Testing Procedures for Group Sequential Clinical Trials with Multiple Survival Endpoints

Rebecca C. Scherzer
Western Michigan University

Follow this and additional works at: http://scholarworks.wmich.edu/dissertations
Part of the Mathematics Commons

Recommended Citation
Dissertations. 985.
http://scholarworks.wmich.edu/dissertations/985

This Dissertation-Open Access is brought to you for free and open access by the Graduate College at ScholarWorks at WMU. It has been accepted for inclusion in Dissertations by an authorized administrator of ScholarWorks at WMU. For more information, please contact maira.bundza@wmich.edu.
TESTING PROCEDURES FOR GROUP SEQUENTIAL CLINICAL TRIALS WITH MULTIPLE SURVIVAL ENDPOINTS

by

Rebecca C. Scherzer

A Dissertation
Submitted to the
Faculty of the Graduate College
in partial fulfillment of the
requirements for the
Degree of Doctor of Philosophy
Department of Statistics

Western Michigan University
Kalamazoo, Michigan
April 2006
TESTING PROCEDURES FOR GROUP SEQUENTIAL CLINICAL TRIALS WITH MULTIPLE SURVIVAL ENDPOINTS

Rebecca C. Scherzer, Ph.D.

Western Michigan University, 2006

This research gives methods for sequential monitoring of survival data in clinical trials with multiple endpoints. We illustrate the use of marginal proportional hazards models and other survival models with various group sequential methods to test multiple survival endpoints at K interim analyses. To adjust for multiplicity at each interim analysis, we consider and extend methods developed by Tang and Geller (1999), Follmann, et al. (1994), and others. These methods are motivated, compared, and evaluated using survival data from a clinical study and using simulation studies.
ACKNOWLEDGEMENTS

I wish to thank my advisor Dr. Daniel Mihalko for his guidance, patience through multiple revisions, and many helpful suggestions throughout this process. I wish to extend the same thanks to my committee members for their careful review, thoughtful comments, and much time devoted to this project. Much thanks also to many past and present colleagues for their suggestions and moral support. And finally, I am deeply grateful to my family and friends for their encouragement, love, and assistance in so many ways.

Rebecca C. Scherzer
# TABLE OF CONTENTS

ACKNOWLEDGEMENTS........................................................................................................... ii  
LIST OF TABLES................................................................................................................... viii  
LIST OF FIGURES ................................................................................................................ x  

CHAPTER  

I. INTRODUCTION.................................................................................................................... 1  

II. SURVIVAL ANALYSIS........................................................................................................ 4  

  2.1 Introduction.................................................................................................................. 4  

  2.2 Kaplan-Meier Estimator and Log-Rank Statistic......................................................... 6  

  2.3 Cox Proportional Hazards Regression........................................................................ 8  

  2.4 Multivariate Methods............................................................................................... 9  

    2.4.1 Accelerated Failure Time Models....................................................................... 10  

    2.4.2 Frailty Models.................................................................................................... 11  

    2.4.3 Additive Hazard Models.................................................................................... 12  

    2.4.4 Marginal Models............................................................................................... 12  

III. GROUP SEQUENTIAL METHODOLOGY.................................................................... 14  

  3.1 Introduction and Theoretical Assumptions............................................................... 14  

  3.2 Early Work and Overview......................................................................................... 17  

    3.2.1 Pocock................................................................................................................ 17  

    3.2.2 O'Brien-Fleming ............................................................................................... 20  

    3.2.3 Lan-DeMets...................................................................................................... 21  

    3.2.4 Wang and Tsiatis............................................................................................. 24
Table of Contents—Continued

CHAPTER

3.2.5 Haybittle-Peto ................................................................. 25
3.2.6 Discussion ................................................................. 26

3.3 GS Methods for Survival Data .................................................. 26
3.3.1 Stratified and Unstratified Log-Rank Test ....................... 27
3.3.2 Error Spending Approach ......................................... 29
3.3.3 Cox Proportional Hazards Regression ......................... 30
3.3.4 Alternatives to the Cox Proportional Hazards Model ....... 31

3.4 GS Methods for Multivariate Data ........................................ 31
3.4.1 Composite Scores ...................................................... 31
3.4.2 Bonferroni Procedure ............................................... 32
3.4.3 Group Sequential Hotelling Test .................................. 32
3.4.4 O'Brien's Global Tests ............................................... 33
3.4.5 Closed Testing Procedures ........................................... 34
3.4.6 Bonferroni-Holm Method ............................................ 36
3.4.7 Hochberg Method .................................................... 38

IV. GROUP SEQUENTIAL METHODOLOGY FOR MULTIVARIATE SURVIVAL ANALYSIS ................................................................. 42

4.1 Introduction ................................................................. 42
4.2 Types of Multivariate Survival Data ...................................... 42
4.2.1 Distinct, Unordered Events ........................................ 43
4.2.2 Distinct, Ordered Events .......................................... 43
4.2.3 Recurring Events .................................................... 43
Table of Contents—Continued

CHAPTER

4.2.4 Competing Risks................................................................. 44

4.3 Analyzing Multivariate Survival Data in the GS Setting........ 44

4.3.1 Marginal Hazards Models (WLW variant)......................... 44

4.3.2 Accelerated Failure Time (AFT) Models ......................... 48

4.3.3 Frailty Models................................................................. 50

4.3.4 Closed Testing Methods .................................................. 53

4.3.5 Bonferroni-Holm and Hochberg Methods......................... 54

4.3.6 Fisher's LSD Analogue..................................................... 56

4.4 Summary of Methods Proposed............................................ 58

V. DATA EXAMPLES........................................................................... 59

5.1 Introduction.............................................................................. 59

5.2 Analysis Results using Closed Testing Method (CTM) ........ 60

5.2.1 Lan-DeMets Error Spending Function Simulating O'Brien-Fleming Shape with WLW Method ........................................ 61

5.2.2 Lan-DeMets Error Spending Function Simulating Pocock Shape with WLW Method ..................................................... 63

5.2.3 Lan-DeMets Error Spending Function Simulating O'Brien-Fleming Shape with AG Method .............................................. 64

5.2.4 Lan-DeMets Error Spending Function Simulating O'Brien-Fleming Shape with PWP Method ............................................ 65

5.2.5 Lan-DeMets Error Spending Function Simulating O'Brien-Fleming Shape with AFT Method............................................. 65

5.2.6 Lan-DeMets Error Spending Function Simulating O'Brien-Fleming Shape with Frailty Method................................. 67
Table of Contents—Continued

CHAPTER

5.3 Analysis Results using Bonferroni-Holm Method (BHM) ...................... 68
  5.3.1 Lan-DeMets Error Spending Function Simulating O'Brien-Fleming Shape with WLW Method ........................................ 69
  5.3.2 Lan-DeMets Error Spending Function Simulating Pocock Shape with WLW Method ...................................................... 70
  5.3.3 Lan-DeMets Error Spending Function Simulating O'Brien-Fleming Shape with AFT Method ........................................... 71
  5.3.4 Lan-DeMets Error Spending Function Simulating O'Brien-Fleming Shape with Frailty Method ........................................ 72

5.4 Analysis Results using Hochberg Method ........................................ 73
  5.4.1 Lan-DeMets Error Spending Function Simulating O'Brien-Fleming Shape with WLW Method ........................................ 73
  5.4.2 Lan-DeMets Error Spending Function Simulating Pocock Shape with WLW Method ...................................................... 75
  5.4.3 Lan-DeMets Error Spending Function Simulating O'Brien-Fleming Shape with AFT Method ........................................... 76
  5.4.4 Lan-DeMets Error Spending Function Simulating O'Brien-Fleming Shape with Frailty Method ........................................ 77

5.5 Analysis Results using Fisher's LSD Analogue ................................. 78
  5.5.1 Lan-DeMets Error Spending Function Simulating O'Brien-Fleming Shape with WLW Method ........................................ 78
  5.5.2 Lan-DeMets Error Spending Function Simulating Pocock Shape with WLW Method ...................................................... 79
  5.5.3 Lan-DeMets Error Spending Function Simulating O'Brien-Fleming with AFT Method .................................................... 79
  5.5.4 Lan-DeMets Error Spending Function Simulating O'Brien-Fleming with Frailty Method .................................................. 80
Table of Contents—Continued

CHAPTER

5.6 Discussion and Summary of Results ............................................................... 81

VI. DATA SIMULATIONS .................................................................................. 84

6.1 General Description of Data Simulation for PBC Data ............................. 84
6.2 Data Simulation Results for PBC Data under Null Hypothesis .............. 90
   6.2.1 O'Brien-Fleming and Pocock Shape with Various Censoring Levels ................................................................. 90
6.3 Data Simulation Results for PBC Data under Null Hypothesis with Non-Constant Correlation Structures ................................................................. 98
6.4 Simulations under Alternative Hypotheses .............................................. 103
6.5 Data Simulation Results for Fisher’s LSD Analogue ............................... 112
6.6 Data Simulation Results for BHM vs Hochberg Methods ...................... 119

VII. SUMMARY, CONCLUSIONS, AND DIRECTIONS FOR FUTURE RESEARCH ........................................................................................................... 126

7.1 Summary of Methodology Developed in This Dissertation .................. 126
7.2 Discussion of Results ............................................................................... 127
7.3 Conclusions and Directions for Further Research ................................. 129

APPENDICES

A. Derivations and Theory ............................................................................. 132
B. Data Sources ............................................................................................. 134
C. SAS and Splus Code and Output for Data Examples and Methodology .... 136
D. SAS Code for Data Simulations ................................................................. 154

BIBLIOGRAPHY .............................................................................................. 171
LIST OF TABLES

