric models and in GEE-like methods such as that of Liang & Zeger (1986). A sandwich estimator appropriate to the Cox's model has been derived by Lin and Wei (1989) and is algebraically equivalent to $\boldsymbol{J}_1' \boldsymbol{J}_2$ but was not described within the jackknife context: a point explored in the subsequent sections of this chapter.

### 6.1.3. **Exact Delta Beta and its Approximation (DFBETA)**

A logical way of evaluating the influence of any observation is by observing the difference in parameter estimates, one made when that particular observation is omitted, and the other made when the full data set is used. This procedure is related to the jackknife method with the exception that the effect of removing $i^{th}$ observation is quantified by the difference between the parameter estimates $\hat{\beta}$ and $\hat{\beta}_{(i)}$. Instead of taking the difference between the average of the jackknife estimates, $\hat{\beta}_{(\bullet)}$ and $\hat{\beta}_{(i)}$, the parameter estimate $\hat{\beta}$, which is obtained from the complete sample size of $n$ is used. In this thesis, the actual influence of the observation termed "delta-beta" is expressed as $\Delta_i\hat{\beta} = \hat{\beta}_{(i)} - \hat{\beta}$; adopting the notation used for the plots in Section 4.3.1 of Collett (2003).

Both the jackknife influence ($\hat{\beta}_{(i)} - \hat{\beta}_{(\bullet)}$) and the delta-beta ($\hat{\beta}_{(i)} - \hat{\beta}$) require computation of $\hat{\beta}_{(i)}$ for their evaluation. The exact computation of $\hat{\beta}_{(i)}$ involves refitting the model each time a subject is omitted, which is computationally expensive when the sample size is large. The effect of removing one observation from a survival data set is more complicated than for other types of data. Mathematically, the log likelihood function for the Cox's model cannot be expressed as the sum of terms in which each term is the contribution by each observation. This implies that any exclusion of an observation cannot be simply modelled by omitting a single corresponding term. In fact, the exclusion of one observation affects the risk sets over which quantities of the form $\exp(\boldsymbol{\beta}'\boldsymbol{x})$ are summed (Collett, 2003). A practical option is given by Cain and Lange (1984) who derived the following approximation of delta-beta as weighted score residuals.

Suppose each individual has a set of data comprising a vector ($t$, $\delta$, $\boldsymbol{x}$) where $t$ is the time from the start of the study until death or censoring, $\delta$ is the censoring indicator, 0 for censored, 1 for death, and $\boldsymbol{x}$ is a row vector of covariates. When there are no ties, the partial likelihood is given by the product of the partial likelihood corresponding to the risk set $R(t_j)$,

$$L(\boldsymbol{\beta}) = \prod_{i \in D} L_i(\boldsymbol{\beta}) = \prod_{i \in D} \left\{ \exp(\boldsymbol{x}_i \boldsymbol{\beta}) / \sum_{l \in R(t_j)} \exp(\boldsymbol{x}_l \boldsymbol{\beta}) \right\}$$

where $D$ is the set of subjects who died at time $t_i$ and $t_i \geq t_j$. Its similar form to that of the earlier equation (3.8) is to be noticed.

Cain & Lange (1984) derived an approximation to $\hat{\beta}_{(i)}$ based on the first-order Taylor series. Suppose that observation $j$ is given weight $w_j$ leading to a likelihood of the form

$$L(\boldsymbol{\beta}) = \prod_{j \in D} L_j(\boldsymbol{\beta}) = \prod_{j \in D} \left\{ \exp(\boldsymbol{x}_j \boldsymbol{\beta}) / \sum_{l \in R(t_j)} w_l \exp(\boldsymbol{x}_l \boldsymbol{\beta}) \right\}^{w_j}. \tag{6.12}$$

Then, the log likelihood is given by

$$\ell(\boldsymbol{\beta}) = \sum_{j \in D} w_j \left\{ (\boldsymbol{x}_j \boldsymbol{\beta}) - \log \sum_{l \in R(t_j)} (w_l \exp(\boldsymbol{x}_l \boldsymbol{\beta})) \right\} \tag{6.13}$$

Now, we focus on subject $i$, and its parameter estimate, $\hat{\beta}_{(i)}$. For approximating $\hat{\beta}_{(i)}$, Cain & Lange choose the weighting $w_j = w$ if $j = i$, $w_j = 1$ otherwise. Further denote the estimated vector of coefficients $\boldsymbol{\beta}$ when this weighting is used, by $\hat{\boldsymbol{\beta}}_i(w)$. If $w = 1$, then all observations are weighted equally and the standard estimation is obtained: $\hat{\boldsymbol{\beta}}_i(1) = \hat{\boldsymbol{\beta}}$. If $w = 0$, then the $i^{\text{th}}$ observation is simply excluded,

so that $\hat{\boldsymbol{\beta}}(0) = \hat{\boldsymbol{\beta}}_{(i)}$. The approximation to delta-beta based on the first-order Taylor series expansion about $w = 1$ is thus approximately equal to the derivative of $\hat{\boldsymbol{\beta}}_i(w)$ where

$$\hat{\boldsymbol{\beta}}_i(w) = \hat{\boldsymbol{\beta}}_i(1) + (w-1)\frac{\partial \hat{\boldsymbol{\beta}}_i(1)}{\partial w}. \tag{6.14}$$

Now substituting $w = 0$ and re-arranging, we have

$$\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)} = \partial \hat{\boldsymbol{\beta}}_i(1) / \partial w, \tag{6.15}$$

as stated by Cain & Lange. The derivative term in equation (6.15) is the approximation used in the infinitesimal jackknife (Miller, 1974; Efron, 1981). It is to be recalled that previously, the score statistic was denoted by $Z$ when evaluated under the null. In this section, $\boldsymbol{U}$ is used instead, as the the score vector is now a function of both $\hat{\boldsymbol{\beta}}(w)$ and $w$. To evaluate $\partial \hat{\boldsymbol{\beta}}_i(1) / \partial w$, their relationship is defined at $\boldsymbol{U}\left(\hat{\boldsymbol{\beta}}(w), w\right) = 0$. The derivative of the score vector is thus given by

$$(\partial \boldsymbol{U} / \partial \hat{\boldsymbol{\beta}})(\partial \hat{\boldsymbol{\beta}} / \partial w) + (\partial \boldsymbol{U} / \partial w)(\partial w / \partial w) = 0, \tag{6.16}$$

and therefore the delta-beta approximation is

$$\partial \hat{\boldsymbol{\beta}} / \partial w = (-\partial \boldsymbol{U} / \partial \hat{\boldsymbol{\beta}})^{-1} \partial \boldsymbol{U} / \partial w. \tag{6.17}$$

It is to be noted that the term in parentheses is the observed information matrix and is denoted by $\boldsymbol{I}$ in this chapter. From equation (6.13), the log likelihood can be re-expressed as

$$\ell(\boldsymbol{\beta}) = \boldsymbol{\beta} \sum_{j \in D} w_j \boldsymbol{x}_j - \sum_{j \in D} w_j \log \left\{ \sum_{l \in R(t_j)} w_l \exp(\boldsymbol{x}_l \boldsymbol{\beta}) \right\}. \tag{6.18}$$

Following the relationship described in Section 1.4, and from equation (6.18), the derivative of the log likelihood with respect to $\boldsymbol{\beta}$ is the score vector $\boldsymbol{U}$ :

$$\boldsymbol{U} = \frac{\partial \ell}{\partial \boldsymbol{\beta}} = \sum_{j \in D} w_j \boldsymbol{x}_j - \sum_{j \in D} w_j \frac{\sum_{l \in R(t_j)} w_l \boldsymbol{x}_l \exp(\boldsymbol{x}_l \boldsymbol{\beta})}{\sum_{l \in R(t_j)} w_l \exp(\boldsymbol{x}_l \boldsymbol{\beta})}. \tag{6.19}$$

To distinguish explicitly between the contributions when subject $i$ is in the risk set and that when subject $i$ is excluded from the risk set, further notation is necessary. Suppose $D_i$ is the set of all subjects, other than subject $i$ who have died, while $\delta_i = 1$ if subject $i$ died and $\delta_j = 1$ if subject $i$ is at risk when subject $j$ died, 0 otherwise. Equation (6.19) can be rewritten as

$$\begin{aligned} \boldsymbol{U} = & \sum_{j \in D_i} \boldsymbol{x}_j + w\delta_i \boldsymbol{x}_i - \sum_{j \in D_i} \frac{\sum_{l \in R(t_{j,i})} \boldsymbol{x}_l \exp(\boldsymbol{x}_l \boldsymbol{\beta}) + w\delta_j \boldsymbol{x}_i \exp(\boldsymbol{x}_i \boldsymbol{\beta})}{\sum_{l \in R(t_{j,i})} \exp(\boldsymbol{x}_l \boldsymbol{\beta}) + w\delta_j \exp(\boldsymbol{x}_i \boldsymbol{\beta})} \\ & - w\delta_i \sum_{j \in D_i} \frac{\sum_{l \in R(t_{j,i})} \boldsymbol{x}_l \exp(\boldsymbol{x}_l \boldsymbol{\beta}) + w\boldsymbol{x}_i \exp(\boldsymbol{x}_i \boldsymbol{\beta})}{\sum_{l \in R(t_{j,i})} \exp(\boldsymbol{x}_l \boldsymbol{\beta}) + w\exp(\boldsymbol{x}_i \boldsymbol{\beta})}, \end{aligned} \tag{6.20}$$

where $R(t_{j,i})$ is the risk set at time $t_j$, excluding subject $i$. Its derivative with respect to $w$ can be obtained via the product rule and subsequently the quotient rule,

$$\begin{aligned} \frac{\partial \boldsymbol{U}}{\partial w} = & \delta_i \boldsymbol{x}_i - \sum_{j \in D_i} \frac{\delta_j \{\exp(\boldsymbol{x}_i \boldsymbol{\beta})\}(A_j \boldsymbol{x}_i - B_j)}{A_j^2} - \delta_i \frac{\sum_{l \in R(t_i)} \boldsymbol{x}_l \exp(\boldsymbol{x}_l \boldsymbol{\beta}) + w\boldsymbol{x}_i \exp(\boldsymbol{x}_i \boldsymbol{\beta})}{\sum_{l \in R(t_i)} \exp(\boldsymbol{x}_l \boldsymbol{\beta}) + w\exp(\boldsymbol{x}_i \boldsymbol{\beta})} \\ & - \frac{w\delta_i \{\exp(\boldsymbol{x}_i \boldsymbol{\beta})\}(A_i \boldsymbol{x}_i - B_i)}{A_i^2}, \end{aligned} \tag{6.21}$$

where

$$A_j = \left\{ \sum_{l \in R(t_{j,i})} \exp(\boldsymbol{x}_l \boldsymbol{\beta}) + w\delta_j \exp(\boldsymbol{x}_i \boldsymbol{\beta}) \right\}, \quad B_j = \left\{ \sum_{l \in R(t_{j,i})} \boldsymbol{x}_l \exp(\boldsymbol{x}_l \boldsymbol{\beta}) + w\delta_j \boldsymbol{x}_i \exp(\boldsymbol{x}_i \boldsymbol{\beta}) \right\},$$

$$A_i = \left\{ \sum_{l \in R(t_{i,i})} \exp(\boldsymbol{x}_l \boldsymbol{\beta}) + w\exp(\boldsymbol{x}_i \boldsymbol{\beta}) \right\}, \quad B_i = \left\{ \sum_{l \in R(t_{i,i})} \boldsymbol{x}_l \exp(\boldsymbol{x}_l \boldsymbol{\beta}) + w\boldsymbol{x}_i \exp(\boldsymbol{x}_i \boldsymbol{\beta}) \right\}.$$

It is to be noted that $R(t_{i,i})$ is the risk set at time $t_i$, excluding subject $i$. Putting $w = 1$ in equation (6.21) for equal weighting, we get

$$\frac{\partial \boldsymbol{U}}{\partial w} = \delta_i \left\{ \boldsymbol{x}_i - \frac{\sum_{l \in R(t_i)} \boldsymbol{x}_l \exp(\boldsymbol{x}_l \boldsymbol{\beta})}{\sum_{l \in R(t_i)} \exp(\boldsymbol{x}_l \boldsymbol{\beta})} \right\} - \sum_{j \in D} \delta_j \left\{ \frac{\exp(\boldsymbol{x}_i \boldsymbol{\beta})}{\sum_{l \in R(t_j)} \exp(\boldsymbol{x}_l \boldsymbol{\beta})} \right\} \left\{ \boldsymbol{x}_i - \frac{\sum_{l \in R(t_j)} \boldsymbol{x}_l \exp(\boldsymbol{x}_l \boldsymbol{\beta})}{\sum_{l \in R(t_j)} \exp(\boldsymbol{x}_l \boldsymbol{\beta})} \right\}. \tag{6.22}$$

The change in the score vector $\boldsymbol{U}$ with respect to changes in $w$, is given by the sum of two parts, as shown in equation (6.22). It is to be noted that the first part is included only when individual $i$ died ($\delta_i = 1$), which is the difference between the covariates for $i$ and the weighted average of covariates for all individuals who are at risk at time $t_i$. This part is indeed the Schoenfeld residual (Schoenfeld, 1984). Meanwhile, the second part is a summation of the effects that changes in $w$ have upon all the risk sets that include individual $i$ at time $t_j$. Graphical representation of individual parts as well as the resulting approximation per equation (6.22) are illustrated and examined by Cain & Lange (1984). Combining equations (6.15), (6.17) and (6.22) provides an expression for $\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)}$, the delta-beta approximation.

An alternative expression is given by Collett (2003, p119) in the form of score residuals,

$$\boldsymbol{R}(\boldsymbol{\beta}) = \delta_i \left\{ \boldsymbol{x}_i - \frac{\sum_{l \in R(t_i)} \boldsymbol{x}_l \exp(\boldsymbol{\beta}' \boldsymbol{x}_l)}{\sum_{l \in R(t_i)} \exp(\boldsymbol{\beta}' \boldsymbol{x}_l)} \right\} - \sum_{t_j \le t_i} \delta_j \left( \frac{\exp(\boldsymbol{\beta}' \boldsymbol{x}_i)}{\sum_{l \in R(t_j)} \exp(\boldsymbol{\beta}' \boldsymbol{x}_l)} \right) \left\{ \boldsymbol{x}_i - \frac{\sum_{l \in R(t_j)} \boldsymbol{x}_l \exp(\boldsymbol{\beta}' \boldsymbol{x}_l)}{\sum_{l \in R(t_j)} \exp(\boldsymbol{\beta}' \boldsymbol{x}_l)} \right\}, \quad (6.23)$$

which is indeed the same as the quantity $W(\boldsymbol{\beta})$ formulated by Wei et al. (1989) as part of equation A.2 in their Appendix. This means that all these three quantities appearing in Wei et al. (1989), Collett (2003) and Cain & Lange (1984) are indeed different ways of expressing the same approximation.

Consequently, putting equation (6.23) into equation (6.15), the approximation to delta-beta can now be expressed as $\boldsymbol{R}\hat{\boldsymbol{I}}$ where $\boldsymbol{R}$ is the vector of score residuals as given above, and $\hat{\boldsymbol{I}}$ is the inverse of the information matrix. The quantity $\boldsymbol{R}\hat{\boldsymbol{I}}$ is an output known as the matrix of DFBETA residuals, readily given in statistical software packages such as SAS. As shown earlier, these DFBETA statistics are a transformation of the score residuals, and are utilized in computing the robust sandwich variance estimators of Lin and Wei (1989) and Wei, Lin, and Weissfeld (1989). Further description follows in later sections.

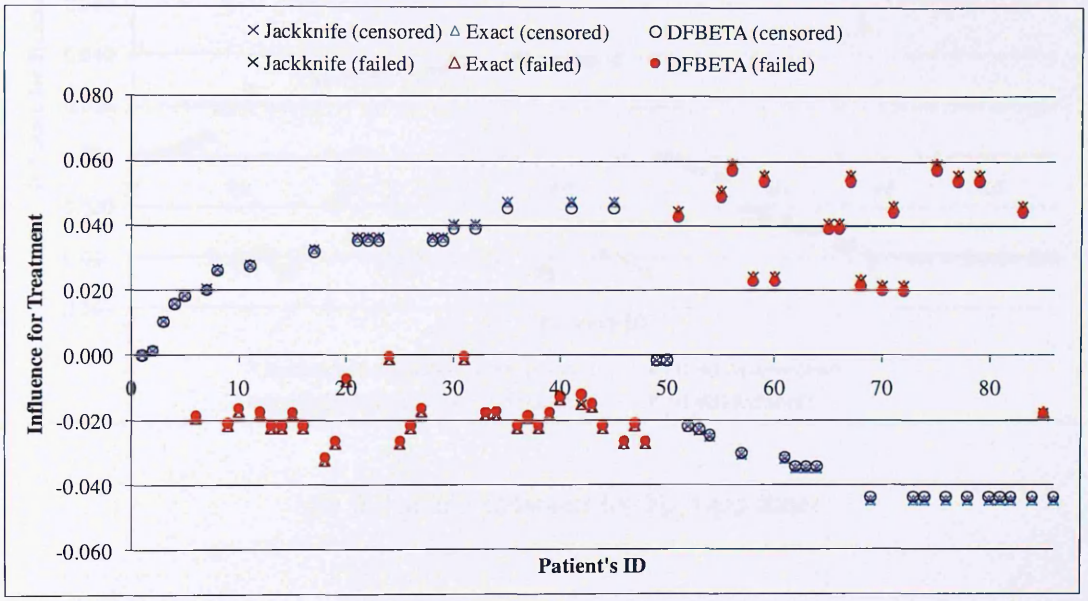
### 6.1.4. **Comparison of Influences: Jackknife, Delta-beta and DFBETA**

The jackknife influence, exact delta-beta and DFBETA have been described in the previous sections. Considering bivariate data ($m$ = 1, 2), a comparison between the jackknife influence ($\hat{\boldsymbol{\beta}}_{m(i)} - \hat{\boldsymbol{\beta}}_{m(\bullet)}$), the exact delta-beta ($\hat{\boldsymbol{\beta}}_{m(i)} - \hat{\boldsymbol{\beta}}_{m}$) and the DFBETA ($\boldsymbol{R}_m \hat{\boldsymbol{I}}_m$), is shown below. The construction of these plots is quite similar to that given in Cain & Lange (1984) and Section 4.3.1 of Collett (2003), except for the inclusion of the jackknife influence. In Figures 6.2 (a) to (c) below, the vertical axis is the quantity we can generally call influence which are evaluated using the jackknife influence, exact delta beta, and DFBETA.

In this section, the bladder cancer data (Section 4.6.2) is used for illustration. Due to the nature of the data set, whereby the times are recorded to the nearest month resulting in many ties, plotting against the rank order of event times as per Cain & Lange (1984) and Collett (2003) is not favoured. Alternatively, we shall examine the influence for treatment against the patient's ID via these three methods. $T_1$ and $T_2$ are the times from start of study to the 1st and 2nd recurrence of tumour respectively based on 86 patients in the bladder cancer study described in Chapter 4 earlier. Influences based on $T_{2G}$ are also plotted for the gap time between the 1st and 2nd recurrences.

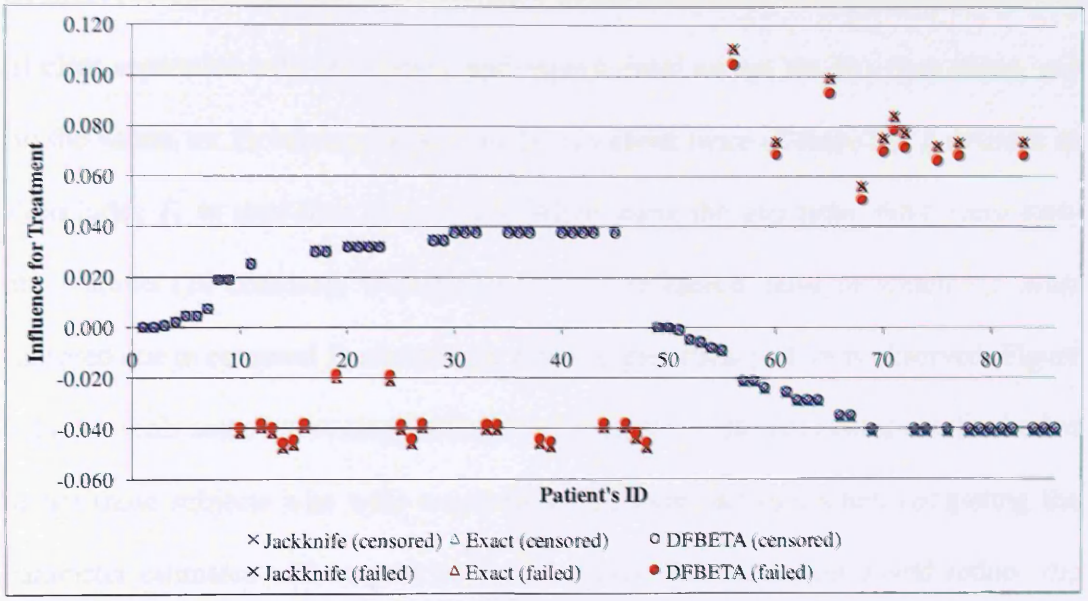
Figure 6.2: Plot of the influence using the jackknife, exact and DFBETA methods for treatment against patient id for the bladder cancer data

(a) failed and censored for $T_1$

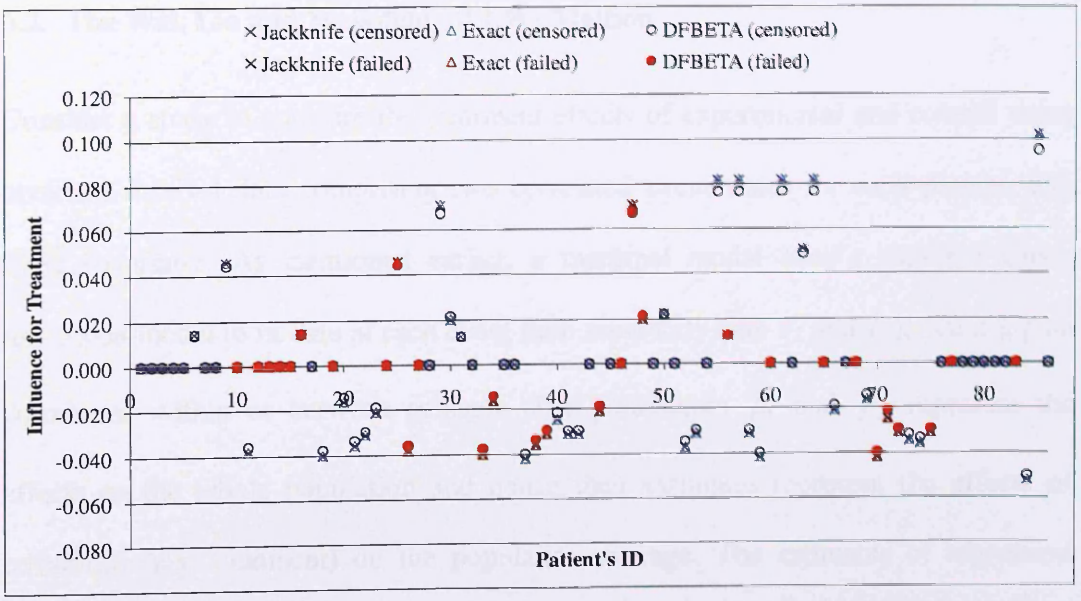



The points in red are influences for the patients who failed and blue indicates those who were censored. Notice a clear separation (above and below 0) between patient ID 1 to 48 on control from those on experimental. It is observed that DFBETA gives slightly lower values than the exact method, and the jackknife. There are two large influences of about 0.06 for treatment observed for patients 56 and 75 on experimental who both failed at month 1. This suggests that the exclusion of either patient changes the hazard of tumour recurrence relative to the baseline hazard, thus making the effect of treatment either slightly more or slightly less significant, but may or may not necessarily be of clinical importance. Meanwhile, patients 49 and 50 on experimental who were censored at month 1, and patients 24 and 31 on control who failed at month 5, show almost zero influence for treatment. Figures 6.2 (b) and (c) show similar plots for $T_2$ and $T_{2G}$.

(b) failed and censored for $T_2$ (total time)




(c) failed and censored for $T_{2G}$ (gap time)




It is evident from these plots that all three methods give very close values of the influence. The jackknife technique gives values closest to the exact method, while the approximate DFBETA approach tends to underestimate slightly for some patients. No incident of overestimation by DFBETA is observed in any of the plots above. This

illustration supports (although it is not a comprehensive justification) the use of DFBETA as an economical approximation to the delta-beta. Two points to note are: (i) clear separation between control and experimental except for $T_{2G}$ (gap time), and (ii) the values for $T_2$ failures (beyond id 48) are about twice of those for $T_1$ failures as $T_2$ includes $T_1$ in total time convention. When using the gap time, there were forty observations (26 censored, 14 failed) with zero influence, most of which $T_{2G}$ were censored due to censored $T_1$ at month 1, while no particular pattern is observed. Figure 6.2 (c) reveals some interesting findings. For example, it did not matter much whether or not those subjects who were censored for $T_1$ were included when computing the parameter estimates with respect to $T_{2G}$. However, their inclusion would reduce the variance of such estimates for $T_{2G}$. Now that the relationship between the jackknife and DFBETA is clearly established, the WLW method is described in the next section.

## 6.2. The Wei, Lin and Weissfeld (WLW) Method

Consider a study to compare the treatment effects of experimental and control using bivariate survival data comprising two correlated event times for each patient with fixed covariates. As mentioned earlier, a marginal model uses a standard Cox's regression model to fit data at each event time separately (say $T_1$ and $T_2$), assuming no correlation within or between subjects. The parameters $\beta_1$ and $\beta_2$ represent the effects on the whole population and hence their estimates represent the effects of covariates (e.g. treatment) on the population average. The estimates of regression coefficients $\hat{\beta}_1$ and $\hat{\beta}_2$ are then used, but their estimated variance is empirically adjusted using a sandwich estimator. This is analogous to the GEE approach of Liang and Zeger (1986), applied to Cox's models by Lin and Wei (1989); Wei, Lin and Weissfeld (1989); Lin (1991); and Lee, Wei and Amato (1992).

WLW offers a solution for analyzing bivariate survival by providing robust variance estimates that allow for multiple event times. To date, none of the mentioned methods, with the exception of WLW, give the estimation of the correlation coefficient and hence the covariance which is our key objective in this research. This justifies the comparison of our method with that of WLW. As already noted, WLW uses the standard Cox's PH model, stratified by type of event, in computing the estimates. Suppose we have the $m^{th}$ type of event, $m = 1, 2,$ let $t_{mi}$ be the event time of the $i^{th}$ subject, $i = 1, 2, ..., n$ and $t_{mi} = min(t_{mi}, c_{mi})$, where $c_{mi}$ is the censoring time. For each observed event time, let $\delta_{mi} = 0$ if $t_{mi} = c_{mi}$ (censored), and 1 otherwise. Now consider a $p$ x $1$ vector of covariates $\boldsymbol{x}_{mi} = (x_{1mi}, ..., x_{pmi})'$ for the $i^{th}$ subject at time $t \geq 0$ with regard to the $m^{th}$ type of event. By analogy to equation (3.8), the partial likelihood specific to the $m^{th}$ event can be expressed as

$$L_m(\boldsymbol{\beta}_m) = \prod_{i=1}^{n} \left[ \frac{\exp\{\boldsymbol{\beta}_m' \boldsymbol{x}_{mi}\}}{\sum_{l \in R_m(t_{mj})} \exp\{\boldsymbol{\beta}_m' \boldsymbol{x}_{ml}\}} \right]^{\delta_{mi}} \tag{6.24}$$

where $\boldsymbol{\beta}_m' = (\beta_{1m}, ..., \beta_{pm})'$ is the event specific regression parameter and $R_m(t_{mj}) = \{l : t_{mj} \leq t_{mi}\}$ is the risk set just before event time $t_{mi}$ with regard to the $m^{th}$ event. In situations of common $\boldsymbol{\beta}$, the subscript $m$ in equation (6.24) is to be omitted. As usual, the maximum partial likelihood estimator $\hat{\boldsymbol{\beta}}_m$ is defined as the solution to the first derivative of the log likelihood: $d\ell / d\boldsymbol{\beta}_m = 0$, and these $\hat{\boldsymbol{\beta}}_m$ s are generally correlated. These maximum likelihood estimates are therefore the values that maximize the likelihood function. Details of a complicated derivation using martingale and counting process theories are not reproduced here and can be found in

Wei et al. (1989). Alternatively, we describe WLW in a simpler way, based on the fundamental influence diagnostic described in the previous section.

As introduced earlier, WLW uses Cox's model directly to obtain the estimates for each event separately, $\hat{\beta}_1$ and $\hat{\beta}_2$, while the adjusted standard error is given by using the robust sandwich variance estimation method. Wei et al. (1989) showed that asymptotically, $\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \boldsymbol{Y})$, where $\boldsymbol{\beta} = (\beta_1, \beta_2)'$ and the estimated covariance matrix $\hat{Y}$ is composed of the submatrices $\hat{\boldsymbol{Y}}_{mm} = (\boldsymbol{R}_m \hat{\boldsymbol{I}}_m)'(\boldsymbol{R}_m \hat{\boldsymbol{I}}_m)$, given by the product of the matrix of DFBETAs with its own transpose (Section 6.1.3). Notice that upon re-arranging the expression, the inverse matrices of information on both sides of the residual scores resemble the bread of a sandwich.

In computing the robust sandwich variance estimators of Wei, Lin, and Weissfeld (1989) and also described in Lin and Wei, (1989), it is more convenient to use the DFBETA statistics than the score residuals, since the latter would require further coding. As described in the previous section, DFBETA is an approximation to the exact delta-beta, which is also close to the jackknife. Hence, alternatively, DFBETA could be replaced by these two influences in WLW, which are explored later in this section. In this study, treatment is considered to be the only covariate; hence the parameter of interest $\beta$ is the treatment effect. It is to be recalled from Section 4.4, that the SAS PROC PHREG procedure is computed such that $\beta$ is negative, when experimental is better than control. Therefore, relating this quantity to the notation used for ZW, the treatment effect $\theta$ = -$\beta$. To illustrate WLW, the following SAS codes and their corresponding outputs for the bladder cancer data are given in Figures 6.3 to 6.5.

Figure 6.3: Part 1: SAS codes and output for analysis of the bladder cancer data using WLW.

```
PROC PHREG data = sim_wlw outest = est1 COVS(AGGREGATE);
  MODEL tstop*status(0) = trt1-trt2;
  TREATMENT: TEST trt1,trt2/AVERAGE e;
  OUTPUT out = out1 dfbeta = dt1-dt2;
  STRATA visit;
  ID id;
RUN;
```

The PHREG Procedure

Testing Global Null Hypothesis: BETA=0

| Test | Chi-Square | DF | Pr > ChiSq |
|---|---|---|---|
| Likelihood Ratio | 3.5494 | 2 | 0.1695 |
| Score (Model-Based) | 3.4887 | 2 | 0.1748 |
| Score (Sandwich) | 2.3520 | 2 | 0.3085 |
| Wald (Model-Based) | 3.4228 | 2 | 0.1806 |
| Wald (Sandwich) | 2.3163 | 2 | 0.3141 |

Analysis of Maximum Likelihood Estimates

| Parameter | DF | Parameter Estimate | Standard Error | StdErr Ratio | Chi-Square | Pr > ChiSq | Hazard Ratio |
|---|---|---|---|---|---|---|---|
| trt1 | 1 | -0.36266 | 0.29721 | 0.982 | 1.4889 | 0.2224 | 0.696 |
| trt2 | 1 | -0.55178 | 0.37247 | 0.952 | 2.1946 | 0.1385 | 0.576 |

Linear Coefficients for Test TREATMENT

| Parameter | Row 1 | Row 2 | Average Effect |
|---|---|---|---|
| trt1 | 1 | 0 | 0.79783 |
| trt2 | 0 | 1 | 0.20217 |
| CONSTANT | 0 | 0 | 0.00000 |

Test TREATMENT Results

| Wald Chi-Square | DF | Pr > ChiSq |
|---|---|---|
| 2.3163 | 2 | 0.3141 |

Average Effect for Test TREATMENT

| Estimate | Standard Error | z-Score | Pr > \|z\| |
|---|---|---|---|
| -0.4009 | 0.2913 | -1.3761 | 0.1688 |

The first line in Figure 6.3 is a standard SAS statement for Cox's PH regression model, with COVS specifying a sandwich-based covariance instead of the usual default, a model-based covariance. Consequently, the first part of the output shows a typical PHREG output, with inclusion of the sandwich Wald and Score tests. It is to be noticed that these sandwich quantities give smaller chi-squares compared to their corresponding model-based quantities. For example, the sandwich-based score test gives chi-square value of 2.3520, which is smaller than 3.4887 of the model-based. Consequently, a global null hypothesis test for the former gives a larger p-value (0.3085) compared to the latter (0.1748). The model-based covariance matrix is simply the inverse of the observed information matrix, while the sandwich counterpart is further described later. The parameter estimates for trt1 and trt2 refer to $\hat{\beta}_1$ and $\hat{\beta}_2$ respectively; their hazards ratios are given by $\exp(\hat{\beta}_1)$ and $\exp(\hat{\beta}_2)$, for example, the hazard ratio for trt1 is exp(-0.3627) = 0.696. Notice that the analysis of MLE gives a Std Error Ratio, which is the ratio of the sandwich-based standard error estimate to the model-based standard error estimate.

The robust sandwich estimate is used in the Wald tests for testing the global null hypothesis, and null hypotheses of individual parameters. These tests are so constructed that the former is given by $\hat{\beta}_1^2 / \text{var}(\hat{\beta}_1) + \hat{\beta}_2^2 / \text{var}(\hat{\beta}_2) \sim \chi_2^2$ and the latter $\hat{\beta}^2 / \text{var}(\hat{\beta}) \sim \chi_1^2$. The AGGREGATE option requests a summing up of the score residuals for each distinct ID pattern in the computation of the robust sandwich covariance estimate. Note that the ID statement on the second last line is necessary for this AGGREGATE option to work: taking into account the fact that the recurrent event times are from the same individual. The MODEL statement specifies the variables of survival times and censoring indicators.

The TREATMENT: TEST statement is used to perform the global test of no treatment effect for each tumour recurrence, while the AVERAGE option is specified to estimate the parameter for the common treatment effect ( $\hat{\beta}$ = -0.4009), and the E option displays the optimal weights for the common treatment effect. These weights are displayed as the average effect for trt1 (0.79783) and trt2 (0.20217) above, and are described further in Section 6.3. The last three lines of the above codes are standard commands for SAS to stratify by visit, identify output of DFBETA by ID and RUN the codes. The resulting p-value for the bladder cancer data above is 0.1688 (two-sided), thus there is no significant evidence to reject the null hypothesis of zero common treatment effect. Next, the DFBETAs computed for each individual in the standard PROC PHREG as shown in Part 1 earlier, are now subjected to further processing using the codes displayed in Figure 6.4.

Figure 6.4: Part 2: SAS codes for summing DFBETA from the output of WLW earlier.

```
PROC MEANS data = out1 noprint;
  BY id;
  VAR dt1-dt2;
  OUTPUT out = out2 sum = dt1-dt2;
RUN;
```

To compute the correlation of our interest in this thesis, these DFBETAs termed as *dt*1 and *dt*2 in SAS, are then summed by ID via PROC MEANS procedure, as displayed in Figure 6.4 above. The sums of DFBETAs per individual are then called out for a standard matrix manipulation using PROC IML, to arrive at the variance-covariance matrix and ultimately the correlation. The codes used for this purpose are listed in Figure 6.5.