3.1 Comparison of Selected Group Sequential Testing Methods ......................... 27
3.2 Boundary Values for Interim Analyses using Error Spending Function .......... 30
5.1 Number of Subjects Experiencing Events of Interest in PBC Study .............. 60
5.2 Boundary Values for Interim Analyses using Lan-DeMets Spending Function with O’Brien-Fleming Shape for PBC Data .................................................. 61
5.3 Boundary Values for Interim Analyses using Lan-DeMets Spending Function with Pocock Shape for PBC Data ................................................................. 63
5.4 Fisher’s LSD Analogue Results for WLW Method for PBC Data, using Lan-DeMets Spending Function with O’Brien-Fleming Shape ............................... 78
5.5 Fisher’s LSD Analogue Results for WLW Method for PBC Data, using Lan-DeMets Spending Function with Pocock Shape ................................................... 79
5.6 Fisher’s LSD Analogue Results for AFT Method for PBC Data, using Lan-DeMets Spending Function with O’Brien-Fleming Shape .................................... 80
5.7 Fisher’s LSD Analogue Results for Frailty Method for PBC Data, using Lan-DeMets Spending Function with O’Brien-Fleming Shape .................................. 81
5.8 Summary of Results using PBC Data .............................................................. 82
5.9 Summary of Results at Second Interim Analysis for each Hypothesis using PBC Data ............................................................................................................. 83
6.1 Simulation under Null Hypothesis for WLW Method with Correlation rho = 0.0 .................................................................................................................. 92
6.2 Simulation under Null Hypothesis for WLW Method with Correlation rho = 0.25 ................................................................................................................ 93
6.3 Simulation under Null Hypothesis for WLW Method with Correlation rho = 0.50 ................................................................................................................ 94
6.4 Simulation under Null Hypothesis for WLW Method with Correlation rho = 0.75 ................................................................................................................ 95
<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.5</td>
<td>Simulation under Null Hypothesis for WLW Method with O’Brien-Fleming Shape and Non-Constant Correlation Structures</td>
<td>101</td>
</tr>
<tr>
<td>6.6</td>
<td>Simulation under Null Hypothesis for WLW Method with Pocock Shape and Non-Constant Correlation Structures</td>
<td>102</td>
</tr>
<tr>
<td>6.7</td>
<td>Simulation under Alternative Hypotheses for WLW Method with Correlation ( \rho = 0.0 )</td>
<td>104</td>
</tr>
<tr>
<td>6.8</td>
<td>Simulation under Alternative Hypotheses for WLW Method with Correlation ( \rho = 0.25 )</td>
<td>105</td>
</tr>
<tr>
<td>6.9</td>
<td>Simulation under Alternative Hypotheses for WLW Method with Correlation ( \rho = 0.50 )</td>
<td>106</td>
</tr>
<tr>
<td>6.10</td>
<td>Simulation under Alternative Hypotheses for WLW Method with Correlation ( \rho = 0.75 )</td>
<td>107</td>
</tr>
<tr>
<td>6.11</td>
<td>Simulation of Fisher’s for WLW Method with O’Brien-Fleming Shape and Correlation ( \rho = 0.0 )</td>
<td>113</td>
</tr>
<tr>
<td>6.12</td>
<td>Simulation of Fisher’s for WLW Method with O’Brien-Fleming Shape and Correlation ( \rho = 0.25 )</td>
<td>114</td>
</tr>
<tr>
<td>6.13</td>
<td>Simulation of Fisher’s for WLW Method with O’Brien-Fleming Shape and Correlation ( \rho = 0.50 )</td>
<td>115</td>
</tr>
<tr>
<td>6.14</td>
<td>Simulation of Fisher’s for WLW Method with O’Brien-Fleming Shape and Correlation ( \rho = 0.75 )</td>
<td>116</td>
</tr>
<tr>
<td>6.15</td>
<td>Simulation of BHM versus Hochberg for WLW Method with O’Brien-Fleming Shape and Correlation ( \rho = 0.0 ) under ( H_a^* : \Delta = a )</td>
<td>120</td>
</tr>
<tr>
<td>6.16</td>
<td>Simulation of BHM versus Hochberg for WLW Method with O’Brien-Fleming Shape and Correlation ( \rho = 0.25 ) under ( H_a^* : \Delta = a )</td>
<td>121</td>
</tr>
<tr>
<td>6.17</td>
<td>Simulation of BHM versus Hochberg for WLW Method with O’Brien-Fleming Shape and Correlation ( \rho = 0.50 ) under ( H_a^* : \Delta = a )</td>
<td>122</td>
</tr>
<tr>
<td>6.18</td>
<td>Simulation of BHM versus Hochberg for WLW Method with O’Brien-Fleming Shape and Correlation ( \rho = 0.75 ) under ( H_a^* : \Delta = a )</td>
<td>123</td>
</tr>
</tbody>
</table>
LIST OF FIGURES

2.1 Survival Distribution........................................................................................... 6

3.1 Pocock Z-Critical Values .................................................................................... 19

3.2 O’Brien-Fleming Z-Critical Values .................................................................... 21

3.3 Closed Testing Schematic ................................................................................... 35

6.1 Histogram of Lognormally Distributed Data for One Event Time..................... 91

6.2 Cumulative Rejection Rate under Null Hypothesis with O’Brien-Fleming Shape ........................................................................................................ 96

6.3 Cumulative Rejection Rate under Null Hypothesis with Pocock Shape .......... 97

6.4 Cumulative Rejection Rate under Null Hypothesis with O’Brien-Fleming Shape and Non-Constant Correlation Structure .................................................. 99

6.5 Cumulative Rejection Rate under Null Hypothesis with Pocock Shape and Non-Constant Correlation Structure ................................................................... 100

6.6 Rejection Rate at Interim Analysis 1 under $H_a: \Delta = a$ with O’Brien-Fleming Shape ................................................................................................................... 108

6.7 Cumulative Rejection Rate at Final Analysis under $H_a: \Delta = a$ with O’Brien-Fleming Shape ........................................................................................................ 109

6.8 Rejection Rate at Interim Analysis 1 under $H_a: \Delta = a$ with Pocock Shape ....... 110

6.9 Cumulative Rejection Rate at Final Analysis under $H_a: \Delta = a$ with Pocock Shape ........................................................................................................ 111

6.10 Fisher’s LSD Rejection Rate of at Least One Endpoint at Interim Analysis 1 with O’Brien-Fleming Shape ................................................................. 117

6.11 Fisher’s LSD Cumulative Rejection Rate of at Least One Endpoint at Final Analysis with O’Brien-Fleming Shape ............................................................. 118
List of Figures—Continued

6.12 BHM vs Hochberg: Rejection Rate of at Least One Endpoint at Interim Analysis 1 with O’Brien-Fleming Shape .......................................................... 124

6.13 BHM vs Hochberg: Cumulative Rejection Rate of at Least One Endpoint at Final Analysis with O’Brien-Fleming Shape ........................................ 125
CHAPTER I
INTRODUCTION

In many clinical trials involving survival outcomes and sequential monitoring, there may be multiple time-to-event endpoints of interest. For example, in a long-term cardiac study, outcomes of interest may include time to major adverse cardiovascular events such as heart attack, stroke, myocardial infarction, death, etc. In such studies a monitoring committee (sometimes referred to as a Data Safety Monitoring Board, or DSMB) is often established to meet on a periodic basis to determine whether the study should be stopped early due to lack of efficacy or a poor safety profile, or whether the study should be allowed to continue. Clearly in such a setting, adjustment must be made both for multiplicity in the endpoint and for interim monitoring, and one must account for possible correlation between the outcomes of interest.

Different approaches for handling multiple failure time data in a non-group sequential setting include methods developed by Wei, Lin, and Weissfeld (1989), Wei and Glidden (1997), Spiekerman and Lin (1998), and others. Wei et al developed marginal proportional hazards distributions, which were later extended by Spiekerman and Lin to allow different hazard functions for distinct events or a common hazard function for recurrent events of the same type. Wei and Glidden provide an overview of methodology for multiple failure time data in the Cox regression setting.

Follmann et al (1994) first discussed the analysis of more than two treatment arms in the context of a trial requiring group sequential monitoring. Their discussion was restricted to pairwise treatment comparisons using a Dunnett or Tukey procedure with modified
versions of Pocock, O’Brien and Fleming, and Lan and Demets group sequential methods.

More recently, Tang and Geller (1999) presented approaches for conducting closed testing in a study with multiple non-survival outcomes in a group sequential setting. Their approach allows flexible stopping times but is limited by the availability of a suitable testing method for the hypotheses to be tested and by the availability of appropriate group sequential critical values.

Methodology for multiple survival time data in a group sequential setting has received far less attention. Williams (1996) discusses various strategies for constructing joint repeated confidence intervals for hazard ratios, and illustrates this using a study with two endpoints with two interim looks. Gangnon et al (1999) also consider an example with two endpoints and two interim looks. Their approach involves splitting the alpha level between the two endpoints and applying an O’Brien-Fleming boundary (1979) to each interim analysis. Their approach relies on simulating boundaries to account for correlation between endpoints. For more than two endpoints they propose a hierarchical adjustment to the boundaries by ordering the endpoints and simulating the boundaries in a sequential fashion while accounting for correlation of earlier endpoints.

In this dissertation we extend some of the methodology reviewed above and describe various strategies for testing between treatment differences in various situations involving multiple survival outcomes in a group sequential setting. In chapters two and three we review the relevant literature for survival analysis and group sequential methodology, respectively. In chapter four we discuss the application of survival methods to the group
sequential setting and discuss their extension to multivariate survival data. In chapter five, data examples are included to illustrate the implementation of some of these procedures and to facilitate their comparison. In chapter six, simulation study results are included to determine which procedures “behave” best in terms of type I error rates, which procedures are most powerful, and which procedures may be most appropriate in various settings and under various conditions. Benefits and drawbacks to these competing procedures are also presented. These results are discussed to determine whether theoretical expectations were met. Finally, an overall discussion is included along with general findings, conclusions, and suggestions for further research.
2.1 Introduction

In this chapter we provide a brief review of the survival analysis theory and popular methodology that will serve as a necessary foundation in later chapters, as we discuss the analysis of multivariate survival data in the context of the group sequential design.

Survival data is the terminology used to describe time-to-event data, often time until death of a patient or failure of a part, or time until some change or transition has occurred. For this paper we are interested specifically in the methodology used to analyze multivariate survival data, the situation in which a subject may experience different types of events or repeated events during the course of a trial. We would like to be able to make use of all of these events (rather than just a single outcome for every patient) so that we will have more power to detect a difference or more information in order to decide whether to end a trial early.

In a clinical trial, we often find that the data available to us has been truncated or censored. It will be of particular use to us later in this paper to utilize the expression “right-censored”, which refers to the case in which the trial has stopped before the event (onset of disease, death, etc.) has occurred for a given subject, or the case in which a subject has dropped out of the trial before the event occurred. Left-censoring, interval-censoring, and truncation are encountered less frequently in the sorts of data we will consider in this paper. However, the reader will find these terms well-described
throughout the literature (Therneau and Grambsch (2000), Klein and Moeschberger
(1997), etc.).

In working with survival data, we usually consider the survival function, $S(t)$, which can
be interpreted as the probability of survival beyond time $t$. $S(t)$ is defined as one minus
the cumulative distribution function (cdf), or:

$$1 - F(t) = 1 - P(T \leq t) = 1 - \int_{-\infty}^{t} f(u) \, du = P(T > t) = \int_{t}^{\infty} f(u) \, du,$$

for $t \geq 0$

Note that this is a monotonically decreasing function of time, as illustrated below in
Figure 2.1, which indicates that the probability of survival gradually decreases as time
increases. The shape of this survival curve can vary greatly depending on the nature of
underlying hazard, or risk, but the median survival time $t_{0.5}$ can always be found by
locating the point at which the survival function $S(t)$ has $P(T > t)$ equal to 0.5. The hazard
function, in turn, gives the instantaneous failure or event rate, and is defined as:

$$h(t) = \lim_{\Delta \to 0} \frac{P(t \leq T \leq t + \Delta \mid T > t)}{\Delta} = \frac{f(t)}{S(t)}, \text{ for } t \geq 0,$$

where $f(t)$ denotes the probability density function (pdf).
2.2 Kaplan-Meier Estimator and Log-Rank Statistic

One method of estimating the survival function is to use the Kaplan-Meier estimator, which can be written as:

\[ \hat{S}(t) = \prod_{j} [1 - \frac{o_j}{n_j}], \text{ for } j=1, \ldots, m \text{ timepoints} \]

where \( o_j \) = number of events at time \( j \), \( n_j \) = number still at risk of experiencing the event at time \( j \). Note that the number \( n_j \) excludes from the risk set subjects who have been censored prior to time \( j \) as well as subjects who have already experienced the event prior to that time.
The log-rank test statistic can be written in a variety of equivalent ways. One very intuitive form is:

\[
S = \left( \sum_{j} (o_{1j} - e_{1j}) \right)^2 \sum_{j} v_{1j} \sim \chi^2(1), \text{ for } j = 1, \ldots, m
\]

where \( o_{1j} \) denotes the number of events at time \( j \) in group 1, and \( e_{1j} \) and \( v_{1j} \) denote the expected value and variance of \( o_{1j} \), respectively; i.e., \( e_{1j} = E(o_{1j}) = \frac{n_{1j} o_{ij}}{n_j} \), and

\[
v_{1j} = V(o_{1j}) = \frac{n_{1j} n_{2j} o_{ij} (n_j - o_{ij})}{n_j^2 (n_j - 1)}.\]

The log-rank test is most powerful when the hazard rates (and hence also the survival curves) for the two treatment groups are proportional, i.e., when:

\[
(2.5) \lambda_1(t) = c \cdot \lambda_2(t),
\]

where \( \lambda_i \) denotes the hazard function for treatment \( i \). The \( n \)-sample log-rank test statistic can be extended from (2.4) and written as:

\[
(2.6) S = \sum_{i=1}^{n} \frac{(o_i - e_i)^2}{e_i} \sim \chi^2(n-1), \text{ for treatment group } i = 1, \ldots, n, \text{ where } o_i \text{ indicates the observed number of events and } e_i \text{ indicates the expected number of events; } e_i \text{ is calculated as } n_{ij} \cdot (o_{ij} / n_j), \text{ where } n_{ij} \text{ is the number of subjects at risk in treatment group } i \text{ at time } j, \text{ and } o_{ij} \text{ and } n_j \text{ are as stated above. It is also possible to form a weighted log-rank statistic, for example, the Gehan-Breslow test (Gehan, 1965 and Breslow, 1974) can place more weight on earlier failure times, making it easier to detect differences.}
\]
between treatment groups in the survival distribution, and reducing the influence of the variability of later failure times to avoid instability. We shall not utilize this method in this dissertation, however, so we do not present the weighted form of the log-rank statistic here.

2.3 Cox Proportional Hazards Regression

Most commonly used in survival analysis is the Cox proportional hazards regression model, which models the hazard for subject $i$ as:

$$\lambda_i(t) = \lambda_0(t) \exp\{\beta' Z_i(t)\}, \ t \geq 0$$

where $\lambda_0(t)$ denotes the baseline hazard function, $Z_i$ is a vector of covariates for subject $i$, and $\beta'$ is a vector of the coefficients or parameters to be estimated. In Chapter 4 of this dissertation, we shall see how this model can be rewritten to account for multiple endpoints of interest.

One key assumption when using the Cox model is the proportional hazards assumption, i.e., the hazard function of the groups of interest should vary only by a multiplicative constant, as in (2.5) above. In other words, the hazard ratio should be constant and should not vary over time. There are various ways to test this assumption, including graphical checks (e.g., using comparative Kaplan-Meier plots and plotting negative log of the hazard ratios by time) and more formal goodness-of-fit tests.

When the proportional hazards assumption is violated, we may need to incorporate time-dependent covariates to the model or stratify by other covariates, or we may need to
consider alternatives such as the accelerated failure model, discussed in Section 2.4.1 and in Chapter 4.

2.4 Multivariate Methods

Multiple survival outcomes can fall into different categories, including the following: unordered and distinct events, recurrences of the same event type, and ordered distinct events. Therneau and Grambsch (2000) review four major categories of handling multiple events per subject:

1. Time to first event, ignoring multiplicity
2. Assume multiple events are independent, conditional on a per-subject coefficient (frailty model)
3. Marginal models approach
4. Model subject’s correlation within Cox framework

Approach (1) is equivalent to forming a composite event, i.e., time to any one of M events is the same as time to first event. Since this approach reduces to a univariate problem and is already well-covered in the literature, we will not focus on this. For this dissertation, we will begin with approach (3) since it is easily implemented with commonly available statistical packages such as SAS PHREG and Splus coxph (as illustrated in more detail in Chapter 5).

In the marginal models approach, we handle multiple events of distinct types (discussed below in an example) in a Cox proportional hazards regression by stratifying by event type and allowing each event type to have its own baseline hazard function. We can also include in the model baseline covariates and treatment effects which are common across
the strata. In several data examples in Chapter 5, we use the robust estimate of variance described by Wei et al (1989).

Wei & Glidden (1997) discuss some of the above methods, and also suggest combining the endpoints using a linear combination, assuming they are scaled in the same direction. Other options include weighting some endpoints more heavily due, for example, to greater severity, which can be accomplished quite easily programmatically.

We shall describe this method in much more detail in Chapter 4 when we show how the methods in this section can be utilized in a group sequential framework, and then how these methods can be used in a *multivariate* group sequential setting.

### 2.4.1 Accelerated Failure Time Models

An alternative to the Cox Proportional Hazards model may be desirable if the proportionality assumption does not hold. One such alternative is the Accelerated Failure Time (AFT) model, which regresses the log-transformed survival times rather than the actual survival times against treatment group and any other covariates. The term “accelerated” is used to denote the fact that covariates included in the model may shift the timing of when the event of interest may occur. In a univariate case (i.e., only one event of interest), the model can be written as:

\[
\log T_i = \beta_0 + X_{i1} \beta_1 + X_{i2} \beta_2 + \ldots X_{iM} \beta_M + \varepsilon_i
\]

where \( \beta_0 \) is the intercept and \( \beta_m \) is the coefficient for the \( m^{th} \) covariate. The error terms \( \varepsilon_i \) are i.i.d. and can be from a variety of parametric models (such as gamma or normal) or nonparametric (which is not as well studied yet per Wei, Ling, and Ying 1990, but...
revisited in the recurrent event setting in Lin, Wei, and Ying (1998). One advantage of this model is that it does not present the problem of neglected covariates. But according to Wei and Glidden (1997), “it is not clear how this multivariate AFT model can handle the case when there are competing risks.” Another disadvantage, per Hougaard (1999), is that “[AFT] is less good in allowing for time-dependent covariates and for arbitrary hazard functions.”

We shall see in later chapters how to apply this model to the group sequential case as well as to the multivariate case.