Figure 6.5: Part 3: SAS codes and output for WLW to compute the estimated covariance and correlation *(Courtesy of Thomas Hamborg)*.

```
PROC IML;
  USE out2;
  READ all var{dt1 dt2} into x;
  v = x` * x;
  RESET noname;
  vname = {"trt1", "trt2"};
  corr = v[1,2]/SQRT(v[1,1]#v[2,2]);
  v[1,1] = 1/v[1,1];
  v[2,2] = 1/v[2,2];
  v[1,2] = v[1,2] # v[1,1] # v[2,2];
  v[2,1] = v[1,2];
  CALL SYMPUT('wlwCorr',LEFT(CHAR(corr,6,4)));
  PRINT, "estimated covariance matrix (WLW)", ,
        v[colname = vname rowname = vname format = 10.5];
  PRINT, "estimated correlation",  corr[colname="corr"];
  CREATE rcov from v[colname = vname rowname = vname];
  APPEND from v[rowname = vname];
  CLOSE rcov;
  QUIT;
RUN;
```

```
estimated covariance matrix (WLW)

                trt1        trt2

trt1        11.32076     5.81223
trt2         5.81223     7.20805

estimated correlation
                corr

             0.6434218
```

It is to be noticed that the variable coded as $x$ above is the DFBETA, labelled as $dt1$ and $dt2$ for the 1st and 2nd events respectively. The computation of "$v = x` * x$" yields the matrix product

$$\begin{pmatrix} dt_{11} & \cdots & dt_{1n} \\ & & \\ dt_{21} & \cdots & dt_{2n} \end{pmatrix} \begin{pmatrix} dt_{11} & dt_{21} \\ \vdots & \vdots \\ dt_{1n} & dt_{2n} \end{pmatrix} = \begin{pmatrix} \hat{\boldsymbol{Y}}_{11} & \hat{\boldsymbol{Y}}_{12} \\ \hat{\boldsymbol{Y}}_{21} & \hat{\boldsymbol{Y}}_{22} \end{pmatrix} \tag{6.25}$$

where $\hat{\boldsymbol{Y}}_{11}$ and $\hat{\boldsymbol{Y}}_{22}$ are the variances of the estimated treatment advantages $\hat{\beta}_1$ and $\hat{\beta}_2$ respectively, and $\hat{\boldsymbol{Y}}_{12} = \hat{\boldsymbol{Y}}_{21}$ = covariance between the two estimates, as shown in Figure 6.5. Taking an inverse square root of $\hat{\boldsymbol{Y}}_{11}$ (11.32) gives exactly the standard error for trt1 (0.2972) in Figure 6.3 and similarly for $\hat{\boldsymbol{Y}}_{22}$ which concerns the 2nd event, trt2. The written SAS program gives exactly the same output as the PHREG codes, but includes the covariance and the correlation, which is estimated by

$$\text{corr}(\hat{\beta}_1, \hat{\beta}_2) = \frac{\text{cov}((\hat{\beta}_1, \hat{\beta}_2)}{\sqrt{\hat{\boldsymbol{Y}}_{11}\hat{\boldsymbol{Y}}_{22}}} = \frac{\hat{\boldsymbol{Y}}_{12,21}}{\sqrt{\hat{\boldsymbol{Y}}_{11}\hat{\boldsymbol{Y}}_{22}}}. \tag{6.26}$$

Part 3 coding shows how this parameter is computed in various steps described above. A complete set of SAS codes to run this program for analysis of real data is attached in Appendix B.

As mentioned earlier, the DFBETA was shown to slightly underestimate the influences given by the exact delta-beta and the jackknife. How would such differences affect the estimated correlation? To answer the question, an investigation is carried out using the bladder cancer data. As already described in Sections 6.1.2 and 6.1.3, computation of both the jackknife and exact delta-beta requires that each of the 86 observations be omitted one at a time and the ($n$ - 1) estimates, $\hat{\beta}_{(i)}$ be found.

Specific codes are needed to exclude each observation in turn and each of the 86 estimates can be achieved by using SAS codes in Figure 6.3, but removing the "COVS (AGGREGATE)" option and the "ID id" accordingly. The only difference is that the jackknife compares $\hat{\beta}_{(i)}$ with the average of the ($n$ - 1) estimates, $\hat{\beta}_{(\bullet)}$, whereas the exact delta-beta compares $\hat{\beta}_{(i)}$ with the estimate from the complete data set $\hat{\beta}$. Using these different values for the influence, the estimates of the two treatment advantages and the covariance between them can be computed using similar codes in Figure 6.5, but replacing the quantity DFBETA by that of the exact delta-beta and jackknife influences accordingly. A complete set of SAS codes to perform this comparison is attached in Appendix C. The results for the bladder cancer data using these three influences are presented for total time and gap time in Table 6.1.

Table 6.1: Various estimated quantities for the bladder cancer data using the methods of jackknife, exact delta-beta and WLW for **total time** (TT).

*(N.B. Multi-place decimals are intentional to show the minute difference)*

| Parameter (s.e.) | Jackknife | Exact Delta-Beta | WLW (DFBETA) |
|---|---|---|---|
| $\hat{\theta}_1$ | 0.3627418 | 0.3626606 | 0.3626606 |
| | (0.305781) | (0.3075757) | (0.2972092) |
| $\hat{\theta}_2$ | 0.5520362 | 0.5517842 | 0.5517842 |
| | (0.388295) | (0.3905788) | (0.3724699) |
| $\text{cov}(\hat{\theta}_1,\hat{\theta}_2)$ | 0.1026257 | 0.1037313 | 0.0712277 |
| $\text{corr}(\hat{\theta}_1,\hat{\theta}_2)$ | 0.6424595 | 0.6424607 | 0.6434218 |

From Table 6.1, the jackknife method gives slightly higher estimates of individual treatment advantage when compared to those for the exact and WLW. As anticipated, the estimates of variance using the exact and jackknife methods are very close to each other, but much higher than those of WLW which uses DFBETA; note

similar observation for the covariance. The correlation values for all methods are comparable to each other and exact to one decimal point (0.6) for TT.

Table 6.2: Various estimated quantities for the bladder cancer data using the methods of jackknife, exact delta-beta and WLW for **gap time** (GT).

| Parameter (s.e.) | Jackknife | Exact Delta-Beta | WLW (DFBETA) |
|---|---|---|---|
| $\hat{\theta}_1$ | 0.3627418 | 0.3626606 | 0.3626606 |
| | (0.305781) | (0.3075757) | (0.2972092) |
| $\hat{\theta}_2$ | 0.17403 | 0.17381 | 0.17381 |
| | (0.4026626) | (0.4026677) | (0.37669) |
| $\text{cov}(\hat{\theta}_1, \hat{\theta}_2)$ | 0.0681932 | 0.0678577 | -0.0055538 |
| $\text{corr}(\hat{\theta}_1, \hat{\theta}_2)$ | -0.037425 | -0.037412 | -0.049606 |

Meanwhile, Table 6.2 shows that the estimates from all three methods have similar values, the jackknife giving the highest estimate of covariance. However, all the covariances and the correlations are essentially zero for the bladder cancer data using the GT model. This small evidence may support the justification for using the DFBETA in practical applications.

### 6.3. Estimation of an Overall Treatment Advantage: WLW

As described in Section 4.4 earlier, a key objective of a clinical trial is to compare the overall treatment advantage, say $\beta$ between experimental and control treatments. The assumption of equal treatment advantage $\beta_1 = \beta_2 = \beta$ is always valid under the null hypothesis since the common value is zero. However, under the alternative where $\beta_1 = \beta_2 = \beta$, but $\beta \neq 0$, it is natural to estimate $\beta$ by a linear combination of $\hat{\beta}$, with the relationship $\hat{\beta} = w_1\hat{\beta}_1 + w_2\hat{\beta}_2$. The choice of weighting which yields the smallest asymptotic variance among all of the linear estimators (Wei & Johnson, 1985) is

adopted in WLW. Solving for the derivative of $\text{var}(\hat{\beta})$ equals to zero, the optimal weighting for WLW is

$$(w_1, w_2)_{WLW} = \left( \frac{y_2 - c_w}{y_1 + y_2 - 2c_w}, \frac{y_1 - c_w}{y_1 + y_2 - 2c_w} \right) \tag{6.27}$$

where $y_1 = \text{var}(\hat{\beta}_1)$, $y_2 = \text{var}(\hat{\beta}_2)$, and $c_w = \text{cov}(\hat{\beta}_1, \hat{\beta}_2)$. This equation is similar to the optimal weighting used for ZW in equation (4.10), except that the variances and covariances relate to Z, in the latter. As shown in the previous section, this optimal weighting is specified by including COVS (AGGREGATE) within the options of PROC PHREG statement. Without specifying the option "AGGREGATE", SAS would treat the recurrent event times from the same individual as if they were from different individuals. Therefore, the covariance in equation (6.27) is taken as zero, and consequently giving a different overall treatment advantage. In fact, when $c_w = 0$, the overall treatment advantage is found to be exactly the same value as our estimate, $\hat{\theta}^* = Z^*/V^*$, but of opposite sign ( $\theta = -\beta$ ) as already defined in Chapter 4. The effects of both types of weighting on the estimated overall treatment advantage are illustrated using real data in Section 6.5.

The option COVS (AGGREGATE) gives the robust sandwich estimate of the covariance matrix, and the score residuals used in computing the middle part of the sandwich estimate are aggregated over identical ID values. This is a means of empirical adjustment for correlation within the data sets without specifying any dependence structure within the model itself. As shown by the standard error ratio in the previous section, the robust sandwich gives a smaller estimate of standard error compared to the model-based. This robust variance estimate was proposed by Lin and Wei (1989) and Reid and Crèpeau(1985), and adopted by WLW (Wei et al., 1989).

As described earlier, the optimal weighting serves to correct for the dependence structure between the multiple events, hence giving the smallest variance. However, the optimal weighting which weights the marginal estimates by the inverse of the covariance matrix has been criticized for its peculiarity of yielding undesirable negative weighting (Pocock, 1997; Tang et al., 1993). This situation is evident in the case of the bladder cancer data for the third recurrence (Wei et al., 1989): the optimal weight for estimating the parameter of the common treatment effect is -0.07547 as cited in SAS 9.2 example. This negative weighting scenario arises when the covariance is larger than the variance as permissible in equation (6.25). Saville et al. (2010) claimed that use of equal weighting for WLW provided better performance than the weighting used by Wei et al. (1989); the former is hence recommended. Nevertheless, our method adopts the latter as it yields the smallest variance estimate and consistent with our comparator, WLW.

Although WLW was not intended exclusively for recurrent events, following its first appearance in illustrating recurrent failures, it has been a popular choice for such cases. There is a mixed opinion among authors as some recommend it while others criticize. For example, Lipschutz & Snapinn (1997), Kelly & Lim (2000), and Metcalfe & Thompson (2006) contended that WLW overestimates the treatment advantage, largely owing to the carry-over effect when using total time model. However, Metcalfe & Thompson (2007) later reviewed and evaluated the semi-parametric model method in the analysis of recurrent event data and concluded that the application of this method to recurrent event data is justified. None of the above mentioned authors employed WLW using the gap time. Wei et al. (1989) have used gap time in their illustration, but no substantial discussion followed. Here, we compare and contrast WLW with ZW for both total time and gap time risk intervals.

## 6.4. Theoretical Comparison: ZW vs. WLW

From earlier sections, the coefficient of covariates is denoted by $\beta$. Since this study considers treatment as the only covariate, the focus is on treatment advantage. It is recalled from Section 4.4 that $\theta$ equals minus $\beta$ and the former notation is now adopted throughout the remainder of this thesis. Restating, WLW provides estimates which asymptotically follow the normal distribution, $\hat{\boldsymbol{\theta}} \sim N(\boldsymbol{\theta}, \boldsymbol{Y})$ where $\boldsymbol{Y}$ is the robust variance-covariance matrix.

As described in Chapter 4, our method is a direct approach which conditions on the successive risk sets and reproduces the familiar form of logrank variance. It capitalizes on the efficient score, $Z$ which is essentially the cumulative treatment advantage. Meanwhile, its null variance is given by the Fisher's information $V$. Asymptotically, $Z$ follows a normal distribution, $Z \sim N(\theta V, V)$. Both quantities $Z$ and $V$ are calculated as interval-censored logrank statistics and their null variances respectively, with unadjusted standard errors. The null hypothesis of no treatment difference is tested using $Z_1 + Z_2$ and its null variance $V_1 + V_2 + 2C_{12}$ where $C_{12}$ is the estimated covariance as derived in Chapter 4.

The WLW procedure used to estimate the variance is applied to the whole data set hence termed population-average method. On the other hand, ZW computes the variance for each set of 2 x 2 tables in turn, and sums up the covariance. In WLW, the Wald test statistic is given by $\hat{\theta} / s.e.(\hat{\theta})$ with an adjusted standard error. Unconditionally (on the risk sets), the score residuals $R$ defined in equation (6.23) are independent, identically distributed random variables, obtained by evaluating the changes in the score vector $\boldsymbol{U}$ with respect to changes in weighting of the $i^{th}$

individual, as shown in Section 6.1.3, resulting in equation (6.22). In WLW, the covariance is evaluated and the parameters are estimated under the alternative hypothesis (Wald test), while those for the combined score test approach of ZW are accomplished under the null.

Consider a simple case where we have subjects randomized to experimental and control groups, with two correlated event types, $m = 1, 2$, and we want to find the treatment advantage $\theta$. As shown in Chapter 4, the treatment advantage can be estimated by $\hat{\theta} = Z / V$, and therefore $\text{var}(\hat{\theta}) = 1/V$ since $\text{var}(Z) \approx V$. The covariance between treatment advantage for the 1st event and that for the 2nd event can be expressed as

$$\text{cov}(\hat{\theta}_1, \hat{\theta}_2) = E\left(\frac{Z_1 Z_2}{V_1 V_2}\right) - E\left(\frac{Z_1}{V_1}\right) E\left(\frac{Z_2}{V_2}\right) = \frac{\text{cov}(Z_1, Z_2)}{V_1 V_2} \tag{6.28}$$

The relationship between the two covariances in equation (6.28) enables the derivation of a corresponding relationship between the correlations:

$$\text{corr}(\hat{\theta}_1, \hat{\theta}_2) = \frac{\text{cov}(\hat{\theta}_1, \hat{\theta}_2)}{\sqrt{\text{var}(\hat{\theta}_1)\,\text{var}(\hat{\theta}_2)}} = \frac{\text{cov}(Z_1, Z_2)}{\sqrt{V_1 V_2}} = \text{corr}(Z_1, Z_2). \tag{6.29}$$

As shown in equation (6.29), the correlation between the treatment advantages of two dependent events is theoretically equivalent to the correlation between two score statistics of those events. A similar relationship was also reported by Whitehead et al. (2010), in deriving a global score test for binary and ordinal endpoints. Despite the key difference in approach to estimating the covariance, both WLW and ZW should yield approximately the same correlation estimate: which is illustrated in the next section using real data. This transformation is made possible by the very definition of the score statistic being the cumulative treatment advantage (Section

1.4). Nevertheless, adopting WLW in an analysis of interval-censored data may yield a lower power of test than that for continuous data. In the following sections, both ZW and WLW are employed for estimating the treatment advantages and constructing tests when dealing with real and simulated data sets respectively.

On the basis of special characteristics of recurrent events described in Section 4.6, we shall compare the choice of risk intervals and stratification for both WLW and ZW. The former is applicable to both total and gap time, while the latter has proved to work well when using total time model. Overestimation of the correlation pertaining to gap time is further investigated in Section 6.6.6. Both methods imply that the baseline hazards are event-specific, while ZW can further model the baseline hazard to be interval-specific too, owing to its interval-censored nature.

Although both methods are developed based on the proportional hazards assumption, they differ slightly in detail. While WLW allows for a proportional hazards ratio specific to the $m^{th}$ event, ZW offers more flexibility in permitting a proportional hazards ratio within each interval, as illustrated in Section 3.1.4. Despite suitability in the analysis of censored multiple endpoints, WLW offers no insight into the interrelationships among event times. It provides inference as to the population-average covariate effects on event times, but does not address the question as to how prior events affect the risk of having future recurrences. Wei and Glidden (1997) suggested that the AG, PWP and frailty models may be appropriate in affording some answers to that question

## 6.5. Applications to Real Data: ZW vs. WLW

To illustrate the use of WLW, we employ the real data sets which have been used for ZW in Chapter 4. As demonstrated earlier, the correlation between two score statistics closely approximates the correlation between the two treatment advantages given by ZW and WLW respectively. Thus, these two correlations are interchangeable for the remainder of this thesis. The estimated correlation and common treatment advantage are compared for three real data sets earlier described in Chapter 4: (i) bladder cancer tumour recurrence, (ii) paired hips replacement revision and (iii) cancer data.

Note that all analyses using ZW were based on 5 and 10 intervals of event times $T_1$ and $T_2$, as commented upon in Chapter 5; the first data set is presented with more parameters for illustration purposes, while those for the other two are summarized using the key parameters only. Similar to Chapter 5, the null hypothesis test is conducted to detect significance at 2.5% (one-sided), so that the two sets of results given by ZW and WLW can be compared accordingly.

### 6.5.1. Recurrent Events: Bladder Cancer Data

For bivariate survival data, WLW computation dictates that two observations be created for each patient: one row of data for each of the two events. For two events of tumour recurrence (refer Section 4.6.2), analyses are carried out using total time (TT) and gap time (GT). WLW is applied to the raw data set as well as to its interval-censored version (as was used for ZW), for investigation purposes. The standard overall treatment advantage, $\hat{\theta}^*$, is also given in this section, for comparison with the optimal overall treatment advantage, $\hat{\theta}$. The results are summarized in Table 6.3 and Table 6.4 for TT and GT respectively.

Table 6.3: Results for the bladder cancer data using total time (TT) for WLW and ZW.

*(ZW results correspond to Table 4.12 and WLW to Table 6.1 earlier).*

| Parameter (s.e.) | WLW TT | WLW TT (5 intervals) | WLW TT (10 intervals) | ZW TT (5 intervals) | ZW TT (10 intervals) |
|---|---|---|---|---|---|
| $\hat{\theta}_1$ | 0.363 | 0.325 | 0.318 | 0.462 | 0.378 |
| | (0.297) | (0.278) | (0.286) | (0.293) | (0.291) |
| $\hat{\theta}_2$ | 0.552 | 0.567 | 0.550 | 0.581 | 0.537 |
| | (0.373) | (0.360) | (0.370) | (0.370) | (0.369) |
| $\hat{\theta}*$ | 0.436 | 0.415 | 0.405 | 0.508 | 0.439 |
| | (0.296) | (0.277) | (0.287) | (0.290) | (0.289) |
| $\hat{\theta}$ | 0.401 | 0.371 | 0.359 | 0.488 | 0.411 |
| | (0.232) | (0.272) | (0.281) | (0.285) | (0.284) |
| p-value | 0.169 | 0.173 | 0.202 | 0.087 | 0.148 |
| $\text{cov}(Z_1, Z_2)$ | 5.812 | 6.101 | 5.995 | 5.659 | 5.771 |
| $\text{corr}(Z_1, Z_2)$ | 0.643 | 0.610 | 0.632 | 0.611 | 0.619 |
| $Z_1$ | 4.1 | 4.2 | 3.9 | 5.4 | 4.5 |
| $Z_2$ | 4.0 | 4.4 | 4.0 | 4.3 | 4.0 |
| $Z*$ | 5.0 | 5.4 | 4.9 | 6.1 | 5.3 |
| $V_1$ | 11.3 | 13.0 | 12.3 | 11.7 | 11.8 |
| $V_2$ | 7.2 | 7.7 | 7.3 | 7.3 | 7.4 |
| $V*$ | 11.4 | 13.0 | 12.2 | 11.9 | 12.0 |

As shown in Table 6.3, the p-values are in agreement which suggest no significant evidence to reject the null hypothesis. It is observed that the finer intervals give smaller estimates of treatment advantage and the optimal $\hat{\theta}$ is smaller than the standard $\hat{\theta}*$. WLW seems to yield a consistent $\hat{\theta}$ (0.4) even when intervals are used, whereas ZW varies by about 0.1. The standard error values are similar for the two methods, except that for the optimal $\hat{\theta}$ (TT), where ZW yields higher values of about 0.29 compared to 0.23 when using WLW (without intervals). The covariances are close to one another and the estimated correlations are consistent at 0.6. Upon

comparing all the estimates from ZW with those from WLW, the use of 10 intervals for the former seems to give the values closest to the latter. This reaffirms the finding in the simulation study of recurrent TT events earlier (Section 5.2.5). It is also worth noting that the values of $Z$ and $V$ for both methods seem to be in good agreement in all cases, hence they are not presented for the subsequent data sets.