2.4.2 Frailty Models

Typically survival models assume that distinct individuals will have survival times that are independent of one another. This assumption can be violated, for example, in littermates, who share a common environment or genetic background, or in the case of a single individual with multiple, distinct failure times (our area of special interest later in this dissertation). The frailty model, which can be viewed as an extension of the Cox proportional hazards model, includes a continuous predictor modelled as a random effect for each subject to describe the risk (“frailty”) for that particular subject, or for all subjects within a subgroup. The model can be written in various ways, including as:

\[ \lambda_{ij}(t) = \lambda_0(t) u_i \exp(\beta'Z_{ij}), \quad i = 1, ..., G; \quad j = 1, ..., n_i \]

where \( \lambda \) denotes the hazard, \( i \) denotes a subject with \( n_i \) events (or an individual member of a subgroup, such as littermates), \( u_i \) denotes the random frailty effect within each subject, \( G \) is the number of subjects, and \( j \) denotes the event indicator within subject.
Essentially this form of the model is just like the Cox model, but with frailty $u_i$ incorporated multiplicatively into the hazard.

We shall explore in Chapter 4 how this can be applied in the group sequential setting to multivariate survival data.

### 2.4.3 Additive Hazard Models

An alternative to the case where the hazard rate is affected in a multiplicative fashion by some covariate is the class of additive hazard models, which can be expressed as:

\[
\lambda(t) = \beta_0(t) + \sum_{k=1}^{p} \beta_k(t)Z_k(t)
\]

where $\beta_k(t)$ is a function of time for each covariate, for $k = 1, \ldots, p$ covariates. The $\beta_k(t)$ function can be estimated using a least-squares non-parametric technique, as illustrated in Klein and Moeschberger (1997). This differs from the likelihood approach used in estimation for multiplicative models such as Cox proportional hazards regression. Since the additive hazards model is designed for cases where there are multiple covariates, rather than multiple outcomes, we do not consider this method further in this dissertation.

### 2.4.4 Marginal Models

In this dissertation, we shall be especially interested in the marginal Cox proportional hazards model first discussed by Wei, Lin, and Weissfeld (1989), which allows modelling of distinct as well as recurrent multivariate failures without assuming or imposing any specific dependence structure between the multiple outcomes. Briefly, the data are fit using an ordinary Cox proportional hazards model, ignoring possible
dependencies between the multiple outcomes. Then, a correction is applied by calculating a robust sandwich estimate of the variance to account for within-subject correlation of the multiple events. The resulting test statistics are then calculated using this robust estimate of variance rather than the naïve initial estimates that assume no dependence structure. We shall reserve our discussion of this method for Section 4.3.1, in which we discuss in detail its application to the group sequential setting.
CHAPTER III

GROUP SEQUENTIAL METHODOLOGY

3.1 Introduction and Theoretical Assumptions

Group Sequential (GS) methodology refers to the statistical methods that are employed in studies in which interim analyses are planned, i.e., analyses that occur while data are still being collected or while a study is still ongoing. In this and in subsequent chapters, we shall most commonly discuss group sequential methods in the context of clinical trials, but the reader should note that these methods also have application in industrial, epidemiological, and other settings. In such studies, multiplicity adjustments must be made due to the interim monitoring so that the type I error rate is not inflated.

As a simple example, we might consider a study in which a single continuous endpoint of interest, blood pressure reduction, is measured. Suppose that half of all enrolled patients are administered a new active treatment, and the remaining half are given an older standard or conventional treatment. In a fixed sample study (i.e., a study in which no interim monitoring is conducted), we simply wait until the conclusion of the study and conduct a two-sample t-test, an F-test, a Z-test, or some non-parametric analogue, depending on how large the sample size is and whether we wish to adjust for predictors or covariates.

For simplicity, suppose that the sample size is large and we have decided to conduct a Z-test of $H_0: \mu_A = \mu_S$ against a two-sided alternative:
(3.1) \[ Z = \frac{\sum X_i - \sum X_j}{\sigma \sqrt{2n}} \sim N(0,1) \text{ under Ho}, \]
where \( X_i \) denotes blood pressure reductions in group A (active), \( X_j \) denotes responses in group S (standard treatment), and assuming equal variance in the two groups and common group sample size \( n \). As usual, we will reject the null hypothesis if \( Z \) exceeds the critical value for our prespecified alpha level (e.g., \(|Z| \geq 1.96\) at \( \alpha = 0.05 \)).

However, if we decide instead to include interim monitoring and we “look” at the data one or more times before the study is complete, then we must pay a penalty in the type I error rate, i.e., we cannot utilize an alpha level of 0.05. The practical consequence of this is that since our z-critical value must be larger than 1.96, it will be harder to reject the null hypothesis. This is obviously as it should be, since we would not want to make a decision about the conduct of the study (e.g., to discontinue early due to lack of efficacy) based on too little data collected or without strong enough evidence.

However, if we are considering multiple looks at the data, then we will have to construct a separate test statistic at each interim look. Then, we must consider what sort of assumptions are needed to make a statement about any sort of joint distribution of the test statistics.

Suppose there are a total of \( K \) planned interim analyses, \( k = 1, \ldots, K \). Consider first the \textit{univariate, continuous} case. For a situation in which there are two treatment groups, A and B, suppose we are interested in testing:

(3.2) \[ \text{Ho: } \mu_A = \mu_B \text{ versus Ha: } \mu_A > \mu_B, \text{ which is equivalent to:} \]

\[ \text{Ho: } \mu_A - \mu_B = \delta = 0 \text{ versus Ha: } \delta > 0. \]
Let $X_i$ denote responses from group A, and $X_j$ denote responses from group B, and suppose $X_i \sim N(\mu_A, \sigma_a^2)$ and $X_j \sim N(\mu_B, \sigma_b^2)$.

Then $\delta = x_i - x_j \sim N(\delta, \sigma_a^2/n_a + \sigma_a^2/n_b)$.

We define the information for $\mu_A - \mu_B$ as: $I_k = (\sigma_a^2/n_a + \sigma_a^2/n_b)^{-1}$

Let $\hat{\delta}^{(k)}$ denote the estimate of $\delta$ at analysis $k$.

Then (per Jennison & Turnbull, 1997, JASA) the sequence $\{\hat{\delta}^{(1)}, \hat{\delta}^{(2)}, \ldots, \hat{\delta}^{(K)}\}$ has a canonical joint distribution, satisfying:

(3.3) $\hat{\delta}^{(1)}, \hat{\delta}^{(2)}, \ldots, \hat{\delta}^{(K)}$ are multivariate normal,

$\hat{\delta}^{(k)} \sim N(\delta, 1/I_k)$ for each $k$,

and $\text{Cov}(\hat{\delta}^{(k1)}, \hat{\delta}^{(k2)}) = 1/I_{k2}$ for $1 \leq k_1 \leq k_2 \leq K$

Note that this holds because the sequence $\{\hat{\delta}^{(1)}, \hat{\delta}^{(2)}, \ldots, \hat{\delta}^{(K)}\}$ is Markov:

i.e., the conditional density $f$ of $\hat{\delta}^{(k)}$ given the previously observed $\hat{\delta}^{(1)}, \ldots, \hat{\delta}^{(k-1)}$, can be written as $f(\hat{\delta}^{(k)} | \hat{\delta}^{(1)}, \ldots, \hat{\delta}^{(k-1)}) = f(\hat{\delta}^{(k)} | \hat{\delta}^{(k-1)})$. Per Jennison and Turnbull (1997), it is evident from (3.3) that $\hat{\delta}^{(i)}$ is independent of $\hat{\delta}^{(i)} - \hat{\delta}^{(i)}$ for any $j > i$, where $1 \leq i < j \leq K$.

We can also state this same property in terms of standardized test statistics: For $Z_1 \sim N(\delta \sqrt{I_1}, 1)$, and a sequence of information levels $I_1, \ldots, I_K$, let $\Delta_k = I_k - I_{k-1}$ (for $k=2, \ldots, K$). It follows from (3.3) that:

(3.4) $Z_k \sqrt{I_1} - Z_{k-1} \sqrt{I_1} \sim N(\delta \Delta_k, \Delta_k)$, independently of $Z_1, \ldots, Z_{k-1}$. 

16
This idea of “independent increments” is an important property since it will allow us to compute GS boundaries by recursive integration rather than by multivariate normal integration. Additional computational and theoretical details are provided in Chapter 19 of Jennison and Turnbull (2000).

The sections that follow review key group sequential results which will be necessary to understand our current research, and illustrate just how one might adjust the alpha level in order to maintain the overall type I error rate.

3.2 Early Work and Overview

Pioneering work in sequential methods for medical studies is generally considered to have started with Armitage (1954, 1969, 1975) and Bross (1952), although modern group sequential methods and research usually build upon the work begun by Pocock (1977) and O’Brien & Fleming (1979). Group sequential methods differ from strictly sequential methods in that in the latter, analyses occur after every single observation, whereas in the former, analyses occur generally at a smaller number of intervals. These papers from the late 1970s laid out guidelines for the first time for Type I error rate and power calculations for studies in which interim monitoring is planned. The work of Pocock and O’Brien & Fleming, along with other early foundational papers, is described in the sections that follow.

3.2.1 Pocock

The idea behind Pocock’s (1977) approach is to use the same nominal alpha level $\alpha'$ at each interim analysis, where $\alpha' < \alpha$, to maintain an overall type I error rate of $\alpha$. 
Pocock’s method requires equally-spaced interim analyses, and requires that the number of interim analyses be specified in advance. Assuming that two-sided testing is planned, at each interim analysis \( k \), the null hypothesis \( H_0: \mu_A = \mu_B \) is rejected in favor of an alternative of \( H_a: \mu_A \neq \mu_B \) if the significance level is less than \( \alpha' = 2[1-\Phi(C_k)] \). The nominal alpha level \( \alpha' \) (and its corresponding critical value \( C_k \)) is calculated using numerical integration using the joint distribution of standardized statistics \( Z_1, ..., Z_K \), where \( K \) denotes the number of interim looks, as usual. Recall from Section 3.1 that \( Z_1, ..., Z_K \) have a canonical joint distribution and have independent increments. We can then say that:

\[
(3.5) \quad Z_1 \sim N(\theta\sqrt{I_1}, 1) \quad \text{and} \quad \\
(3.6) \quad Z_k\sqrt{I_k} - Z_{k-1}\sqrt{I_{k-1}} \sim N(\theta(I_k - I_{k-1}), (I_k - I_{k-1})) \text{ for any subsequent } k \leq K.
\]

Since the increments are equally sized as well as independent, a constant \( \alpha' \) can be calculated for each interim analysis. In practice the calculation of \( \alpha' \) is done automatically using any of a variety of statistical packages (e.g., EAST, Splus, PASS) or by referencing tables, rather than by writing code to perform numerical integration. Pocock (1977, 1982) provides tabled values for several choices of overall alpha and \( K \) (the total number of interim looks), and references Armitage et al (1969) for details of the calculations. In Armitage et al (1969), numerical quadrature was applied using Simpson’s rule and Newton-Cotes formula to calculate the joint distribution using weighted sums in place of the integrals.

As an example, consider a case where four interim looks are planned with an overall alpha level of 0.05. In this case, the alpha level at each stage would be 0.0182 (as tabled
in Pocock, 1977, 1982), implying a z-critical value of 2.361, and yielding a flat significance bound (as illustrated in Figure 3.1). Thus at each stage, we would compare the calculated standard normal Z-statistic against the z-critical value 2.361. As long as the critical value is not exceeded, the study is allowed to continue. Tabled critical values (and derivations) for Pocock’s method and subsequent methods discussed in this chapter are provided in many sources, including Jennison & Turnbull (2000).

![Figure 3.1. Pocock Z-critical Values](image)

Advantages of the Pocock method include relative ease of use (since a single critical value suffices for all interim looks), and disadvantages include the fact that group sizes are required to be approximately equal at each stage, and the constant significance bound means that we have “spent” too much of the type I error rate to have much left for the final comparison.
3.2.2 O’Brien-Fleming

The O’Brien-Fleming (1979) approach improves on Pocock’s method by giving us ever-increasing alpha levels. In other words, for early interim looks, the z-critical values are relatively large, making it more difficult to reject the null hypothesis. As the study progresses and as the data accumulate, the z-critical values become smaller (and the nominal alpha levels increase). The O’Brien-Fleming method uses the same standardized statistics $Z_k$ as the Pocock method, and as with Pocock’s method, alpha levels are calculated using the joint distribution of $Z_1, \ldots, Z_K$. The increasing nominal alpha levels at each interim analysis are calculated as $\alpha_k' = 2[1-\Phi\{C_k\sqrt{(K/k)}\}]$ instead of as $\alpha' = 2[1-\Phi\{C_k\}]$. The constants $C_k$, which are tabled in O’Brien-Fleming (1979), Jennison and Turbull (2000), and elsewhere, are again calculated using numerical integration.

In a case where a total of four analyses (three interim and one final) are planned at a significance level of 0.05, the two-sided z-critical value at each stage is: $Z_k = 2.024 \sqrt{(4/k)}$, for $k = 1, \ldots, 4$, as illustrated in Figure 3.2. As stated above, the constant $C_k = 2.024$ is tabled or can be calculated using numerical integration.

Advantages of the O’Brien-Fleming method include the fact that it is more conservative at the beginning and also more powerful than the Pocock method. Disadvantages include the fact that group sizes are still required to be approximately equal at each stage, and the number and spacing of interim looks must still be planned in advance.
For 4 Interim Looks

Figure 3.2.  *O’Brien-Fleming Z-critical Values*

### 3.2.3 Lan-DeMets

While the O’Brien-Fleming method can be seen as an improvement over the Pocock method in terms of flexibility in error spending, both approaches suffer from the requirements that the group sizes included in each interim look be fairly constant and that the number of interim looks must be planned in advance.

Lan and DeMets (1983) proposed to deal with this problem by constructing a flexible error spending function that would not require advance specification of the number of interim looks, nor would the interim looks need to be equally spaced. As we shall illustrate through data examples in Chapter 5, this error-spending approach is applicable to a variety of data types, including survival data, the area of interest for this paper. This error-spending function can be constructed as:

\[
\alpha_1^*(t) = 2 - 2\Phi(z_{\alpha/2}/\sqrt{t}), \text{ as an analogue to the O’Brien-Fleming method, or as}
\]

\[
\alpha_2^*(t) = \alpha \log(1+(e-1) t), \text{ to simulate the Pocock method,}
\]

(3.7) \hspace{1cm} (3.8)
where in both cases $\alpha$ denotes the overall type I error rate, and $0 \leq t \leq 1$ denotes the fractional time (of the total length of the study) at which the interim look occurs. When group sizes are roughly equal, Lan and DeMets demonstrate that these two functions yield critical values that are close to those achieved by the O’Brien-Fleming and Pocock methods, respectively. As discussed earlier, both methods allow flexibility in the spacing and timing of interim looks.

A variation on the Lan-DeMets function yields a family of error-spending functions:

$$\alpha_3^*(t) = \alpha \min \{t^\rho, 1\}, \text{ for } \rho > 0 \quad (3.9)$$

Certain choices of $\rho$ can yield boundaries similar to those of O’Brien-Fleming and Pocock, as discussed in Kim & DeMets (1987), Jennison & Turnbull (2000), and elsewhere. We shall utilize one variant of (3.7) in Chapter 5 in an example analyzing multivariate survival data. Many other variants are possible, with the key property being that $\alpha(t)$ must be a monotonically increasing function of time ($t$), where $t \in [0,1]$.

Regardless of which Lan-DeMets spending function is selected, the type I error rate will be preserved. We can show this starting with the assumptions discussed earlier in Section 3.1.

As before, we assume $\delta_1, \ldots, \delta_K$ have a canonical joint distribution, but now the information levels $I_1, \ldots, I_K$ may be unequally spaced. For each $k$, we assume $I_k$ may be influenced by external factors, such as the accrual rate of suitable subjects, but not by the previously observed test statistics, $\delta_1, \ldots, \delta_{k-1}$. 
Let \( \alpha(t) \) (the Lan-DeMets error spending function) denote any monotonically increasing function of \( t \) for \( t \) in \([0,1]\) with \( \alpha(0)=0 \) and \( \alpha(1)=\alpha \).

Why is the type-1 error rate preserved, even if the number and spacing of the interim looks are allowed to change?

The probability of rejection under the null at any time during the study can be written as:

\[
\alpha = P_0(Z_{1} \geq c_1) + P_0(Z_{1}<c_1,Z_{2} \geq c_2) + P_0(Z_{1}<c_1,Z_{2}<c_2,Z_{3} \geq c_3) + \ldots \\
+ P_0(Z_{1}<c_1,Z_{2}<c_2, \ldots, Z_{K-1}<c_{K-1},Z_{K} \geq c_K) \\
= \alpha(t_1) + [\alpha(t_2)-\alpha(t_1)] + \ldots + [\alpha(t_K)-\alpha(t_{K-1})].
\]

If we change the spacing and number of the \( K \) interim looks, we now have \( K' \) interim looks (where \( K \neq K' \)) at times: \( t'_1, t'_2, \ldots, t'_{K'} \).

Then we have probability of rejection under the null, at any time during the study:

\[
\alpha(t'_1') + [\alpha(t'_2')-\alpha(t'_1')] + \ldots + [\alpha(t'_{K'})-\alpha(t'_{K'-1})] = \alpha(t'_{K'}) \leq \alpha
\]

One caveat is that one must not use the observed interim results (i.e., the observed \( Z_k \) test statistics) to make the decision to modify the spacing or number of interim looks, or the canonical distribution theory will no longer hold, although impact on the type I error rate is slight, per Lan and DeMets (1989) and Proschan et al (1992). We shall show through specific data examples in Chapter 5 how the Lan-DeMets spending function can be applied to survival data.
3.2.4 Wang and Tsiatis

Wang and Tsiatis (1987) extended this earlier work by introducing a class of two-sided stopping boundaries which are approximately optimal. Their group sequential test rejects $H_0$ at interim analysis $k$ if:

$$|S_k| \geq \Gamma (\alpha, K, \Delta) k^\Delta; \quad k = 1, ..., K,$$

for a total of $K$ interim analyses, where $S_k$ is actually the Z-statistic (3.1) discussed earlier in this chapter. The constant $\Gamma (\alpha, K, \Delta)$, chosen to achieve level $\alpha$ of significance, is calculated using numerical recursive integration to solve:

$$P(|S_1|<\alpha_1, ..., |S_K|<\alpha_K| \mu_a - \mu_b = 0) = 1 - \alpha$$

An advantage of this method is that depending on the choice of $\Delta$, one can obtain boundaries of different shapes, including that of Pocock ($\Delta = 0.5$) or O’Brien-Fleming ($\Delta = 0.0$) boundaries. Wang and Tsiatis report that an O’Brien-Fleming type boundary is most desirable in cases where the study is expected to run to completion, i.e., where $\delta = \mu_a - \mu_b$ is small, since this will lead to a lower maximum sample size by the end of the study. A disadvantage of this method is that it is limited to normally distributed data, and hence may not be applicable to survival data, our area of interest.
3.2.5 Haybittle-Peto

A very simple, easily implemented, and now commonly used method was proposed by Haybittle (1971) and Peto et al (1976). This group sequential test rejects Ho at interim analysis \( k \) if:

\[
|Z_k| \geq 3; \quad k = 1, ..., K,
\]

where \( Z_k \) is the usual two-sided z-statistic. This test is equivalent to stopping if the p-value at any interim analysis \( k \) is less than \( \alpha = 0.001 \). This is a simple, ad hoc procedure, which for the final analysis uses a z-critical value of 1.96 to achieve an approximate 0.05 significance level. The obvious downside to this method is that it is very difficult to stop the trial early, because the alpha level is so low. On the other hand, the impact on the final alpha level (after the completion of the study) is minimal, so this method may be preferred in studies in which early stopping is not anticipated.