Table 6.4: Results for the bladder cancer data using gap time (GT) for WLW and ZW *(ZW results correspond to Table 4.13 and WLW to Table 6.2 earlier).*

| Parameter (s.e.) | WLW GT | WLW GT (5 intervals) | WLW GT (10 intervals) | ZW GT (5 intervals) | ZW GT (10 intervals) |
|---|---|---|---|---|---|
| $\hat{\theta}_1$ | 0.363 | 0.325 | 0.318 | 0.462 | 0.378 |
| | (0.297) | (0.278) | (0.286) | (0.293) | (0.291) |
| $\hat{\theta}_2$ | 0.174 | 0.281 | 0.264 | 0.193 | 0.161 |
| | (0.377) | (0.349) | (0.360) | (0.376) | (0.375) |
| $\hat{\theta}*$ | 0.290 | 0.308 | 0.297 | 0.360 | 0.297 |
| | (0.228) | (0.221) | (0.228) | (0.219) | (0.224) |
| $\hat{\theta}$ | 0.289 | 0.308 | 0.297 | 0.358 | 0.295 |
| | (0.233) | (0.221) | (0.228) | (0.219) | (0.224) |
| p-value | 0.204 | 0.163 | 0.193 | 0.103 | 0.186 |
| cov($Z_1$, $Z_2$) | -0.443 | 0.358 | 0.417 | -0.919 | -0.488 |
| corr($Z_1$, $Z_2$) | -0.050 | 0.035 | 0.043 | -0.101 | -0.053 |
| $Z_1$ | 4.1 | 4.2 | 3.9 | 5.4 | 4.5 |
| $Z_2$ | 1.2 | 2.3 | 2.0 | 1.1 | 1.1 |
| $Z*$ | 5.6 | 6.2 | 5.7 | 5.8 | 5.9 |
| $V_1$ | 11.3 | 13.0 | 12.3 | 11.7 | 11.8 |
| $V_2$ | 7.1 | 8.2 | 7.7 | 7.1 | 7.1 |
| $V*$ | 19.3 | 20.5 | 19.2 | 19.7 | 20.0 |

Meanwhile, for the gap time (Table 6.4), the p-values consistently indicate no significant evidence to reject the null hypothesis. It is to be noted that $\theta$ here has a different meaning compared to that of the total time (Section 4.6.1). The score statistic when using GT is considerably smaller than TT as it is indeed measuring a cumulative treatment effect using a different time scale. For example, when using WLW, $Z_2 = 1.2$ for GT compared to $Z_2 = 4.0$ for TT (Table 6.3). Consequently, the estimated treatment effect for $T_2$ for GT is also smaller compared to those for TT earlier. This is expected owing to the carry-over effect of $T_1$ unto $T_2$ when using the latter time.

Both methods give very small correlations, which suggest that there is no evidence to show that the time from randomization to first recurrence $(0, T_1)$ correlates with the time between the first and second recurrence $(T_1, T_2)$. Earlier, the analysis using total time has shown some evidence that the time from randomization to the first event $(0, T_1)$ is highly correlated with the time from randomization to the second recurrence $(0, T_1 + T_{2G})$: corr $(Z_1, Z_2) = 0.6$. This high correlation when using total time is largely attributable to the dominating $T_1$. As explained above, these two models have different interpretation as their times are measured differently. The model of choice hence depends on the questions of interest, as already discussed in Section 4.6.1.

### 6.5.2. **Paired Organs: Hip Replacement Revision**

A detailed description of this data set can be seen in Section 4.5.1: it is to be remembered that we consider the different types of cup positioning with respect to the acetabulum as different "treatments". Thus, it is anticipated that the treatment advantage is zero as both "treatments" are supposed to be ethically equivalent. The results for both ZW and WLW are summarized in Table 6.5.

Table 6.5: Results for the hip replacement revision data using WLW and ZW

*(ZW results correspond to Table 4.5).*

| Parameter (s.e) | WLW | ZW (5 interval) | ZW (10 interval) |
|---|---|---|---|
| $\hat{\theta}_1$ | -0.016 | -0.166 | -0.143 |
| | (0.223) | (0.231) | (0.230) |
| $\hat{\theta}_2$ | 0.214 | 0.062 | 0.144 |
| | (0.257) | (0.260) | (0.254) |
| $\hat{\theta}$ | 0.072 | -0.073 | -0.022 |
| | (0.199) | (0.201) | (0.199) |
| p-value | 0.360 | 0.643 | 0.544 |
| $\text{cov}(Z_1, Z_2)$ | 7.187 | 5.947 | 6.235 |
| $\text{corr}(Z_1, Z_2)$ | 0.412 | 0.357 | 0.364 |

Table 6.5 shows that the p-values are qualitatively in agreement: there is no treatment advantage, as anticipated. The setting of 10 intervals gives the estimates closest to those for WLW. This is consistent with an earlier finding in the simulation study for paired organs where 10 intervals gave better accuracy based on the key performance measures (Section 5.2.2). The methods give similar correlations near 0.4 (that for ZW is slightly lower), indicating that the treatment advantages for the two hips are moderately correlated.

### 6.5.3. **Generally Related Indicators and PFS: Cancer Data**

The cancer data used here can be referred to in Section 4.5.2. The correlated events of disease progression and death are analyzed in two ways: (i) taking each as a separate event with its own censoring variable, labelled as an indicator (Section 4.3.4) and (ii) considering a progression-free survival whereby either a progression or death is considered as an event, labelled as a PFS. Output from WLW for each case is now compared to that earlier obtained using ZW, in Table 6.6.

Table 6.6: Results for the cancer data using WLW and ZW (5 intervals)

*(ZW results correspond to those in Table 4.8 and Table 4.9).*

| Parameter (s.e.) | WLW (Indicators) | ZW (Indicators) | WLW (PFS) | ZW (PFS) |
|---|---|---|---|---|
| $\hat{\theta}_1$ | 0.022 | -0.014 | 0.090 | 0.219 |
| | (0.122) | (0.125) | (0.112) | (0.120) |
| $\hat{\theta}_2$ | 0.133 | 0.189 | 0.133 | 0.189 |
| | (0.121) | (0.127) | (0.121) | (0.127) |
| $\hat{\theta}$ | 0.078 | 0.086 | 0.110 | 0.206 |
| | (0.086) | (0.107) | (0.082) | (0.108) |
| p-value | 0.363 | 0.422 | 0.182 | 0.056 |
| $\text{cov}(Z_1, Z_2)$ | 27.790 | 27.294 | 43.450 | 34.802 |
| $\text{corr}(Z_1, Z_2)$ | 0.412 | 0.433 | 0.591 | 0.530 |

When considering the case of related indicators, all parameter estimates using both methods are in good agreement. The p-values conclude that there is no significant evidence to reject the $H_0$, zero treatment advantage. The correlation estimates are accurately matched at 0.4 (ZW slightly higher). Meanwhile, the results from PFS analysis in the last two columns show that ZW gives higher estimates for the treatment effect when comparedwith WLW. Nevertheless, the p-values are leading towards the same conclusion not to reject $H_0$ at the 2.5% (one-sided) level of

significance. Both methods give higher correlation when evaluating either progression or death as an event than those in the indicators case, due to the imposed dependence on their censoring variables: for PFS, 0.59 (WLW) and 0.53 (ZW). It is worth noting that the covariance for WLW is higher than that for ZW whereas the variances of $Z_1$ (given by $V_1$) are about the same.

In summary, the above applications to real data seem to suggest that our method is comparable to WLW. Whether or not these two methods are equally accurate is answered in the subsequent sections.

### 6.6. Simulation and Results

Thus far, the analyses of ZW and WLW have been justified by theory and illustrated by practical application to real data sets. The accuracy of the former has been evaluated in Chapter 5. Now, the latter is also applied to the same simulated data sets as have earlier been analyzed using ZW. It is to be noted that WLW requires two sets of observations for each invididual, as explained in Section 6.5.1.The same key measures as in Section 5.2 are used: (i) type I error, conventionally labelled as $\alpha$, (ii) power (1 - $\beta$), and (iii) ratio of $\rho_{(est)}$ to $\rho_{(sample)}$. It is recalled that each setting covers both the null and alternative, with sample size $n$ = 1000, and is based on the overall results in Section 5.2.8, the results for ZW when using 5 intervals being compared with the simulation results for WLW in this chapter.

In Chapter 5, the simulation was formulated to give a theoretical power which varies with $d$ under the alternative of $\theta = \theta_R$ where $\theta_R$ corresponds to $n$ = 1,000 for ZW specifically. The reader is reminded that the setting of $\theta_R$ was based on the log hazard ratio within subject, $\theta_W$, whereas the simulation design yields an output of $\theta_B$, which is the log hazard ratio between subjects. Therefore, each power is comparable to its

corresponding theoretical power, TP. For a detailed description of the theoretical power the reader is referred to Section 5.1. To compare the performance of ZW with that of WLW, the same data sets are now used for WLW, without recalculation of $\theta_R$. Also, as in the earlier setting for ZW, the type I error rate is set at 0.025 (one-sided).and the level of accuracy sought is 0.025 ± 0.003 ($N$ = 10,000 replicates). The outputs for the cases of complete (uncensored), paired, indicators, PFS, recurrent TT and recurrent GT, are tabulated and discussed in turn.

### 6.6.1. **Complete or Uncensored**

With reference to Section 5.1, the multiplier constant $d$ is varied in such a way that $d \in \{1,5,10\}$ corresponds to low, medium and high correlation respectively. The target treatment effect, $\theta$ is determined for each correlation as governed by $d$. For each varying correlation, 10,000 simulations in total were run under each hypothesis and the mean value of each key performance measure for the uncensored case is presented in Table 6.7. It is to be noted that the correlation ratios are later summarized in Figure 6.6 (a).

Table 6.7: Simulation results for the complete case under the null and alternative using WLW and ZW *(ZW results correspond to Table 5.4).*

| WLW: $H_0$ | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $d$ | $\theta$ | $\hat{\theta}_1$ | $\hat{\theta}_2$ | $\hat{\theta}$ | $\alpha_1$ | $\alpha_2$ | $\alpha_{12}$ | N/A | $\rho_{(est)}$ | $\rho_{(sample)}$ |
| 1 | 0.000 | 0.000 | 0.000 | 0.000 | 0.025 | 0.021 | 0.023 | N/A | 0.023 | 0.027 |
| 5 | 0.000 | -0.001 | -0.001 | -0.001 | 0.025 | 0.026 | 0.026 | N/A | 0.410 | 0.419 |
| 10 | 0.000 | -0.001 | -0.001 | -0.001 | 0.025 | 0.025 | 0.025 | N/A | 0.708 | 0.710 |
| WLW: $H_1$ | | | | | | | | | | |
| $d$ | $\theta$ | $\hat{\theta}_1$ | $\hat{\theta}_2$ | $\hat{\theta}$ | 1-$\beta_1$ | 1-$\beta_2$ | 1-$\beta_{12}$ | TP | $\rho_{(est)}$ | $\rho_{(sample)}$ |
| 1 | 0.174 | 0.171 | 0.171 | 0.171 | 0.68 | 0.68 | 0.92 | 0.89 | 0.023 | 0.027 |
| 5 | 0.209 | 0.139 | 0.139 | 0.139 | 0.51 | 0.50 | 0.65 | 0.58 | 0.410 | 0.420 |
| 10 | 0.221 | 0.089 | 0.088 | 0.088 | 0.24 | 0.23 | 0.27 | 0.25 | 0.708 | 0.709 |
| ZW: $H_0$ @ 5 intervals | | | | | | | | | | |
| $d$ | $\theta$ | $\hat{\theta}_1$ | $\hat{\theta}_2$ | $\hat{\theta}$ | $\alpha_1$ | $\alpha_2$ | $\alpha_{12}$ | N/A | $\rho_{(est)}$ | $\rho_{(sample)}$ |
| 1 | 0.000 | 0.000 | 0.000 | 0.000 | 0.023 | 0.021 | 0.023 | N/A | 0.023 | 0.026 |
| 5 | 0.000 | -0.001 | -0.001 | -0.001 | 0.024 | 0.025 | 0.026 | N/A | 0.382 | 0.412 |
| 10 | 0.000 | -0.001 | -0.001 | -0.001 | 0.024 | 0.027 | 0.025 | N/A | 0.664 | 0.693 |
| ZW: $H_1$ @ 5 intervals | | | | | | | | | | |
| $d$ | $\theta$ | $\hat{\theta}_1$ | $\hat{\theta}_2$ | $\hat{\theta}$ | 1-$\beta_1$ | 1-$\beta_2$ | 1-$\beta_{12}$ | TP | $\rho_{(est)}$ | $\rho_{(sample)}$ |
| 1 | 0.174 | 0.171 | 0.171 | 0.171 | 0.67 | 0.67 | 0.92 | 0.89 | 0.028 | 0.027 |
| 5 | 0.209 | 0.139 | 0.139 | 0.139 | 0.50 | 0.49 | 0.65 | 0.58 | 0.384 | 0.412 |
| 10 | 0.221 | 0.089 | 0.088 | 0.088 | 0.24 | 0.23 | 0.28 | 0.25 | 0.664 | 0.694 |

As shown in the above table, under the null, the estimates of treatment advantage are essentially zero and the type I error rates are well within the 95% probability interval (0.022, 0.028). The power exceeds its theoretical value (TP) and the estimated correlation is very close to its sample value for each setting. It is to be noticed that the assumption of equal $\theta$ is satisfactorily met, hence a clear advantage for the global test. These results show similarities to those for ZW and so do the reasons (Section 5.2.1).

### 6.6.2. **Paired Organs**

The simulation results for the paired organs using WLW are summarized in Table 6.8, while the correlation ratios for comparison with ZW are presented in Figure 6.6 (b).

Table 6.8: Simulation results for the paired case under the null and alternative using WLW and ZW *(ZW results correspond to Table 5.5).*

| WLW: $H_0$ | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $d$ | $\theta$ | $\hat{\theta}_1$ | $\hat{\theta}_2$ | $\hat{\theta}$ | $\alpha_1$ | $\alpha_2$ | $\alpha_{12}$ | N/A | $\rho_{(est)}$ | $\rho_{(sample)}$ |
| 1 | 0.000 | 0.000 | 0.000 | 0.000 | 0.024 | 0.023 | 0.027 | N/A | 0.027 | 0.018 |
| 5 | 0.000 | 0.000 | 0.000 | 0.000 | **0.028** | 0.024 | 0.026 | N/A | 0.431 | 0.434 |
| 10 | 0.000 | 0.001 | 0.001 | 0.001 | 0.027 | 0.026 | 0.025 | N/A | 0.688 | 0.681 |
| WLW: $H_1$ | | | | | | | | | | |
| $d$ | $\theta$ | $\hat{\theta}_1$ | $\hat{\theta}_2$ | $\hat{\theta}$ | 1-$\beta_1$ | 1-$\beta_2$ | 1-$\beta_{12}$ | TP | $\rho_{(est)}$ | $\rho_{(sample)}$ |
| 1 | 0.241 | 0.236 | 0.238 | 0.237 | 0.65 | 0.66 | 0.91 | 0.89 | 0.027 | 0.016 |
| 5 | 0.287 | 0.191 | 0.191 | 0.191 | 0.48 | 0.48 | 0.61 | 0.58 | 0.430 | 0.433 |
| 10 | 0.305 | 0.119 | 0.119 | 0.119 | 0.23 | 0.22 | 0.24 | 0.24 | 0.705 | 0.700 |
| ZW: $H_0$ @ 5 intervals | | | | | | | | | | |
| $d$ | $\theta$ | $\hat{\theta}_1$ | $\hat{\theta}_2$ | $\hat{\theta}$ | $\alpha_1$ | $\alpha_2$ | $\alpha_{12}$ | N/A | $\rho_{(est)}$ | $\rho_{(sample)}$ |
| 1 | 0.000 | 0.000 | 0.000 | 0.000 | 0.025 | 0.023 | 0.026 | N/A | 0.035 | 0.027 |
| 5 | 0.000 | 0.001 | 0.000 | 0.000 | **0.028** | 0.025 | **0.028** | N/A | 0.421 | 0.437 |
| 10 | 0.000 | 0.001 | 0.002 | 0.001 | 0.027 | 0.025 | 0.026 | N/A | 0.674 | 0.685 |
| ZW: $H_1$ @ 5 intervals | | | | | | | | | | |
| $d$ | $\theta$ | $\hat{\theta}_1$ | $\hat{\theta}_2$ | $\hat{\theta}$ | 1-$\beta_1$ | 1-$\beta_2$ | 1-$\beta_{12}$ | TP | $\rho_{(est)}$ | $\rho_{(sample)}$ |
| 1 | 0.241 | 0.235 | 0.236 | 0.235 | 0.65 | 0.65 | 0.90 | 0.89 | 0.041 | 0.022 |
| 5 | 0.287 | 0.190 | 0.190 | 0.190 | 0.47 | 0.48 | 0.61 | 0.57 | 0.422 | 0.437 |
| 10 | 0.305 | 0.120 | 0.120 | 0.120 | 0.23 | 0.22 | 0.25 | 0.25 | 0.692 | 0.704 |

The results for the paired organs are very similar to those for the complete case, except that the power is slightly lower because, probably, of the censoring in the former, and that the type I error ($\alpha_1$ = 0.028 at $d$ = 5) is slightly inflated. Again, a similar trend to that for ZW is observed, including the slightly higher correlation ($d$ = 10) under the alternative compared to the null.

### 6.6.3. **Related Indicators**

Resembling the analysis of cancer data comprising TTP and OS (Section 4.5.2), the simulation results for these related indicators are presented in Table 6.9 and Figure 6.6 (c).

Table 6.9: Simulation results for the indicators under the null and alternative using WLW and ZW *(ZW results correspond to Table 5.6).*

| WLW: $H_0$ | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $d$ | $\theta$ | $\hat{\theta}_1$ | $\hat{\theta}_2$ | $\hat{\theta}$ | $\alpha_1$ | $\alpha_2$ | $\alpha_{12}$ | N/A | $\rho_{(est)}$ | $\rho_{(sample)}$ |
| 1 | 0.000 | -0.001 | -0.001 | -0.001 | **0.020** | 0.023 | 0.023 | N/A | 0.023 | 0.034 |
| 5 | 0.000 | -0.001 | -0.001 | -0.001 | **0.021** | 0.027 | 0.026 | N/A | 0.409 | 0.413 |
| 10 | 0.000 | -0.001 | -0.001 | -0.001 | 0.025 | 0.024 | 0.025 | N/A | 0.695 | 0.696 |
| WLW: $H_1$ | | | | | | | | | | |
| $d$ | $\theta$ | $\hat{\theta}_1$ | $\hat{\theta}_2$ | $\hat{\theta}$ | $1-\beta_1$ | $1-\beta_2$ | $1-\beta_{12}$ | TP | $\rho_{(est)}$ | $\rho_{(sample)}$ |
| 1 | 0.194 | 0.189 | 0.189 | 0.189 | 0.67 | 0.67 | 0.92 | 0.89 | 0.023 | 0.035 |
| 5 | 0.245 | 0.161 | 0.161 | 0.161 | 0.51 | 0.50 | 0.66 | 0.57 | 0.408 | 0.417 |
| 10 | 0.275 | 0.106 | 0.106 | 0.106 | 0.24 | 0.23 | 0.26 | 0.24 | 0.695 | 0.694 |
| ZW: $H_0$ @ 5 intervals | | | | | | | | | | |
| $d$ | $\theta$ | $\hat{\theta}_1$ | $\hat{\theta}_2$ | $\hat{\theta}$ | $\alpha_1$ | $\alpha_2$ | $\alpha_{12}$ | N/A | $\rho_{(est)}$ | $\rho_{(sample)}$ |
| 1 | 0.000 | 0.000 | -0.001 | -0.001 | 0.022 | 0.023 | 0.023 | N/A | 0.023 | 0.032 |
| 5 | 0.000 | -0.001 | -0.001 | -0.001 | 0.022 | 0.026 | 0.027 | N/A | 0.387 | 0.404 |
| 10 | 0.000 | -0.001 | -0.001 | -0.001 | 0.025 | 0.025 | 0.026 | N/A | 0.670 | 0.688 |
| ZW: $H_1$ @ 5 intervals | | | | | | | | | | |
| $d$ | $\theta$ | $\hat{\theta}_1$ | $\hat{\theta}_2$ | $\hat{\theta}$ | $1-\beta_1$ | $1-\beta_2$ | $1-\beta_{12}$ | TP | $\rho_{(est)}$ | $\rho_{(sample)}$ |
| 1 | 0.194 | 0.189 | 0.189 | 0.189 | 0.66 | 0.66 | 0.91 | 0.89 | 0.028 | 0.034 |
| 5 | 0.245 | 0.161 | 0.161 | 0.161 | 0.50 | 0.49 | 0.65 | 0.57 | 0.389 | 0.406 |
| 10 | 0.275 | 0.107 | 0.107 | 0.107 | 0.24 | 0.23 | 0.26 | 0.24 | 0.671 | 0.685 |

For WLW, upon comparing with the paired case, slightly better type I errors (more conservative) and higher powers, are found to have been achieved for the indicators. Notably, the power exceeds the theoretical power in all settings and the correlations are consistent. Similar trends are observed in the results for the two methods, but the type I error rates when using WLW, are slightly more conservative than those using ZW.

### 6.6.4. **Progression-Free Survival**

As already described in Sections 4.3.4 and 4.5.2, PFS analysis considers either TTP or death (OS), whichever occurs first, as an event. This condition imposed a strong dependence between the two endpoints (PFS, OS); hence higher correlation compared to the indicators case. Consequently, their results differ, as shown in Table 6.10 and Figure 6.6 (d) for PFS.

Table 6.10: Simulation results for the PFS under the null and alternative using WLW and ZW *(ZW results correspond to Table 5.7).*

| WLW: $H_0$ | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $d$ | $\theta$ | $\hat{\theta}_1$ | $\hat{\theta}_2$ | $\hat{\theta}$ | $\alpha_1$ | $\alpha_2$ | $\alpha_{12}$ | N/A | $\rho_{(est)}$ | $\rho_{(sample)}$ |
| 1 | 0.000 | 0.000 | 0.000 | 0.000 | 0.025 | 0.024 | 0.024 | N/A | 0.608 | 0.611 |
| 5 | 0.000 | 0.000 | 0.000 | 0.000 | 0.025 | 0.025 | 0.025 | N/A | 0.779 | 0.784 |
| 10 | 0.000 | -0.001 | -0.001 | -0.001 | 0.025 | 0.024 | 0.025 | N/A | 0.879 | 0.881 |
| WLW: $H_1$ | | | | | | | | | | |
| $d$ | $\theta$ | $\hat{\theta}_1$ | $\hat{\theta}_2$ | $\hat{\theta}$ | $1\text{-}\beta_1$ | $1\text{-}\beta_2$ | $1\text{-}\beta_{12}$ | TP | $\rho_{(est)}$ | $\rho_{(sample)}$ |
| 1 | 0.223 | 0.213 | 0.222 | 0.216 | 0.84 | 0.80 | 0.90 | 0.88 | 0.608 | 0.613 |
| 5 | 0.252 | 0.156 | 0.165 | 0.158 | 0.56 | 0.52 | 0.59 | 0.53 | 0.779 | 0.786 |
| 10 | 0.275 | 0.100 | 0.106 | 0.101 | 0.24 | 0.23 | 0.25 | 0.22 | 0.879 | 0.881 |
| ZW: $H_0$ @ 5 intervals | | | | | | | | | | |
| $d$ | $\theta$ | $\hat{\theta}_1$ | $\hat{\theta}_2$ | $\hat{\theta}$ | $\alpha_1$ | $\alpha_2$ | $\alpha_{12}$ | N/A | $\rho_{(est)}$ | $\rho_{(sample)}$ |
| 1 | 0.000 | 0.000 | -0.001 | 0.000 | 0.025 | 0.023 | 0.024 | N/A | 0.589 | 0.601 |
| 5 | 0.000 | 0.000 | 0.000 | 0.000 | 0.026 | 0.025 | 0.026 | N/A | 0.755 | 0.779 |
| 10 | 0.000 | -0.001 | -0.001 | -0.001 | 0.024 | 0.023 | 0.025 | N/A | 0.869 | 0.882 |
| ZW: $H_1$ @ 5 intervals | | | | | | | | | | |
| $d$ | $\theta$ | $\hat{\theta}_1$ | $\hat{\theta}_2$ | $\hat{\theta}$ | $1\text{-}\beta_1$ | $1\text{-}\beta_2$ | $1\text{-}\beta_{12}$ | TP | $\rho_{(est)}$ | $\rho_{(sample)}$ |
| 1 | 0.223 | 0.214 | 0.225 | 0.218 | 0.83 | 0.81 | 0.90 | 0.89 | 0.592 | 0.601 |
| 5 | 0.252 | 0.156 | 0.167 | 0.159 | 0.55 | 0.52 | 0.59 | 0.53 | 0.755 | 0.779 |
| 10 | 0.275 | 0.102 | 0.108 | 0.103 | 0.24 | 0.23 | 0.25 | 0.23 | 0.869 | 0.882 |

Again, the findings for WLW are similar to the corresponding results for ZW. As anticipated, the marginal powers are now closer to the global power as the estimates of $\theta$ tend to vary slightly. Nevertheless, it is to be noticed that the correlation values remain consistent.

### 6.6.5. **Recurrent Events TT**

For the recurrent event using total time, highly dependent events are imposed as $T_1$ is directly contained within $T_2$ (Section 4.6.1). The simulation results for recurrent TT using WLW and ZW are given in Table 6.11 and in Figure 6.6 (e).

Table 6.11: Simulation results for recurrent events TT under the null and alternative using WLW and ZW *(ZW results correspond to Table 5.8).*

| WLW: $H_0$ | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $d$ | $\theta$ | $\hat{\theta}_1$ | $\hat{\theta}_2$ | $\hat{\theta}$ | $\alpha_1$ | $\alpha_2$ | $\alpha_{12}$ | N/A | $\rho$ (est) | $\rho$ (sample) |
| 1 | 0.000 | 0.000 | 0.001 | 0.000 | 0.026 | 0.025 | 0.026 | N/A | 0.556 | 0.557 |
| 5 | 0.000 | 0.001 | 0.000 | 0.001 | 0.027 | 0.025 | **0.028** | N/A | 0.745 | 0.753 |
| 10 | 0.000 | 0.001 | 0.001 | 0.001 | 0.025 | 0.025 | 0.026 | N/A | 0.856 | 0.856 |
| WLW: $H_1$ | | | | | | | | | | |
| $d$ | $\theta$ | $\hat{\theta}_1$ | $\hat{\theta}_2$ | $\hat{\theta}$ | 1-$\beta_1$ | 1-$\beta_2$ | 1-$\beta_{12}$ | TP | $\rho$ (est) | $\rho$ (sample) |
| 1 | 0.293 | 0.283 | **0.437** | 0.318 | 0.87 | **0.96** | 0.95 | 0.94 | 0.552 | 0.552 |
| 5 | 0.311 | 0.192 | 0.240 | 0.200 | 0.54 | 0.57 | 0.58 | 0.55 | 0.744 | 0.750 |
| 10 | 0.317 | 0.117 | 0.131 | 0.118 | 0.23 | 0.23 | 0.24 | 0.22 | 0.851 | 0.853 |
| ZW: $H_0$ @ 5 intervals | | | | | | | | | | |
| $d$ | $\theta$ | $\hat{\theta}_1$ | $\hat{\theta}_2$ | $\hat{\theta}$ | $\alpha_1$ | $\alpha_2$ | $\alpha_{12}$ | N/A | $\rho$ (est) | $\rho$ (sample) |
| 1 | 0.000 | 0.000 | 0.001 | 0.000 | 0.026 | 0.025 | 0.027 | N/A | 0.539 | 0.557 |
| 5 | 0.000 | 0.001 | 0.001 | 0.001 | **0.028** | 0.025 | **0.028** | N/A | 0.730 | 0.751 |
| 10 | 0.000 | 0.001 | 0.001 | 0.001 | 0.026 | 0.025 | 0.026 | N/A | 0.848 | 0.855 |
| ZW: $H_1$ @ 5 intervals | | | | | | | | | | |
| $d$ | $\theta$ | $\hat{\theta}_1$ | $\hat{\theta}_2$ | $\hat{\theta}$ | 1-$\beta_1$ | 1-$\beta_2$ | 1-$\beta_{12}$ | TP | $\rho$ (est) | $\rho$ (sample) |
| 1 | 0.293 | 0.282 | **0.430** | 0.317 | 0.86 | **0.96** | 0.95 | 0.94 | 0.543 | 0.555 |
| 5 | 0.311 | 0.192 | 0.238 | 0.200 | 0.53 | 0.57 | 0.58 | 0.55 | 0.730 | 0.749 |
| 10 | 0.317 | 0.118 | 0.131 | 0.120 | 0.24 | 0.24 | 0.24 | 0.23 | 0.848 | 0.856 |

Again, we see findings similar to those for earlier cases, but with a type I error rate for $d$ = 5 slightly inflated, but still within the 95% PI (0.022, 0.028). The global power exceeds the theoretical power at all correlation settings. These results, compared with those for ZW, are very similar indeed. For example, the marginal estimate $\hat{\theta}_2$ is larger than the global $\hat{\theta}$ and the marginal power 1 - $\beta_2$ is higher than

the global power ($d$ =1); for its detailed description the reader can be referred to Section 5.2.5.

### 6.6.6. **Recurrent Events GT**

As explained in Section 4.6.1, the recurrent GT model is technically different from the TT model. The results for recurrent GT are summarized in Table 6.12 and Figure 6.6 (f).

Table 6.12: Simulation results for recurrent events GT under the null and alternative using WLW and ZW *(ZW results correspond to* Table 5.9*)*.

| WLW: $H_0$ | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $d$ | $\theta$ | $\hat{\theta}_1$ | $\hat{\theta}_2$ | $\hat{\theta}$ | $\alpha_1$ | $\alpha_2$ | $\alpha_{12}$ | N/A | $\rho_{(est)}$ | $\rho_{(sample)}$ |
| 1 | 0.000 | -0.001 | 0.000 | 0.000 | 0.026 | 0.025 | 0.027 | N/A | 0.017 | 0.014 |
| 5 | 0.000 | 0.001 | 0.000 | 0.000 | 0.026 | 0.023 | 0.025 | N/A | 0.223 | 0.221 |
| 10 | 0.000 | 0.001 | 0.000 | 0.001 | 0.026 | 0.027 | 0.026 | N/A | 0.352 | 0.345 |
| WLW: $H_1$ | | | | | | | | | | |
| $d$ | $\theta$ | $\hat{\theta}_1$ | $\hat{\theta}_2$ | $\hat{\theta}$ | 1-$\beta_1$ | 1-$\beta_2$ | 1-$\beta_{12}$ | TP | $\rho_{(est)}$ | $\rho_{(sample)}$ |
| 1 | 0.235 | 0.231 | 0.227 | 0.229 | 0.71 | 0.57 | 0.90 | 0.88 | 0.017 | 0.017 |
| 5 | 0.271 | 0.182 | 0.137 | 0.163 | 0.49 | 0.25 | 0.56 | 0.49 | 0.222 | 0.224 |
| 10 | 0.289 | **0.117** | 0.069 | 0.096 | **0.23** | 0.10 | 0.22 | 0.19 | 0.350 | 0.348 |
| ZW: $H_0$ @ 5 intervals | | | | | | | | | | |
| $d$ | $\theta$ | $\hat{\theta}_1$ | $\hat{\theta}_2$ | $\hat{\theta}$ | $\alpha_1$ | $\alpha_2$ | $\alpha_{12}$ | N/A | $\rho_{(est)}$ | $\rho_{(sample)}$ |
| 1 | 0.000 | -0.001 | 0.000 | 0.000 | 0.025 | 0.025 | 0.026 | N/A | 0.040 | 0.030 |
| 5 | 0.000 | 0.001 | 0.000 | 0.000 | 0.026 | 0.022 | 0.022 | N/A | **0.305** | 0.235 |
| 10 | 0.000 | 0.001 | 0.000 | 0.001 | 0.026 | 0.027 | **0.020** | N/A | **0.512** | 0.352 |
| ZW: $H_1$ @ 5 intervals | | | | | | | | | | |
| $d$ | $\theta$ | $\hat{\theta}_1$ | $\hat{\theta}_2$ | $\hat{\theta}$ | 1-$\beta_1$ | 1-$\beta_2$ | 1-$\beta_{12}$ | TP | $\rho_{(est)}$ | $\rho_{(sample)}$ |
| 1 | 0.235 | 0.229 | 0.228 | 0.228 | 0.70 | 0.58 | 0.90 | 0.88 | 0.045 | 0.030 |
| 5 | 0.271 | 0.180 | 0.137 | 0.162 | 0.49 | 0.25 | 0.53 | 0.49 | **0.306** | 0.238 |
| 10 | 0.289 | **0.116** | 0.069 | 0.096 | **0.23** | 0.10 | 0.19 | 0.19 | **0.511** | 0.355 |

When compared with the results for ZW, the type I error rates for WLW are less conservative and the global powers for WLW are higher than theoretical powers in all settings. Like ZW, WLW yields marginal power 1 - $\beta_1$ that exceeds the global

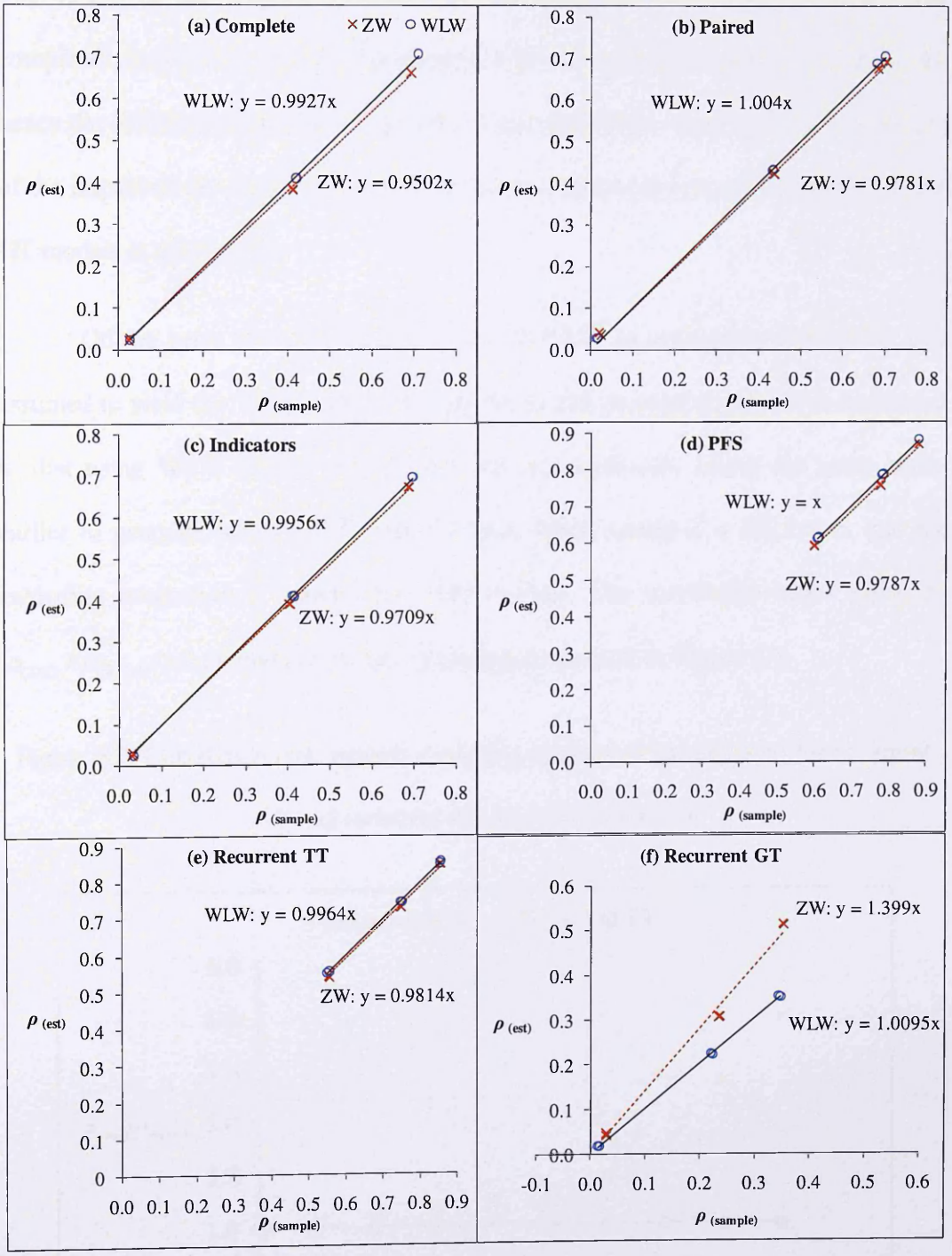
power ($d$ = 10); the marginal $\hat{\theta}_1$ is bigger than the global $\hat{\theta}$. Overall, the latter method obviously wins concerning the accuracy of the correlation estimates, whilst ZW gave an overestimation for recurrent GT.