The calculation of the final alpha level can be illustrated for a simple case involving one interim analysis and one final analysis, i.e., where \( K=2 \). (The calculations for the case where \( K>2 \) are aided by the fact that the sequence of test statistics is Markov, as noted earlier.) To achieve an overall type I error rate of 0.05 when \( K=2 \), we require:

\[
P(|Z_1| > 3 \mid \delta=0) + P(|Z_1|<3, |Z_2|>c_2 \mid \delta=0) = 0.05.
\]

At the first interim analysis we have \( P(|Z_1| > 3 \mid \delta=0) \approx 0.001 \), so we must solve for \( c_2 \), the boundary at the second analysis needed for an overall 0.05 error rate, i.e.:
\[ P(|Z_1| > 3 \mid \delta=0) + P(|Z_1|<3, |Z_2|>c_2 \mid \delta=0) = 0.001 + P(|Z_1|<3, |Z_2|>c_2 \mid \delta=0) = 0.05 = > \]
\[ P(|Z_1|<3, |Z_2|>c_2 \mid \delta=0) = 0.049 \]

It can be shown (Jennison & Turnbull, 2000) that \( Z_1 \) and \( Z_2 \) are bivariate normal, with covariance \( 1/\sqrt{2} \). We can then determine the actual boundary value \( c_2 \) by solving:

\[ (3.16) \quad P(|Z_1|<3, |Z_2|>c_2 \mid \delta=0) = 2 \int_{-3-\infty}^{c_2} \exp\left(\frac{z_1^2 + z_2^2 - \sqrt{2}z_1z_2}{\pi}\right) dz_2 dz_1 \]

### 3.2.6 Discussion

The early methods discussed above in this section serve as the core or foundation on which we shall build a class of group sequential methods to utilize in a situation requiring testing of multiple survival endpoints. Much flexibility is possible based on the requirements of the study, and each method is less or more suitable depending on the design of the study, and the expectations for how it will progress, as illustrated below in Table 3.1.

### 3.3 GS Methods for Survival Data

We next shift our attention away from early work to our more specific area of interest, which is the analysis of survival data in the context of a group sequential study. We shall consider several methods in this section which are familiar from the non-sequential setting, and show how they are adapted. Later, in Chapter 4, we discuss how these methods can be applied more specifically to multivariate survival data.
Table 3.1. Comparison of Selected Group Sequential Testing Methods

<table>
<thead>
<tr>
<th>Test</th>
<th>Benefits</th>
<th>Drawbacks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pocock</td>
<td>Easier to stop early</td>
<td>(1) Large impact on final alpha level</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(2) Cannot deviate from preplanned number and spacing of interim looks</td>
</tr>
<tr>
<td>O’Brien-Fleming</td>
<td>Fairly low impact on final alpha level since</td>
<td>(1) Difficult to stop early, but easier as trial progresses</td>
</tr>
<tr>
<td></td>
<td>little is spent at earlier analyses, and often</td>
<td>(2) Cannot deviate from preplanned number and spacing of interim looks</td>
</tr>
<tr>
<td></td>
<td>more powerful than Pocock.</td>
<td></td>
</tr>
<tr>
<td>Lan-DeMets</td>
<td>Flexibility in number and spacing of interim</td>
<td>Caution needed when changing number or spacing so as not to invalidate</td>
</tr>
<tr>
<td></td>
<td>looks, without affecting type I error</td>
<td>distribution theory.</td>
</tr>
<tr>
<td>Wang and Tsiatis</td>
<td>Can mimic popular methods such as O’Brien-Flem</td>
<td>Cannot deviate from preplanned number and spacing of interim looks; may</td>
</tr>
<tr>
<td></td>
<td>ing and Pocock using different choices of Δ</td>
<td>not be applicable to survival data.</td>
</tr>
<tr>
<td>Haybittle-Peto</td>
<td>Minimal impact on final alpha level, and easy</td>
<td>(1) Difficult to stop early</td>
</tr>
<tr>
<td></td>
<td>to apply.</td>
<td>(2) Cannot deviate from preplanned number and spacing of interim looks</td>
</tr>
</tbody>
</table>

3.3.1 Stratified and Unstratified Log-Rank Test

As discussed in Chapter 2, the n-sample unstratified log-rank test statistic can be written as:

\[
S = \sum_{i=1}^{n} \frac{(o_i - e_i)^2}{e_i} \sim \chi^2(n-1),
\]

for treatment group \(i = 1, ..., n\), where \(o_i\) indicates the observed number of events and \(e_i\) indicates the expected number of events. Jennison and Turnbull (2000) show how this
can be rewritten in a group sequential setting so that the log-rank statistics have approximately a jointly multivariate normal distribution for $k = 1, \ldots, K$ interim analyses:

$$Z_k = \frac{\sum_{i=1}^{d_{ik}} (\hat{\delta}_{B,k} - \delta_{i,k})}{\left(\sum_{i=1}^{d_{ik}} v_{i,k}\right)^{1/2}} = \frac{\sum_{i=1}^{d_{ik}} (\hat{\delta}_{B,k} - \delta_{i,k})}{\sqrt{I_k}} = \frac{S_k}{\sqrt{I_k}} \sim N(\theta \sqrt{I_k}, 1),$$

where $\delta_{iA,k}$ and $\delta_{iB,k}$ denote the number of subjects that fail at time $\tau_{i,k}$ on treatment arms A and B, respectively, and $\theta = \log(\lambda)$, the log of the hazard ratio. At each interim analysis $k$, $d_k$ distinct failure times $\tau_{1,k} < \tau_{2,k} \ldots < \tau_{d_k,k}$ are observed. Formulas for the expectation $e_{ik}$ and variance $v_{ik}$ and more details are provided in Jennison and Turnbull (2000). Note that this follows the canonical joint distribution (3.3) discussed earlier. The stratified version may be rewritten as:

$$Z_k = \frac{\sum_{m=1}^{M} \sum_{i=1}^{d_{mk}} (\hat{\delta}_{B,mk} - \delta_{i,mk})}{\left(\sum_{m=1}^{M} \sum_{i=1}^{d_{mk}} v_{i,mk}\right)^{1/2}},$$

for $m = 1, \ldots, M$ strata, where $\delta_{iA, mk}$ and $\delta_{iB, mk}$ denote the number of subjects that fail at time $\tau_{i,k}$ in strata $m$ on treatment arms A and B, respectively. We wish to test the null hypothesis of no difference in hazard rates $h_A$ and $h_B$ between the two treatment groups, i.e., $H_0: h_A(t) = h_B(t)$ for all $t > 0$, or $H_0: h_{Am}(t) = h_{Bm}(t)$ for $t > 0$ and $m = 1, \ldots, M$, for the stratified version, against an alternative $H_a: h_{Am}(t) = \lambda \cdot h_{Bm}(t)$, where $\lambda$ is a common hazard ratio. In the interim setting we select our group sequential method as desired (e.g., Pocock, O’Brien-Fleming, etc.), and simply compare $Z_k$ against the appropriate critical value. Again, additional details are provided in Jennison and Turnbull (2000).
3.3.2 Error Spending Approach

As discussed above in Section 3.2.3, methods developed by Lan and DeMets allow flexibility in the spacing of interim analyses and also offer the benefit of not requiring that the number of interim analyses be predetermined. This can be particularly useful in a study involving survival endpoints, as the information (which is proportional to the number of events or failures) may accumulate unevenly, unlike a trial involving a more straightforward continuous or categorical outcome. Information may often accumulate unevenly in a survival study since more events may be observed at one interim analysis than at another, whereas information is collected at regular predetermined intervals in the non-survival setting.

We now present a simple example to illustrate how an error spending function can be applied to a survival study. Suppose the planned duration of our study is 48 months, and we are interested in taking an interim look at 12 months and again at 36 months, as well as a final analysis at the end of the study. Suppose also that we are interested in one-sided testing. Our alpha level must then be partitioned between a total of three analyses. Using a simple Lan-DeMets error spending function of \( f(t) = \alpha \cdot \min(t, 1) \) with \( \alpha = 0.05 \), we can obtain approximate alpha-level boundaries and one-sided z-critical values for each analysis as indicated below in Table 3.2. At interim analysis one, our nominal alpha level is simply \( \alpha \cdot t = 0.05 \cdot (12/48) = 0.0125 \). We then obtain our Z-critical value from a normal table, or calculate it as \( c_1 = \Phi^{-1}(1-0.0125) = 2.241 \). The calculations at interim analysis two require that we calculate the joint probability \( P(Z_1 > 2.241, Z_2 \geq c_2) = (36/48 - 12/48) \times 0.05 = 0.0250 \). Calculation of the boundary value \( c_2 \) then requires that we consider the
joint distribution of $Z_1$ and $Z_2$, i.e., we must calculate $\text{Corr}(Z_1, Z_2)$. For interim analyses $k>2$, we utilize the Markov property discussed earlier. Specialized Fortran code and further calculation details are provided in Jennison and Turnbull (2000). We could easily substitute different adjusted z-critical values into this procedure, if desired, such as Pocock (1977) or O’Brien-Fleming (1979).