### 6.6.7. **Summary of Correlation Ratios**

In Section 5.2.7, the correlation ratios obtained from ZW were plotted for each case, comparing the results when five and ten intervals were used ($k$ = 5, 10). It was concluded that $k$ = 5 gave a better estimate than $k$ = 10. In this section, only the former is considered and the correlation ratios for ZW are now compared to those of WLW. The plots of the sample correlation $\rho_{(\text{sample})}$ versus the derived correlation $\rho_{(\text{est})}$ for the six cases are given in Figures 6.6 (a) to (f).

Again, a linear plot with $y = x$ is an ideal situation indicating that the estimated correlation, $\rho_{(\text{est})}$ for either method is exactly the same as the correlation $\rho_{(\text{sample})}$ observed from its own samples of 10,000 replications. In all cases, WLW performs consistently well with ratios of 0.99 to 1.00, while those for ZW vary from 0.95 to 0.98 for most cases, except for the recurrent GT with 1.40. Figure 6.6 (a) shows that for the complete data, WLW gives a ratio of 0.99 compared to 0.95 for ZW. The former performs well at varying degrees of correlation, while the latter tends to perform slightly less well at higher correlations. In Figures 6.6 (b) to (f), the correlation ratios when using WLW are 1.00 for all the five cases. Meanwhile, when using ZW, the correlation ratios for the paired, related indicators, recurrent TT and recurrent GT are 0.98, 0.97, 0.98, 0.98 and 1.40 respectively.

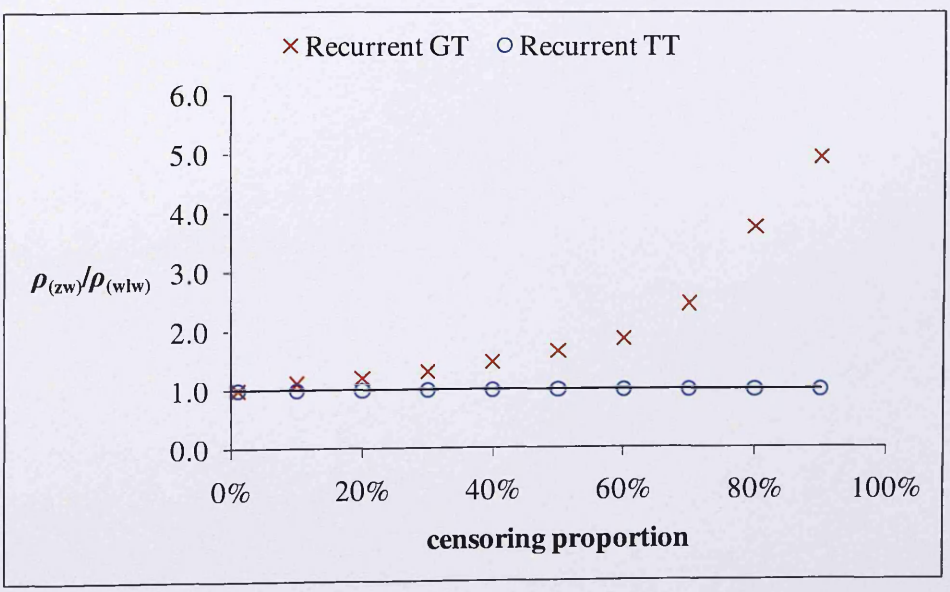
Figure 6.6: Plots of the sample correlation $\rho_{(\text{sample})}$ versus the derived correlation $\rho_{(\text{est})}$ for the six cases using WLW and ZW ($k = 5$).



To investigate further the 40 % overestimation for recurrent GT when using ZW, a comparison between the components that make up the correlation is performed. The values of the variances $V_1$, $V_2$ and the covariance $C_{12}$ when using ZW are compared to those for WLW. Apparently, both methods give consistent variances; hence the difference is solely attributable to the covariance, $C_{12}$. Next, an exploration of the impact of the censoring proportion on the correlation ratio of recurrent TT and GT models is undertaken.

On the basis of the evident accuracy of WLW in estimating correlation, it is assumed to yield the "true" correlation, and hence the estimate using ZW is compared to that using WLW in this exercise without any replicates. Using the same codes earlier to generate recurrent TT and GT data, while setting $d$ = 10, $k$ = 5, and the censoring proportion is varied from 0.01 to 0.90. The correlation ratios given by $\rho_{(ZW)} / \rho_{(WLW)}$ are plotted against the censoring proportion in Figure 6.7.

Figure 6.7: Correlation ratio against censoring proportion (percentage) for recurrent TT and recurrent GT ($d$ = 10, under $H_0$)



Ideally, all the points should lie on the horizontal line at $y = 1$, indicating accurate estimation of correlation: as observed for the case of recurrent TT. However, it is clear that for the recurrent GT, estimate of the correlation is only accurate when only 0.1% is censored. It is to be recalled that the earlier simulation was set at a censoring proportion of 40%, giving a correlation of 1.4 times higher as shown here. Upon increasing the censoring proportion, the correlation overestimation gets bigger, indicated by the upward gradient of the points. For example, the correlation is about 5 times higher when the data set is 90% censored. Closer examination unveils that the covariances when using ZW are substantially larger than those given by WLW, while the variances remain consistent. In Section 4.3.1, it was explained that, with heavy censoring, the contribution of the marginal failures diminishes and the estimator relies largely on the risks sets and the combined failures. This finding renders the ZW formulation for covariance estimation unsuitable for recurrent GT.

## 6.7. Overall Results: WLW vs ZW

In the previous sections, the results for each of the six cases investigated have been described in turn. This section now compares WLW to ZW in terms of their overall performance based on the key measures specified: type I error, power and correlation ratio. Under each hypothesis, 18 sets of data (six cases each with three varied correlations) are used to summarize the three key measures for each method. Figure 6.8 (a) and (b) respectively display the type I error rates for both methods, identified by cases and degrees of correlation. Figure 6.8 (c) shows how the power varies for all six cases with the degrees of correlation imposed, while the last chart summarizes all the correlation values for both methods, the gradient measuring the overall correlation ratio of $\rho_{(est)}/\rho_{(sample)}$ for each method. Consistently with the previous chapter, a diagonal line of $y = x$ serves as a reference line for an ideal situation; points below or above it are described accordingly.

Figure 6.8: Type 1 error rates, powers and correlation ratios of ZW versus WLW.

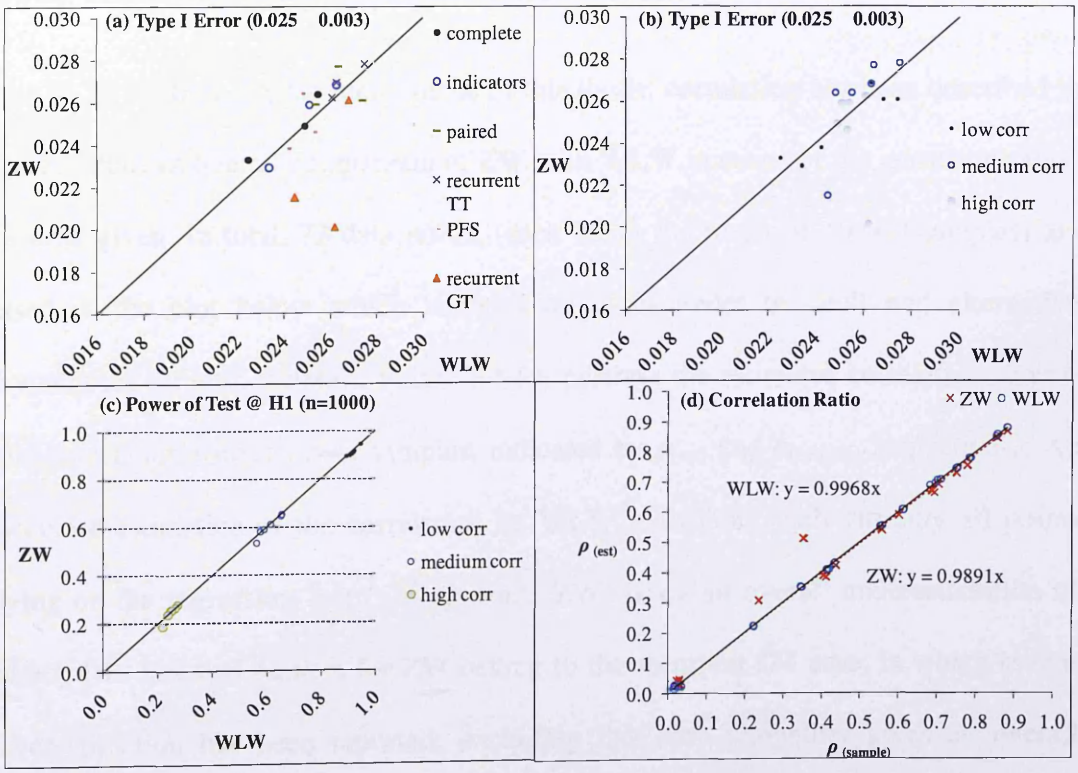


Figure 6.8 (a) shows that both methods give type I error rates within the 95% probability interval (0.022, 0.028). The points above the diagonal line ($y = x$) indicate the better performance of WLW in detecting any departure from the null hypothesis, and vice versa. It is evident that WLW seems to perform better than ZW by the predominance of points above the line leaving only seven points underneath. Figure 6.8 (b) shows no particular trend for type I error rates with regard to varied correlations.

The power of ZW is comparable to that of WLW as is evident in Figure 6.8 (c). Both methods give powers equal to or more than 0.89 ($1 - \beta = 0.90 \pm 0.01$) at low correlation ($d = 1$), but when a higher correlation is imposed by increasing the subject effect, the power reduces accordingly for both methods. It is to be recalled that such behaviour is expected owing to the discrepancy in $\theta$ setting as explained in Section 5.1 earlier. The appearance of two points slightly below the diagonal line for power suggests that WLW is only slightly better than ZW: ZW suffers a slight power loss owing to its interval nature, as already noted.

In Section 4.3, the main theme of this thesis, correlation has been described in great detail; an overall comparison of ZW with WLW in terms of the correlation ratio is now given. In total, 72 data points (each being the mean of 10,000 samples) are used in the plot below which includes all cases under the null and alternative hypotheses for both methods. Figure 6.8 (d) portrays the estimated correlation against the correlation from its own samples, indicated by $\rho_{(est)}$ and $\rho_{(sample)}$ respectively. An accurate estimation of the correlation by WLW is evident, with virtually all points lying on the regression line ($y = x$), while ZW shows an overall underestimation of about 1%. The two outliers for ZW belong to the recurrent GT case, in which severe overestimation has been reported; excluding this case altogether gives an overall

underestimation by 2%. Theoretical equality of the correlation between the two estimates of the treatment advantage and that between two score statistics given in equation (6.29) is hereby validated accordingly.

All in all, the two methods demonstrated similar trends in all the six cases explored, with the exception of an overestimation of the correlation for recurrent GT when using ZW.

## 6.8. Discussion

ZW provides a conceptually straightforward approach to the analysis of multivariate survival data. The benefits of our method are (i) ease of use: simple computation and (ii) good interpretability: straightforward derivation based on marginal analyses, unlike WLW, which is based on non-standard adjusted statistics. By design, ZW is capable of analyzing naturally interval-censored data whereas WLW was not specifically intended to cater for such data. Meanwhile the disadvantages when dealing with continuous data are that ZW requires categorization into intervals, and consequently might lose a little power. Despite their technical differences (Section 6.4), extensive simulations show that our new method is accurate, consistent and comparable to WLW in the five main bivariate cases investigated.

As reported earlier, ZW works well with modelling of total time, but as far as recurrent events gap time is concerned, the correlation estimate is accurate only for a very small censoring proportion. As shown in Section 6.6.6, ZW overestimates increasingly with increasing censoring proportion. This could be due to inaccurate estimation of the probabilities of each combined outcome for the paired intervals for GT. As already noted in Chapter 4, the covariance formulation implies that, with heavy censoring, the contribution of the marginal failures diminishes and the estimator

relies largely on the risks sets and the combined failures. That this proposed method is not applicable to recurrent gap time is an interesting finding, but not one explainable here.

With regard to WLW, it actually uses concepts similar to the jackknife (Section 6.1.4), although the original paper by Wei et al. (1989) makes no reference to any such similarity. The much emphasized issue of overestimation by WLW (Kelly & Lim, 2000), is rather an inevitable scenario when using total time convention for recurrent events since the effect for $T_1$ is carried over into $T_2$. Kelly & Lim (2000) concluded that WLW is most appropriate for data arising from different types of event from the same individual rather than for recurrent events, while our simulations show that no one case affirms its superiority: it performs equally well in all six cases explored. Kelly & Lim (2000) also commented that the within subject correlation was not satisfactorily accounted for by employing the robust variance, but that the reason is unknown. Metcalfe & Thompson (2007) also argued on the same point, but commentary is rather unclear on how else to address it. As derived in Section 4.3, our proposed method embeds within-subject correlation directly in the computation of covariance which considers the pair of intervals for each subject exclusively.

WLW assumes continuous survival times while ZW is based on interval-censored survival times. Nevertheless, the former has been applied to interval-censored data using the real data sets in Section 6.5 which showed acceptable results. Extensions of WLW for multivariate interval-censored data have earlier been studied by Goggins & Finkelstein (2000), and Kim & Xue (2002). Hence its application here serves as further evidence of its flexibility in the analysis of bivariate survival data.

As earlier noted, both ZW and WLW are marginal models which assume proportional hazards. In marginal models with total time risk intervals, the information in the early part of the total observation period will be used several times: once for each event. This induces a greater weight on the first recurrence considered (as observed in Figure 6.3, Section 6.2). The alternative frailty models which include a common random effect to adjust for within-subject correlation are available in the literature (Vaupel et al., 1979; Clayton & Cuzick, 1985; Klein, 1992). They estimate the covariate effect on the recurrence, conditionally on a specified dependence structure. Such models can be of interest for a full modelling of the recurrent event process, but are of far less interest in studies which focus on the average covariate effect.

In terms of usability, ZW requires simple calculations which can be achieved by a MS Excel spreadsheet or something like it, whereas WLW involves complex formulation which necessitates some programming or standard software packages. It is to be noted that only the SAS software package was considered in this study, for suitability and convenience purposes. Other software packages for analyzing correlated survival data are namely Stata, S-Plus and R, MLwiN, and WinBUGS. A good review of their different capabilities, including those of SAS, is provided by Kelly (2004).

Workable for practical applications to real data and comparable to the established WLW, our proven method ought to contribute a new alternative method for the analysis of correlated survival data.

# Chapter 7. Conclusions and Further Work

Our proposed method, ZW has proved useful for finding estimates of correlations and treatment effects. Its performance is comparable to that of WLW and it has a clearer interpretation. To my knowledge, this is the first successful development of the global score test methodology for bivariate survival data. The derivation of the covariance between two score statistics using an interval-censored model, viewed as a viable alternative to the logrank, is a new approach. The relationship between the old jackknife technique and WLW has also been clarified in this thesis. The new method, ZW should be a useful contribution of knowledge to survival analysis. However, it does not outperform the established method, WLW.

The whole procedure of ZW enables estimation of correlations, and thus has many practical uses. Among these are (i) combined null hypothesis testing to establish whether a linear combination of effects is equal to zero and (ii) global null hypothesis testing, establishing whether all effects are equal to zero, both of which have been demonstrated in this thesis. Nevertheless, various problems were encountered and should be investigated further, such as the unsuitability of ZW for handling recurrent events gap time.

Two key questions relating to the issues discussed in Section 5.1 are: "How to formulate the correct treatment advantage between-subject for simulation purposes?" and "How to achieve equality of the proportional hazards ratios?". In situations where these ratios are unequal (as observed in the cases of PFS and recurrent events), a method to estimate the theoretical power, comparable with the method achieved by Bolland et al. (2009) for binary data, should be devised. It is also essential to account adequately for the within-subject correlation or between-subject correlation or both.

Essentially, the application of the global score approach to interval-censored survival data is valid as long as the assumption of proportional hazards is adequately met. That having been said, the global score test approach can be adapted to combine normally distributed, Poisson-distributed or other types of endpoint which can also be mixed together. This promising method can be further developed to allow adjustment for combining composite endpoints, for example, two different types of endpoint arising from binary and survival data whereby the treatment effects are measured on log-odds ratio and log hazard ratio respectively.

In situations where the failures are unequally distributed across intervals, ZW may not be as efficient as established for the equal failure distribution. As noted earlier, the global test methodology may be most suitable when none of the endpoints captures all of the information on a patient's condition, for example, a case involving the multiple stroke scales cited in Chapter 2. Similar features of outcome measures exist in mental health studies; hence it could be a potential application area of this methodology.

In this study, apart from the treatment advantages, only the correlation between the two score statistics (as well as) was investigated. For further understanding of correlated endpoints, perhaps it is worthwhile to consider the correlation between $T_1$ and $T_2$ directly, as was attempted by Fleischer et al. (2009). Connections between ZW and frailty modelling should also be investigated to yield better insight into subject effects. Unfortunately, due to resource constraints, some extended investigations were not feasible and hence have to be considered as further work. These include adjustment for covariates and the construction of joint confidence regions for multiple hazard ratios. Implementations of sequential and multiple testing procedures are worth investigating as these are becoming increasingly popular in view

of minimizing the cost of trials. Additionally, comparison of multiple active treatments to a common control may be attempted.

I am pleased that the results from this thesis are encouraging, although some investigations could not be completed according to the original plan. Nevertheless, that does not diminish the knowledge and skills I have learned throughout this study. I now look forward to continuing teaching and collaborating on new research projects. I am already excited about the huge potential for growth of medical statistics, across Asia generally and within Malaysia specifically.

# APPENDIX

## A. SAS Codes for ZW Specific for Analysis of Hip Revision Data

```
*Hips bilateral data using ZW for 2 intervals;

%macro hct( imin , imax, last);

data bilat;
      set bil;
array cut1{&imax}(15      35);
array cut2{&imax}(15      35);
      t1a = 1;  t2a = 1;
            do i = &imin to &imax;
                  t1a = t1a + (t1 > cut1(i));
                  t2a = t2a + (t2 > cut2(i));
            end;
      t1=t1a; t2=t2a;
      keep id trt t1  t2 cens1 cens2;
run;

data add;
      set bilat end=lastval;
      do j = &imin to &imax;
            do k = &imin to &imax;
      if trt=2 then do;

      array affe{&imax,&imax } ffe&imin-ffe&last;
      atffe= (t1=j)*(cens1=1)*(t2=k)*(cens2=1);
      affe {j,k}  = sum(atffe, affe {j,k}, 0);
      retain ffe&imin-ffe&last 0;

      array afse {&imax,&imax }  fse&imin-fse&last;
      atfse= (t1=j)*(cens1=1)*(t2>k);
      afse {j,k}  = sum(atfse, afse {j,k}, 0);
      retain fse&imin-fse&last 0;

      array afme{&imax,&imax } fme&imin-fme&last;
      atfme= (t1=j)*(cens1=1)*(t2=k)*(cens2=0);
      afme {j,k}  = sum(atfme, afme {j,k}, 0);
      retain fme&imin-fme&last 0;

      array amfe{&imax,&imax } mfe&imin-mfe&last;
      atmfe= (t1=j)*(cens1=0)*(t2=k)*(cens2=1);
      amfe {j,k}  = sum(atmfe, amfe {j,k}, 0);
      retain mfe&imin-mfe&last 0;

      array asse {&imax,&imax } sse&imin-sse&last;
      atsse= (t1>j)*(t2>k);
      asse {j,k}  = sum(atsse, asse {j,k}, 0);
      retain sse&imin-sse&last 0;

      array asfe {&imax,&imax } sfe&imin-sfe&last ;
      atsfe= (t1>j)*(t2=k)*(cens2=1);
      asfe {j,k}  = sum(atsfe, asfe {j,k}, 0);
      retain sfe&imin-sfe&last 0;

      array asme {&imax,&imax } sme&imin-sme&last;
```

```
      atsme= (t1>j)*(t2=k)*(cens2=0);
      asme {j,k}  = sum(atsme, asme {j,k}, 0);
      retain sme&imin-sme&last 0;

      array amse {&imax,&imax } mse&imin-mse&last;
      atmse= (t1=j)*(t2>k)*(cens1=0);
      amse {j,k}  = sum(atmse, amse {j,k}, 0);
      retain mse&imin-mse&last 0;

      array ao1e {&imax}o1e&imin-o1e&imax;
      ao1e{j} = sum( affe {j,&imin},afse {j,&imin},afme {j,&imin});
      retain o1e&imin-o1e&imax 0;

      array ao2e {&imax}o2e&imin-o2e&imax;
      ao2e{k} = sum( affe {&imin, k}, asfe {&imin,k},amfe {&imin,k});
      retain o2e&imin-o2e&imax 0;

      array as1e {&imax}s1e&imin-s1e&imax;
      as1e{j} = sum( asfe {j,&imin},asse {j,&imin},asme {j,&imin});
      retain s1e&imin-s1e&imax 0;

      array as2e {&imax}s2e&imin-s2e&imax;
      as2e{k} = sum( afse {&imin, k}, asse {&imin,k}, amse
{&imin,k});
      retain s2e&imin-s2e&imax 0;

            end;
      end;
end;
      do j = &imin to &imax;
            do k = &imin to &imax;
      if trt=1 then do;

      array affc{&imax,&imax } ffc&imin-ffc&last;
      atffc= (t1=j)*(cens1=1)*(t2=k)*(cens2=1);
      affc {j,k}  = sum(atffc, affc {j,k}, 0);
      retain ffc&imin-ffc&last 0;

      array afsc {&imax,&imax }  fsc&imin-fsc&last;
      atfsc= (t1=j)*(cens1=1)*(t2>k);
      afsc {j,k}  = sum(atfsc, afsc {j,k}, 0);
      retain fsc&imin-fsc&last 0;

      array afmc{&imax,&imax } fmc&imin-fmc&last;
      atfmc= (t1=j)*(cens1=1)*(t2=k)*(cens2=0);
      afmc {j,k}  = sum(atfmc, afmc {j,k}, 0);
      retain fmc&imin-fmc&last 0;

      array amfc{&imax,&imax } mfc&imin-mfc&last;
      atmfc= (t1=j)*(cens1=0)*(t2=k)*(cens2=1);
      amfc {j,k}  = sum(atmfc, amfc {j,k}, 0);
      retain mfc&imin-mfc&last 0;

      array assc {&imax,&imax } ssc&imin-ssc&last;
      atssc= (t1>j)*(t2>k);
      assc {j,k}  = sum(atssc, assc {j,k}, 0);
      retain ssc&imin-ssc&last 0;

      array asfc {&imax,&imax } sfc&imin-sfc&last ;
      atsfc= (t1>j)*(t2=k)*(cens2=1);
      asfc {j,k}  = sum(atsfc, asfc {j,k}, 0);
```

```
      retain sfc&imin-sfc&last 0;

      array asmc {&imax,&imax } smc&imin-smc&last;
      atsmc= (t1>j)*(t2=k)*(cens2=0);
      asmc {j,k}  = sum(atsmc, asmc {j,k}, 0);
      retain smc&imin-smc&last 0;

      array amsc {&imax,&imax } msc&imin-msc&last;
      atmsc= (t1=j)*(t2>k)*(cens1=0);
      amsc {j,k}  = sum(atmsc, amsc {j,k}, 0);
      retain msc&imin-msc&last 0;

      array ao1c {&imax}o1c&imin-o1c&imax;
      ao1c{j} = sum( affc {j,&imin},afsc {j,&imin},afmc {j,&imin});
      retain o1c&imin-o1c&imax 0;

      array ao2c {&imax}o2c&imin-o2c&imax;
      ao2c{k} = sum( affc {&imin, k}, asfc {&imin,k},amfc {&imin,k});
      retain o2c&imin-o2c&imax 0;

      array as1c {&imax}s1c&imin-s1c&imax;
      as1c{j} = sum( asfc {j,&imin},assc {j,&imin},asmc {j,&imin});
      retain s1c&imin-s1c&imax 0;

      array as2c {&imax}s2c&imin-s2c&imax;
      as2c{k} = sum( afsc {&imin, k}, assc {&imin,k},amsc {&imin,k});
      retain s2c&imin-s2c&imax 0;

            end;
      end;
end;
      IF lastval ~= 1 THEN DELETE;
keep f: s: o: m:
;
run;

data summary;
      set add;
      do j = &imin to &imax;
            do k = &imin to &imax;
array ao1e {&imax } o1e&imin-o1e&imax;
array ao1c {&imax } o1c&imin-o1c&imax;
array ao1 {&imax } o1x&imin-o1x&imax;
array as1e {&imax}s1e&imin-s1e&imax;
array as1c {&imax}s1c&imin-s1c&imax;
array ar1e {&imax}r1e&imin-r1e&imax;
array ar1c {&imax}r1c&imin-r1c&imax;
array ar1 {&imax}r1x&imin-r1x&imax;
array aq1 {&imax}q1x&imin-q1x&imax;
array aZ1 {&imax}Z1x&imin-Z1x&imax;
array aV1 {&imax}V1x&imin-V1x&imax;
ao1{j}= sum(ao1c{j}, ao1e{j});
ar1e{j}= sum(ao1e{j}, as1e{j});
ar1c{j}= sum(ao1c{j}, as1c{j});
ar1{j}= sum(ar1c{j}, ar1e{j});
aq1{j}= -log(1-(ao1{j}/ar1{j}));
aZ1{j} = aq1{j}*(ar1e{j}*ao1c{j}-ar1c{j}*ao1e{j})/ao1{j};
aV1{j} = aq1{j}**2*ar1e{j}*ar1c{j}*(ar1{j}-ao1{j})/
(ao1{j}*(ar1{j}-1));
array ao2e {&imax } o2e&imin-o2e&imax;
array ao2c {&imax } o2c&imin-o2c&imax;
```

```
array ao2 {&imax } o2x&imin-o2x&imax;
ao2{k}= sum(ao2c{k}, ao2e{k});
array as2e {&imax}s2e&imin-s2e&imax;
array as2c {&imax}s2c&imin-s2c&imax;
array ar2e {&imax}r2e&imin-r2e&imax;
array ar2c {&imax}r2c&imin-r2c&imax;
array ar2 {&imax}r2x&imin-r2x&imax;
array aq2 {&imax}q2x&imin-q2x&imax;
array aZ2 {&imax}Z2x&imin-Z2x&imax;
array aV2 {&imax}V2x&imin-V2x&imax;
ao2{k}= sum(ao2c{k}, ao2e{k});
ar2e{k}= sum(ao2e{k}, as2e{k});
ar2c{k}= sum(ao2c{k}, as2c{k});
ar2{k}= sum(ar2c{k}, ar2e{k});
aq2{k}= -log(1-(ao2{k}/ar2{k}));
aZ2{k} = aq2{k}*(ar2e{k}*ao2c{k}-ar2c{k}*ao2e{k})/ao2{k};
aV2{k} = aq2{k}**2*ar2e{k}*ar2c{k}*(ar2{k}-ao2{k})/
(ao2{k}*(ar2{k}-1));
array affe {&imax,&imax } ffe&imin-ffe&last;
array afse {&imax,&imax } fse&imin-fse&last;
array asse {&imax,&imax } sse&imin-sse&last;
array asfe {&imax,&imax } sfe&imin-sfe&last;
array arte {&imax,&imax } rte&imin-rte&last;
array affc {&imax,&imax } ffc&imin-ffc&last;
array afsc {&imax,&imax } fsc&imin-fsc&last;
array assc {&imax,&imax } ssc&imin-ssc&last;
array asfc {&imax,&imax } sfc&imin-sfc&last;
array artc {&imax,&imax } rtc&imin-rtc&last;
array aff {&imax,&imax } ff&imin-ff&last;
array afs {&imax,&imax } fs&imin-fs&last;
array ass {&imax,&imax } ss&imin-ss&last;
array asf {&imax,&imax } sf&imin-sf&last;
array art {&imax,&imax } rt&imin-rt&last;
array aft1 {&imax,&imax } f1t&imin-f1t&last;
array aft2 {&imax,&imax } f2t&imin-f2t&last;
array aC {&imax,&imax } C&imin-C&last;
aff{j,k}= sum(affe{j,k},affc{j,k});
asf{j,k}= sum(asfe{j,k},asfc{j,k});
afs{j,k}= sum(afse{j,k},afsc{j,k});
ass{j,k}= sum(asse{j,k},assc{j,k});
arte{j,k}= sum(affe{j,k},afse{j,k},asse{j,k},asfe{j,k});
artc{j,k}= sum(affc{j,k},afsc{j,k},assc{j,k},asfc{j,k});
art{j,k}= sum(arte{j,k},artc{j,k});
aft1{j,k}= sum(aff{j,k},afs{j,k});
aft2{j,k}= sum(aff{j,k},asf{j,k});
aC{j,k}=(aq1{j}*aq2{k}*((ar1e{j}*ar2e{k}*artc{j,k}+(ar1c{j}*ar2c{k}*a
rte{j,k}))*((aff {j,k }*art{j,k}-(aft1 {j,k}*aft2 {j,k }))))
      /(ao1{j}*ao2{k}*(art{j,k}**2)));
Z1= sum(of Z1x&imin-Z1x&imax);
V1= sum(of V1x&imin-V1x&imax);
Z2= sum(of Z2x&imin-Z2x&imax);
V2= sum(of V2x&imin-V2x&imax);
Cov=sum(of C&imin-C&last);
Cor=Cov/sqrt(V1*V2);
Zs= sum(Z1, Z2)*sum(V1,V2)/sum(sum(V1,V2), 2*Cov);
Vs= sum(V1,V2)**2/sum(sum(V1,V2), 2*Cov);
theta1=Z1/V1;
theta2=Z2/V2;
thetas=Zs/Vs;*standard theta;
thetaz=((V1-Cov)/(sum(V1, V2,-2*Cov)))*theta1 + ((V2-Cov)/
(sum(V1, V2,-2*Cov)))*theta2;*optimal theta;
```

```
vthetaz= (V1*V2-Cov**2)/(V1*V2*(sum(V1, V2,-2*Cov)));* variance of
the optimal theta;
p1=(1-probnorm(Z1/sqrt(V1)));
p2=(1-probnorm(Z2/sqrt(V2)));
p3=(1-probnorm(Zs/sqrt(Vs)));
pz=(1-probnorm(thetaz/sqrt(vthetaz)));
end;
      end;
keep
      theta: p: Vs Zs  Z1 V1  Z2 V2  Cov  Cor vthetaz;
run;

proc print data = summary ;   *ZW method;
      title "ZW bilat at &imax intervals";
      var Z1 Z2 Zs V1 V2 Vs Cov Cor p: theta: vthetaz;
run;

proc means data = summary ;   *ZW method;
      title "ZW bilat at &imax intervals";
      var Z1 Z2 Zs V1 V2 Vs Cov Cor p: theta: vthetaz;
run;

%mend;
%hct( imin=1 , imax=2, last=4);
```

## B. SAS Codes for WLW Specific for Analysis of Bladder Cancer Data

```
*Bladder Cancer (TT) Data Using WLW;
options nolabel;

proc sort data = blad_tt;
  by id;
run;

data blad_tt;
       set blad tt;
  trt1 = trt*(visit = 1);
  trt2 = trt*(visit = 2);
run;

ods output parameterestimates=estw;
ods output testaverage=eta;

PROC PHREG data = blad_tt outest = est1 covs (aggregate);
  MODEL tstop*status(0) = trt1-trt2;
  TREATMENT: test trt1,trt2/average e;*gives estimate of global
treatment effect ;
  OUTPUT out = out1 dfbeta = dt1-dt2/ order = data;
  STRATA visit;
  ID id;
RUN;

PROC MEANS data = out1 noprint;
  BY id;
  VAR dt1-dt2;
  OUTPUT out = out2 sum = dt1-dt2;
RUN;

PROC IML;
  USE out2;
  READ all var{dt1 dt2} into x;
  v = x` * x;
  RESET noname;
  vname = {"trt1", "trt2"};
  corr = v[1,2]/SQRT(v[1,1]#v[2,2]);
  v[1,1] = 1/v[1,1];
  v[2,2] = 1/v[2,2];
  v[1,2] = v[1,2] # v[1,1] # v[2,2];
  v[2,1] = v[1,2];
  CALL SYMPUT('wlwCorr',LEFT(CHAR(corr,6,4)));
  PRINT, "estimated covariance matrix (WLW)", ,
            v[colname = vname rowname = vname format = 10.5];
  PRINT, "estimated correlation",  corr[colname="corr"];
  CREATE rcov from v[colname = vname rowname = vname];
  APPEND from v[rowname = vname];
  CLOSE rcov;
QUIT;
RUN;

data rcov1;*split for merging later;
      set rcov;
      if VNAME='trt2' then delete;
      if VNAME='trt1' then VNAME = ' est ' ;
      rename trt1=v1w trt2=covz;
run;
```

```
data rcov2;
      set rcov;
      if VNAME='trt1' then delete;
      if VNAME='trt2' then VNAME = ' est ' ;
      rename trt1=covz trt2=v2w;
run;

proc sort data=estw  (keep= parameter estimate stderr probchisq );
      by parameter;
run;

data estw1;
      set estw;
      if parameter='trt2' then delete;
      if parameter='trt1' then parameter = ' est ' ;
      rename estimate=w1 probchisq=p1 stderr=s1 parameter=VNAME;
run;

data estw2;
      set estw;
      if parameter='trt1' then delete;
      if parameter='trt2' then parameter = ' est ' ;
      rename estimate=w2 probchisq=p2 stderr=s2 parameter=VNAME;;
run;

data eta1;
      set eta;
      if Label='TREATMENT' then Label = ' est ' ;
      drop zscore;
      rename estimate=w3 probz=p3 stderr=s3 Label=VNAME;;
run;

data est;
      merge estw1 estw2 eta1 rcov1 rcov2;
      by vname ;
      y1=1/v1w;
      y2=1/v2w;
      z1w=-w1/y1;
      z2w=-w2/y2;
      covw=covz*y1*y2;
      corw=covw/sqrt(y1*y2);
      vpw = sum(v1w, v2w);
      zpw = sum(z1w, z2w);
      vsw = (vpw)**2/ sum(vpw, 2*covz);
      zsw = (zpw*vpw)/ sum(vpw, 2*covz);
      theta1w = -w1;
      theta2w = -w2;
      thetasw = zsw/vsw;
      eta3 = -w3;
      p1 = p1;
      p2=p2;
      p1w = (1-probnorm(-w1/s1));
      p2w = (1-probnorm(-w2/s2));
      psw = (1-probnorm(eta3/s3));
run;

proc means data=est;
      title "Bladder WLW TT ";
      var w: v1w v2w vsw z1w z2w zsw c: p:  theta: eta3 ;
run;
```