### Table 3.2. Boundary Values for Interim Analyses using Error Spending Function

<table>
<thead>
<tr>
<th>Analysis</th>
<th>Timepoint (months)</th>
<th>Nominal $\alpha$-level</th>
<th>Cumulative Alpha</th>
<th>Z-critical Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>12</td>
<td>12/48 x 0.05 = 0.0125</td>
<td>0.0125</td>
<td>2.241</td>
</tr>
<tr>
<td>2</td>
<td>36</td>
<td>(36/48 – 12/48) x 0.05 = 0.0250</td>
<td>0.0375</td>
<td>1.897</td>
</tr>
<tr>
<td>Final</td>
<td>48</td>
<td>(48/48 – 36/48) x 0.05 = 0.0125</td>
<td>0.0500</td>
<td>1.899</td>
</tr>
</tbody>
</table>

Calculation of the z-critical values is done using the recursive formulae of Armitage et al (1969), taking advantage of the fact that the $Z_k$ sequence is Markov:

\[
Z_k \sqrt{I_k} - Z_{k-1} \sqrt{I_{k-1}} \sim N(\theta(I_k - I_{k-1}), (I_k - I_{k-1}))
\]

We implement this method as follows: At interim analysis $k$ ($k = 1, 2, \text{ or } 3$), we reject $H_0$ if $Z_k = S_k / \sqrt{I_k}$ (from 3.18) exceeds the z-critical value tabled above. If the observed test statistic is less than the z-critical value at analysis $k$, the trial will proceed to the next analysis $k+1$.

#### 3.3.3 Cox Proportional Hazards Regression

The error-spending approach described in 3.3.2 can be applied also to more complicated survival models, such as the Cox proportional hazards model. In a group sequential setting, the model given in the previous chapter (2.5) can be rewritten as:

\[
\lambda_{mi}(t) = \lambda_{m0}(t) \exp\{\beta_0' Z_{mi}(t)\}, \ t \geq 0
\]
We shall show the joint distribution properties of the test statistics arising from this model later in Chapter 4. Note, however, that a z-statistic can be formed as described in the non-sequential setting in Chapter 2, and can then be compared with the appropriate critical values as described in Section 3.3.2.

### 3.3.4 Alternatives to the Cox Proportional Hazards Model

As discussed in Chapter 2, an alternative to the Cox Proportional Hazards model may be desirable if the proportional hazards assumption is violated. Several such alternatives were presented, including accelerated failure time models, additive hazard models and frailty models. We shall reserve discussion of their possible application to the group sequential setting for Chapter 4, which is devoted to discussion of multivariate survival data in the group sequential setting.

### 3.4 GS Methods for Multivariate Data

Now that we have discussed the analysis of survival data in the context of a group sequential study, we shall consider available methods for analysis of multivariate data in a group sequential setting. The reader should note that this discussion now shifts to techniques which are not specifically or necessarily intended for survival data, although in later sections we discuss when and how these techniques can be adapted to be useful also for multivariate survival data.

#### 3.4.1 Composite Scores

One method of dealing with the problem of multiple outcomes is the formation of a single composite score, using a linear combination or by assigning different weights to
the individual response variables. In effect, this reduces the problem from a multivariate to a univariate one, so that the methods described earlier in this chapter may be applied directly using a univariate test statistic. This is discussed in the survival, non-sequential setting by Wei and Glidden (1997), who suggest, among other options, the formation of a global, numerical score to combine information across multiple survival outcomes. The obvious downside of this method is the loss of information available with a multivariate approach. Additionally, there may be situations where it is not feasible or meaningful to combine endpoints, such as when endpoints do not run in the same direction.

3.4.2 Bonferroni Procedure

To maintain the prespecified type I error rate, one can use a Bonferroni correction (Dunn, 1961) by forming a separate test statistic for each variable at each interim analysis and rejecting the null hypothesis at that time based on the result of any individual endpoint. As in the non-sequential setting, the alpha level for that interim analysis will be decreased by dividing by the number of endpoints, i.e., if the alpha level at interim analysis k is 0.025 and there are \( m \) endpoints, the new adjusted alpha at that timepoint is \( \alpha' = 0.025/m \). This method has the advantage of being very simple to implement, but is inefficient and too conservative, particularly in the typical situation where multivariate outcomes are correlated.

3.4.3 Group Sequential Hotelling Test

Another option for handling multivariate data is the Hotelling test, which is well studied in the group sequential setting (Jennison and Turnbull, 1991, 1997, and elsewhere).
Briefly, for interim analyses \( k = 1, \ldots, K \), if the underlying data are multivariate normal, the expression:

\[
W_k = Z_k' \cdot \Sigma_k^{-1} \cdot Z_k
\]  

is distributed as \( \chi^2_p \) under \( H_0 \), where \( \Sigma_k \) denotes the correlation matrix and \( Z_k \) is a vector of standardized z-statistics for \( p \) endpoints. Jennison and Turnbull (1991) show that the sequence \( W_1 \ldots W_K \) is Markov, meaning that it holds the independent increment properties described earlier in this chapter.

Although this procedure is less conservative than the Bonferroni procedure described above, it allows only a global test of treatment difference at each analysis, which is not as useful as we would like if our goal is to make more specific inferences about individual endpoints. Additionally, this method assumes that the underlying data are multivariate normal, which is generally not applicable to survival data, our area of interest in this paper.

### 3.4.4 O’Brien’s Global Tests

Also well studied (by Tang et al 1989, 1993 and elsewhere) in the group sequential setting for multivariate data is the method originated by O’Brien (1984), which has both ordinary least squares (OLS) and generalized least squares (GLS) variants. Another similar variant is the approximate likelihood ratio test (Tang et al 1993).

A downside of this method is the requirement that all of the endpoints run in the same direction, i.e., if a large value for one endpoint is clinically desirable, the same should hold true for the other endpoints. As is true with the Hotelling procedure, the O’Brien
tests allow only a global test of treatment difference at each analysis and require the underlying data to be multivariate normal.

3.4.5 Closed Testing Procedures

As mentioned briefly in Chapter 1, Tang and Geller (1999) have developed methodology for closed testing in a group sequential study (illustrated in a schematic in Figure 3.3 below), enabling one to make inferences about multiple endpoints. Their method extends the closed testing procedure first discussed by Marcus, Peritz, and Gabriel (1976) and later by Lehmacher, Wassmer, and Reitmeir (1991), which was designed for a fixed sample study. Briefly, their approach involves the following steps:

1. Form a global hypothesis of all endpoints at interim analysis one:
   \[ H_0: \mu_{i1} - \mu_{i2} = 0 \text{ for all } i = 1, 2, ..., M \text{ endpoints, i.e., test that mean treatment difference is zero using a global test statistic.} \]

2. If hypothesis in step 1 is not rejected, continue the trial to interim analysis two and repeat step 1. This step is repeated for each succeeding interim analysis until a rejection occurs, or until the final analysis.

3. If the hypothesis in step 1 is rejected, the trial is stopped and the procedure steps down to test all (M-1) dimensional hypotheses. If one of these (M-1) dimension hypotheses is rejected, the procedure continues and tests the (M-2) dimensional hypotheses that are implied, and so on until single-endpoint testing is done. (Note however that if a certain hypothesis is not rejected, that we cannot test any of the subhypotheses implied by it, as indicated by the arrows in Figure 3.3. For
example, if we fail to reject $H_0^{(1,2,3)}$ we will not be able to proceed to test $H_0^{(1,2)}$, $H_0^{(1,3)}$, or $H_0^{(2,3)}$.

This procedure strongly controls the type-1 error rate, and has advantages of increased power or greater flexibility in interpreting results compared with previous multivariate group sequential methodology. Tang and Geller also describe a variation on this approach which would allow one to continue the study if, in step 3, some hypotheses were not rejected. We describe in detail later in Chapter 4 how Tang and Geller’s method can be extended to multivariate survival data.

Figure 3.3. *Closed Testing Schematic*

3.4.6 Bonferroni-Holm Method

The Bonferroni-Holm method (Holm, 1979), which is also known as Holm’s stepdown Bonferroni method, was extended from the fixed sample setting to the group sequential setting for multiple treatments (rather than multiple endpoints) by Follmann et al (1994). We shall discuss in Chapter 4 how to adapt this to the multiple endpoint scenario. SAS code is presented in Appendix C to conduct this analysis in the group sequential setting.

We describe briefly here how this method was implemented by Follmann et al (1994):

(1) Suppose that K interim analyses are planned for M treatment groups, of which one is a placebo or control group, and the other M-1 are active groups. Additionally, assume that we are interested only in the M-1 pairwise comparisons of each active group with placebo. Follmann et al propose to stop the study at interim analysis k if either of the following occurs:

(a) Any active treatment group is superior to placebo, OR
(b) All active treatment groups are inferior to placebo.

i.e., they test \( H_0: \mu_i = \mu_1 \) versus \( H_{A_i}: \mu_i \neq \mu_1 \) for \( i = 2, \ldots, M \), where \( \mu_1 \) denotes the mean response for the placebo group. Follmann et al suggest that an individual treatment arm might be dropped if it was found to be inferior to placebo, but that the trial itself would continue. They suggest that only in the extreme (and unlikely) circumstance that all groups were inferior would the trial stop at an interim point.

(2) At interim analysis k, we perform the following three steps:

(a) Re-order the M-1 pairwise p-values from smallest to largest, such that:

\[
p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(M-1)}.\]
(b) Compare \( p(1) \) against \( \alpha^* / M-1 \), \( p(2) \) against \( \alpha^* / M-2 \), …, \( p(M-1) \) against \( \alpha^* \), where \( \alpha^* \) is the nominal alpha-level calculated for interim analysis \( k \). (An additional step is added to ensure that the ordering of the adjusted comparisons is the same as the ordering of the unadjusted comparisons, as described below.) Recall that the alpha-level may vary at each interim analysis, depending on which group sequential procedure is selected.

(c) If \( p(1) \) is significant and if it indicates active is superior, we stop the trial, per (1a) above. If \( p(1) \) is significant and if it indicates active is inferior to placebo, then we proceed to check \( p(2) \). If all pairwise comparisons indicate superiority of placebo over active, then we stop the trial (for lack of efficacy), per (1b) above. If \( p(1) \) is non-significant, then all comparisons are non-significant, and we proceed to interim analysis \( k+1 \), repeating steps (2a) through (2c).