## C. SAS Codes for Comparison between The Jackknife, Exact Delta-Beta and DFBETA Influences

```
*Compare Jackknife, Exact Delta-Beta & DFBETA for Bladder Data (TT);

options ps = 50 ls = 78;
libname blad 'F:\Recovered\Data\Blad';

%macro id(id);

data blad_tt12;
      set blad.blad_12;
      if id ne &i;
run;

ods output parameterestimates=estw;

proc phreg data = blad tt12 outest = est12 ;
  model tstop*status(0) = trt1-trt2;
  Trt1= Trt * (Visit=1);
  Trt2= Trt * (Visit=2);
  strata visit;
run;

proc sort data=estw  (keep= parameter estimate stderr probchisq );
      by parameter;
run;

data estw1;
      set estw;
      if parameter='trt2' then delete;
      if parameter='trt1' then parameter =  &i  ;
      rename estimate=w1 probchisq=p1 stderr=s1 parameter=VNAME;
run;

data estw2;
      set estw;
      if parameter='trt1' then delete;
      if parameter='trt2' then parameter = &i  ;
      rename estimate=w2 probchisq=p2 stderr=s2 parameter=VNAME;
run;

data cov;
      merge estw1 estw2;
by vname;
run;

data covall;
      set covall cov;
run;

data result;
      set result est12;/*Show all 86 of JK's n-1 individual
estimates*/
run;

%mend;
%macro iter (n);
```

```
%do i = 0 %to &n;
%id(&i);
%end;

%mend;
%iter(n=86);

options nolabel;

proc printto;
run;

proc means data=result;
      output out=ave mean=trt1-trt2;/*gives JK's overall estimate*/
run;

data calc;
      merge covall ave;
      if trt1= '.' then trt1=-0.3627418;*JK's average of 86 (n-1
estimates);
      if trt2='.' then trt2=-0.5520362;
      ex1=-0.3626606-w1;*exact delta beta;
      ex2=-0.5517842-w2;
      vex1=(w1--0.3626606)**2;*exact var;
      vex2=(w2--0.5517842)**2;
      covex=sqrt(vex1)*sqrt(vex2);
      jk1=trt1-w1;
      jk2=trt2-w2;
      vjk1=(w1-trt1)**2;*JK's var;
      vjk2=(w2-trt2)**2;
      covjk=sqrt(vjk1)*sqrt(vjk2);
      keep w1 w2 trt1 trt2 ex1 ex2 jk1 jk2 vjk1 vjk2 covjk vex1 vex2
covex;
run;

proc means data=calc sum;
run;

data blad.calc1;
      set calc;
run;

data blad.dtcomp;
      merge blad.dt_all blad.calc1 ;
      keep id dt1 dt2 jk1 jk2 ex1 ex2 ;
run;

proc iml;
  use blad.dtcomp;
  read all var{dt1 dt2} into x;
  v = x` * x;
  reset noname;
  vname = {"trt1", "trt2"};
  corr = v[1,2]/SQRT(v[1,1]#v[2,2]);
  v[1,1] = 1/v[1,1];
  v[2,2] = 1/v[2,2];
  v[1,2] = v[1,2] # v[1,1] # v[2,2];
  v[2,1] = v[1,2];
  CALL SYMPUT('wlwCorr',LEFT(CHAR(corr,6,4)));
  PRINT, "estimated covariance matrix DFBETA WLW", ,
```

```
      v[colname = vname rowname = vname format = 10.5]
corr[colname="corr"];
  CREATE rcov from v[colname = vname rowname = vname];
  APPEND from v[rowname = vname];
  CLOSE rcov;
QUIT;
RUN;

proc iml;
  use blad.dtcomp;
  read all var{ex1 ex2} into x;
  v = x` * x;
  reset noname;
  vname = {"trt1", "trt2"};
  corr = v[1,2]/SQRT(v[1,1]#v[2,2]);
  v[1,1] = 1/v[1,1];
  v[2,2] = 1/v[2,2];
  v[1,2] = v[1,2] # v[1,1] # v[2,2];
  v[2,1] = v[1,2];
  CALL SYMPUT('wlwCorr',LEFT(CHAR(corr,6,4)));
  PRINT, "estimated covariance matrix: EXACT", ,
      v[colname = vname rowname = vname format = 10.5]
corr[colname="corr"];
  CALL SYMPUT('wlwCorr',LEFT(CHAR(corr,6,4)));
  PRINT, "estimated correlation: EXACT",  corr[colname="corr"];
  CREATE rcov from v[colname = vname rowname = vname];
  APPEND from v[rowname = vname];
  CLOSE rcov;
QUIT;
RUN;

proc iml;
  use blad.dtcomp;
  read all var{jk1 jk2} into x;
  v = x` * x;
  reset noname;
  vname = {"trt1", "trt2"};
  corr = v[1,2]/SQRT(v[1,1]#v[2,2]);
  v[1,1] = 1/v[1,1];
  v[2,2] = 1/v[2,2];
  v[1,2] = v[1,2] # v[1,1] # v[2,2];
  v[2,1] = v[1,2];
  CALL SYMPUT('wlwCorr',LEFT(CHAR(corr,6,4)));
  PRINT, "estimated covariance matrix: JACKKNIFE", ,
      v[colname = vname rowname = vname format = 10.5]
corr[colname="corr"];
  CREATE rcov from v[colname = vname rowname = vname];
  APPEND from v[rowname = vname];
  CLOSE rcov;
QUIT;
RUN;
```

# REFERENCES


Abulibdeh, H., Turnbull, B. W., and Clark, L. C. (1990). Analysis of multitype recurrent events in longitudinal-studies - application to a skin-cancer prevention trial. *Biometrics***46**, 1017-1034.

Allison, P. D. (2001). *Survival Analysis Using the SAS System a Practical Guide*, 5 edition. Cary, N.C. : SAS Institute, c1995.

Andersen, P. K., and Gill, R. D. (1982). Cox regression-model for counting-processes - a large sample study. *Annals of Statistics***10**, 1100-1120.

Anderson, P. K., Borgan, O., Gill, R. D., and Keiding, N. (1993). *Statistical Models Based on Counting Processes*. New York: Springer-Verlag.

Anderson, P. K., and Keiding, N. (2006). *Survival and event history analysis*. Chichester: John Wiley & Sons Ltd.

Armitage, P., and Gehan, E. (1974). Statistical-methods for identification and use of prognostic factors. *International Journal of Cancer*, 16-36.

Azzalini, A. (1996). *Statistical inference : based on the likelihood*. London: Chapman & Hall.

Bijwaard, G. E., Franses, P. H., and Paap, R. (2006). Modeling purchases as repeated events. *Journal of Business & Economic Statistics***24**, 487-502.

Bolland, K. (2003). The Design and Analysis of Neurological Trials Yielding Repeated Ordinal Data. The University of Reading.

Bolland, K., Whitehead, J., Cobo, E., and Secades, J. J. (2009). Evaluation of a sequential global test of improved recovery following stroke as applied to the ICTUS trial of citicoline. *Pharmaceutical Statistics***8**, 136-149.

Box-Steffensmeier, J. M., and Zorn, C. (2002). Duration models for repeated events. *Journal of Politics***64**, 1069-1094.

Cain, K. C., and Lange, N. T. (1984). Approximate case influence for the proportional hazards regression-model with censored-data. *Biometrics***40**, 493-499.

Collett, D. (2003). *Modelling Survival Data in Medical Research*, Second edition. USA: Chapman & Hall/CRC.

Cook, R. J., and Lawless, J. F. (1997). An overview of statistical methods for multiple failure time data in clinical trials - Discussion. *Statistics in Medicine***16**, 841-843.

Cox, D. R. (1972). Regression Models and Life-Tables. *Journal of the Royal Statistical Society Series B-Statistical Methodology***34**, 187-&.

Cox, D. R. (1975). Partial likelihood. *Biometrika***62**, 269-276.

Cox, D. R., and Hinkley, D. V. (1974). *Theoretical statistics*. London: Chapman and Hall.

Cox, D. R. D. O. ( 1984). *Analysis of Survival Data*. London: Chapman and Hall.

Davalos, A. (2007). Protocol 06PRT/3005:ICTUS study: International Citicoline Trial on acute Stroke (NCT00331890) Oral citicoline in acute ischemic stroke. *Lancet Protocol Reviews*.

Davalos, A., Catillo, J., Alvarez-Sabin, J., Secades, J.J., Mercadal, J., Lopez, S., Cobo, E., Warach, S., Sherman, D., Clark, W.M., Lozano, R. . ; (2002). Oral Citicoline in Acute Ischemic Stroke: An Individual Patient Data Pooling Analysis of Clinical Trials. *Stroke***33**, 2850-2857.

Dodge (2003). *The Oxford Dictionary of Statistical Terms*: OUP.

Dupont, W. D. (2009). *Statistical modeling for biomedical researchers*. USA: Cambridge University Press.

Efron, B. (1977). The efficiency of Cox's likelihood function for censored data. *Journal of the American Statistical Association***72**, 557-565.

Efron, B. (1981). Nonparametric estimates of standard error - the jackknife, the bootstrap and other methods. *Biometrika***68**, 589-599.

Efron, B., and Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.

Eisenhauer, E. A., Therasse, P., Bogaerts, J., *et al.* (2009). New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1). *European Journal of Cancer***45**, 228-247.

Finkelstein, D. M. (1986). A proportional hazards model for interval-censored failure time data. *Biometrics***42**, 845-854.

Finkelstein, D. M., Schoenfeld, D. A., and Stamenovic, E. (1997). Analysis of multiple failure time data from an aids clinical trial. *Statistics in Medicine***16**, 951-961.

Fleischer, F., Gaschler-Markefski, B., and Bluhmki, E. (2009). A statistical model for the dependence between progression-free survival and overall survival. *Statistics in Medicine***28**, 2669-2686.

Gaumetou, E., Zilber, S., and Hernigou, P. (2011). Non-simultaneous bilateral hip fracture: Epidemiologic study of 241 hip fractures. *Orthopaedics & Traumatology: Surgery & Research***97**, 22-27.

Genser, B., and Wernecke, K. D. (2005). Joint modelling of repeated transitions in follow-up data - A case study on breast cancer data. *Biometrical Journal***47**, 388-401.

Goggins, W. B., and Finkelstein, D. M. (2000). A proportional hazards model for multivariate interval-censored failure time data. *Biometrics***56**, 940-943.

Gutierrez, E., Lozano, S., and Gonzalez, J. R. (2011). A recurrent-events survival analysis of the duration of Olympic records. *Ima Journal of Management Mathematics***22**, 115-128.

Heitjan, D. F., and Rubin, D. B. (1991). Ignorability and coarse data: some biomedical examples. *The Annals of Statistics***19**, 2244-2253.

Hougaard, P. (1999). Fundamentals of survival data. *Biometrics***55**, 13-22.

Hougaard, P. (2000). *Analysis of multivariate survival data.* New York: Springer.

Huber, P. J. (1967). The Behavior of Maximum Likelihood Estimates under Nonstandard Conditions. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 221–233.

Kalbfleisch, J. D., and Prentice, R. L.(1973). Marginal likelihoods based on coxs regression and life model. *Biometrika***60**, 267-278.

Kaplan, E. L., and Meier, P. (1958). Nonparametric-Estimation from Incomplete Observations. *Journal of the American Statistical Association***53**, 457-481.

Kelly, P. J. (2004). A review of software packages for analyzing correlated survival data. *American Statistician***58**, 337-342.

Kelly, P. J., and Lim, L. L. Y. (2000). Survival analysis for recurrent event data: An application to childhood infectious diseases. *Statistics in Medicine***19**, 13-33.

Kim, M. Y., and Xue, X. N. (2002). The analysis of multivariate interval-censored survival data. *Statistics in Medicine***21**, 3715-3726.

Klein, J. P., and Moeschberger, M. L. (1997). *Survival analysis : techniques for censored and truncated data.* New York: Springer.

Lee, E. W., Wei, L. J., Amato, D. A., and Leurgans, S. (1992). Cox-type regression-analysis for large numbers of small-groups of correlated failure time observations. In *Survival Analysis : State of the Art*, J. P. Klein, and P. K. Goel (eds), 237-247.

Li, Q. H., and Lagakos, S. W. (1997). Use of the Wei-Lin-Weissfeld method for the analysis of a recurring and a terminating event. *Statistics in Medicine***16**, 925-940.

Liang, K. Y., and Zeger, S. L. (1986). Longitudinal data-analysis using generalized linear-models. *Biometrika***73**, 13-22.

Lim, R., Sun, Y., Im, S. A., *et al.* (2011). Cetuximab plus irinotecan in pretreated metastatic colorectal cancer patients: The ELSIE study. *World Journal of Gastroenterology***17**, 1879-1888.

Lin, D. Y., and Wei, L. J. (1989). The robust inference for the cox proportional hazards model. *Journal of the American Statistical Association***84**, 1074-1078.

Lipschutz, K. H., and Snapinn, S. M. (1997). An overview of statistical methods for multiple failure time data in clinical trials - Discussion. *Statistics in Medicine***16**, 846-848.

Lipsitz, S. R., Laird, N. M., and Harrington, D. P. (1990). Using the jackknife to estimate the variance of regression-estimators from repeated measures studies. *Communications in Statistics-Theory and Methods***19**, 821-845.

Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemother. Rep.***50**, 163-170.

McCullagh, P. (1980). Regression-models for ordinal data. *Journal of the Royal Statistical Society Series B-Methodological***42**, 109-142.

Metcalfe, C., and Thompson, S. G. (2006). The importance of varying the event generation process in simulation studies of statistical methods for recurrent events. *Statistics in Medicine***25**, 165-179.

Metcalfe, C., and Thompson, S. G. (2007). Wei, Lin and Weissfeld's marginal analysis of multivariate failure time data: should it be applied to a recurrent events outcome? *Statistical Methods in Medical Research***16**, 103-122.

Miller, R. G. (1974). Jackknife - review. *Biometrika***61**, 1-15.

Nelson, W. B. (2003). *Recurrent Events Analysis for Product Repairs, Disease Recurrences, and Other Applications*. Philadelphia: ASA.

Neuhaus, J. M., Kalbfleisch, J. D., and Hauck, W. W. (1991). A comparison of cluster-specific and population-averaged approaches for analyzing correlated binary data. *International Statistical Review***59**, 25-35.

Ng, E. T. M., and Cook, R. J. (1999). Adjusted score tests of homogeneity for Poisson processes. *Journal of the American Statistical Association***94**, 308-319.

O'Brien, P. (1984). Procedures for Comparing Samples with Multiple Endpoints. *Biometrics***40**, 1079-1087.

Oakes, D. (2001). Biometrika Centenary: Survival analysis. *Biometrika***88**, 99-142.

Peto, R., and Peto, J. (1972). Asymptotically Efficient Rank Invariant Test Procedures. *Journal of the Royal Statistical Society Series a-General***135**, 185-&.

Pocock, S. J., Geller, N. L., and Tsiatis, A. A. (1987). The analysis of multiple endpoints in clinical trials. *Biometrics***43**, 487-498.

Pocock SJ, G. N., and Tsiatis AA. (1987). The Analysis of Multiple Endpoints in Clinical Trials. *Biometrics***43**, 487-498.

Pocock, S. J. *C. T. A. P. A.* (1983). *Clinical Trials. A Practical Approach.* Chichester: Wiley.

Prentice, R. L., and Gloeckler, L. A. (1978). Regression-analysis of grouped survival data with application to breast-cancer data. *Biometrics***34**, 57-67.

Prentice, R. L., Williams, B. J., and Peterson, A. V. (1981). On the regression-analysis of multivariate failure time data. *Biometrika***68**, 373-379.

Quenouille, M. H. (1949). Approximate tests of correlation in time-series-3. *Proceedings of the Cambridge Philosophical Society***45**, 483-484.

Reid, N., and Crepeau, H. (1985). Influence functions for proportional hazards regression. *Biometrika***72**, 1-9.

Roychoudhuri, R., Putcha, V., and Moller, H. (2006). Cancer and laterality: A study of the five major paired organs (UK). *Cancer Causes & Control***17**, 655-662.

Saville, B. R., Herring, A. H., and Koch, G. G. (2010). A robust method for comparing two treatments in a confirmatory clinical trial via multivariate time-to-event methods that jointly incorporate information from longitudinal and time-to-event data. *Statistics in Medicine***29**, 75-85.

Scharfstein, D., Tsiatis, A., and Robins, J. (1997). Semiparametric efficiency and its implication on the design and analysis of group sequential studies. *Journal of the American Statistical Association***92**, 1342-1350.

Storb, R., Deeg, H. J., Whitehead, J., *et al.* (1986). Methotrexate and Cyclosporine Compared with Cyclosporine Alone for Prophylaxis of Acute Graft Versus Host-Disease after Marrow Transplantation for Leukemia. *New England Journal of Medicine***314**, 729-735.

Sun, J. (2006). *The statistical analysis of interval-censored failure time data.* USA: Springer.

Tang, D. I., Gnecco, C., and Geller, N. L. (1989). Design of Group Sequential Clinical-Trials with Multiple Endpoints. *Journal of the American Statistical Association***84**, 776-779.

TenHave, T. R. (1996). A mixed effects model for multivariate ordinal response data including correlated discrete failure times with ordinal responses. *Biometrics***52**, 473-491.

Thall, P. F., and Lachin, J. M. (1988). Analysis of recurrent events - nonparametric methods for random-interval count data. *Journal of the American Statistical Association***83**, 339-347.

Therasse, P., Arbuck, S. G., Eisenhauer, E. A., *et al.* (2000). New guidelines to evaluate the response to treatment in solid Tumors. *Journal of the National Cancer Institute***92**, 205-216.

Therneau, T. M., and Grambsch, P., M. (2000). *Modeling survival data : extending the Cox model.* New York: Springer.

Tilley, B. C., Marler, J., Geller, N. L., *et al.* (1996). Use of a global test for multiple outcomes in stroke trials with application to the national institute of neurological disorders and stroke t-PA stroke trial. *Stroke***27**, 2136-2142.

Tukey, J. W. (1958). Bias and confidence in not-quite large samples. *Annals of Mathematical Statistics***29**, 614-614.

Turnbull, B. W. (1976). The empirical distribution with arbitrarily grouped censored and truncated data. *Journal of the Royal Statistical Society Series B-Methodological***38**, 290-295.

Wei, L. J., and Glidden, D. V. (1997). An overview of statistical methods for multiple failure time data in clinical trials. *Statistics in Medicine***16**, 833-839.

Wei, L. J., and Johnson, W. E. (1985). Combining dependent tests with incomplete repeated measurements. *Biometrika***72**, 359-364.

Wei, L. J., Lin, D.Y. and Weissfeld, L. (1989). Regression Analysis of Multivariate Incomplete Failure Time Data by Modeling Marginal Distributions. *Journal of the American Statistical Association***84**, 1065-1073.

White, H. (1982). Maximum-likelihood estimation of mis-specified models. *Econometrica***50**, 1-25.

Whitehead, J. (1989). The Analysis of Relapse Clinical-Trials, with Application to a Comparison of 2 Ulcer Treatments. *Statistics in Medicine***8**, 1439-1454.

Whitehead, J. (1997). *The design and analysis of sequential clinical trials*, Revised Second Edition edition. Chichester: John Wiley & Sons Ltd.

Whitehead, J., Branson, M., and Todd, S. (2010). A combined score test for binary and ordinal endpoints from clinical trials. *Statistics in Medicine*, 521-532.

Whitehead, J., and Thomas, P. (1997). A sequential trial of pain killers in arthritis: Issues of multiple comparisons with control and of interval-censored survival data. *Biopharmaceutical Statistics***7**, 333-353.

Wroblewski, B. M., and Siney, P. D. (1992). Charnley low-friction arthroplasty in the young patient.*Clinical Orthopaedics and Related Research*, 45-47.

Yan, Y., Andriole, G. L., Humphrey, P. A., and Kibel, A. S. (2002). Patterns of multiple recurrences of superficial (Ta/T1) transitional cell carcinoma of bladder and effects of clinicopathologic and biochemical factors. *Cancer***95**, 1239-1246.