As a simpler way to implement the Bonferroni-Holm method, adjusted p-values can be calculated and compared directly with the nominal alpha level. The \( M-1 \) pairwise p-values are reordered from smallest to largest as \( p(1), p(2), \ldots, p(M-1) \). Suppose there are \( M-1 = 5 \) pairwise p-values. Each raw p-value \( p(m) \) is then adjusted as follows:

\[
\begin{align*}
\hat{p}_{(1)}^* &= 5 \cdot p(1), \text{ the smallest.} \\
\hat{p}_{(2)}^* &= \max(\hat{p}_{(1)}^*, 4 \cdot p(2)) \\
\hat{p}_{(3)}^* &= \max(\hat{p}_{(2)}^*, 3 \cdot p(3)) \\
\hat{p}_{(4)}^* &= \max(\hat{p}_{(3)}^*, 2 \cdot p(4)) \\
\hat{p}_{(5)}^* &= \max(\hat{p}_{(4)}^*, p(5)), \text{ the largest.}
\end{align*}
\]

This method has the advantage over Tang and Geller’s procedure of not requiring composite hypotheses (as illustrated in Figure 3.3), which makes implementation of the test and interpretation of the results much simpler, while still controlling the type-1 error rate. Additionally, Bonferroni-Holm controls the familywise error rate without requiring
the assumption that the multiple hypotheses are independent, which may be an advantage in situations where the independence assumption is suspect or clearly violated (e.g., potentially correlated multiple endpoints). We describe in detail later in Chapter 4 how the group sequential version of this method can be extended to multivariate survival data.

A key issue when implementing stepwise procedures such as the Bonferroni-Holm is control of directional error, sometimes known as the Type III error. In our setting, a Type III error is made by correctly concluding that two treatments differ for an endpoint, but incorrectly concluding the direction of the difference (i.e., concluding that active is superior to placebo when in fact placebo is superior.) This is particularly important to avoid since we are making the decision of whether to stop an entire trial for efficacy or lack of efficacy. Fortunately Type III errors are relatively rare, and the probability of a directional error is asymptotically negligible (Finner, 1999). Under certain conditions (Holm 1979 and Shaffer 1980), it can be proved that:

$$\sup_{\mu} P_{\mu} \{\text{directional error}\} \leq \alpha$$

Shaffer (1980) proved that the Type I and Type III error rates are simultaneously controlled at level $\alpha$ if the independence assumption is met. Finner (1999) extended this result to a larger class of stepdown and stepup procedures, including the Bonferroni-Holm and Hochberg methods, under less rigid assumptions.

**3.4.7 Hochberg Method**

Hochberg’s step-up method (Hochberg, 1988) is similar to the Holm procedure described above in that it uses the same critical values, but works in the reverse direction by starting
with the largest p-value and working “up” to the smallest. Due to the reversed over of p-value adjustment, Hochberg’s method can produce smaller p-values and hence can be more powerful than Holm’s method. A disadvantage of Hochberg’s method is that it does not guarantee control of the familywise error rate (FWE) as Holm’s does, particularly when p-values cannot be assumed to be independent; however, simulation studies (e.g., Sarkar and Chang, 1997) have shown that Hochberg’s method is often conservative anyway (i.e., leading to FWE < \( \alpha \)). Finner and Roters (1998) offer a theoretical comparison of the asymptotic properties of these two procedures.

While Hochberg’s method was not extended to the group sequential setting by Follmann et al (1994) as Holm’s method was, we can illustrate briefly here how that may be accomplished. We shall then discuss in Chapter 4 how to adapt this to the multiple endpoint scenario. SAS code is presented in Appendix C to conduct this analysis in the group sequential setting. The implementation of Hochberg’s method is as follows:

1. Identical to Follmann’s discussion for Holm, we propose to stop the study at interim analysis k if either of the following occurs:

   (a) Any endpoint shows active treatment group is superior to placebo, OR

   (b) All endpoints show active treatment group is inferior to placebo.

2. At interim analysis k, we perform the following three steps:

   (a) Re-order the M-1 pairwise p-values from largest to smallest, such that:

   \[ p_{(M-1)} \geq p_{(M-2)} \geq \ldots \geq p_{(1)}. \]

   (b) Compare \( p_{(M-1)} \) against \( \alpha^* \) (where \( \alpha^* \) is the nominal alpha-level calculated for interim analysis k), \( p_{(M-2)} \) against \( \alpha^*/2 \), \ldots, \( p_{(1)} \) against \( \alpha^*/(M-1) \). Recall that the
alpha-level may vary at each interim analysis, depending on which group sequential procedure is selected. As with the Bonferroni-Holm method, an additional step is added to ensure that the ordering of the adjusted comparisons is the same as the ordering of the unadjusted comparisons, as described below.

(c) Note that this step follows Bonferroni-Holm step 2c, although the actual adjusted p-values may or may not turn out to be the same. If \( p_{(M-1)} \) is significant and if it indicates active is superior, we stop the trial, per (1a) above. Note that if \( p_{(M-1)} \) is significant, then all p-values are significant due to the reordering from largest to smallest. If all pairwise comparisons indicate superiority of placebo over active, then we stop the trial (for lack of efficacy), per (1b) above. If \( p_{(1)} \) is non-significant, then all comparisons are non-significant, and we proceed to interim analysis \( k+1 \), repeating steps (2a) through (2c).

As a simpler way to implement the Hochberg method, adjusted p-values can be calculated and compared directly with the nominal alpha level. The M-1 pairwise p-values are reordered from largest to smallest as \( p_{(M-1)}, p_{(M-2)}, \ldots, p_{(1)} \). Suppose there are M-1 = 5 pairwise p-values. Each raw p-value \( p_{(m)} \) is then adjusted as follows:

\[
\begin{align*}
p^*_5 &= p_{(5)}, \text{ the largest.} \\
p^*_4 &= \min(p^*_5, 2 \cdot p_{(4)}) \\
p^*_3 &= \min(p^*_4, 3 \cdot p_{(3)}) \\
p^*_2 &= \min(p^*_3, 4 \cdot p_{(2)}) \\
p^*_1 &= \min(p^*_2, 5 \cdot p_{(1)}), \text{ the smallest.}
\end{align*}
\]

This method shares the advantages of the Bonferroni-Holm method mentioned in the previous section. A simple consideration of the adjusted p-values above shows how Hochberg’s method can be more powerful. Note that where the Bonferroni-Holm method would calculate an adjusted p-value \( p^*_{(3)} = \max(p^*_2, 3 \cdot p_{(3)}) \), Hochberg’s method
calculates $p^*_{(3)} = \min(p^*_{(4)}, 3 \cdot p_{(3)})$. We describe in Chapter 4 how the group sequential version of this method can be extended to multivariate survival data.
CHAPTER IV

GROUP SEQUENTIAL METHODOLOGY FOR MULTIVARIATE SURVIVAL ANALYSIS

4.1 Introduction

In the sections that follow, we discuss various types of multivariate survival data, review existing strategies for their analysis in a non-group sequential setting, and then discuss new ways to analyze multivariate survival data in the group sequential setting.

4.2 Types of Multivariate Survival Data

To recall the discussion in Chapter 2, multiple survival outcomes can fall into different categories, including the following: unordered and distinct events, recurrences of the same event type, and ordered distinct events. In this dissertation, we shall focus on distinct events which are typically unordered, although the situation may occur where a subject drops out of the risk set after experiencing a given event (e.g., death). Therneau and Grambsch (2000) review four major strategies of handling multiple events per subject in the non-sequential setting:

1. Time to first event, ignoring multiplicity
2. Assume multiple events are independent, conditional on a per-subject coefficient (frailty model)
3. Marginal models approach
4. Model subject’s correlation within Cox framework

In the sections below, we shall discuss methodology for handling multiple survival endpoints per subject.
4.2.1 Distinct, Unordered Events

As an example, we consider a long-term cardiac study, in which outcomes of interest may include time to major adverse cardiovascular events such as heart attack, stroke, myocardial infarction, death, etc. In such an example, the events are distinct rather than recurring, and there is not necessarily any inherent ordering, i.e., a heart attack could occur before a stroke or after a stroke. The focus of the rest of this dissertation will be on this event type.

4.2.2 Distinct, Ordered Events

Ordered events may occur in a situation where one endpoint is considered to take priority over another, for example, when one endpoint is designated to be of primary clinical importance, while others are secondary, tertiary, and so forth. There may be other situations where one outcome cannot occur before another for underlying reasons intrinsic to the nature of the data. If the primary event is expected to occur for very few patients, additional endpoints may be considered to increase the power to detect a difference between treatment groups or to reduce the number of enrolled patients.

4.2.3 Recurring Events

Typically when considering recurring events, we are interested in a single type of survival outcome which may occur multiple times for a patient over the course of a long-term study, such as multiple hospitalizations or heart attacks. In such a situation, the recurrent events will not be independent of one another for a given patient, and so we must address the issue of intrasubject correlation in selecting our model. Since the focus of this dissertation is on distinct events, we will not provide additional detailed discussion on recurring events. In the non-sequential study setting, the reader may refer to the work
4.2.4 Competing Risks

As an example of competing risks, one may be interested in tracking time until first hospitalization, and there may be a variety of possible reasons for hospitalization. Each patient is at risk of hospitalization, but a given patient will have only one cause of hospitalization (i.e., the possible causes are “competing” risks). Again, the focus of this dissertation is on distinct multiple events, so we provide no further discussion on competing risks.

4.3 Analyzing Multivariate Survival Data in the GS Setting

Having considered the variety of types of multivariate survival data, we now discuss how to analyze distinct, unordered multiple event data by applying survival methodology to the group sequential study setting.

4.3.1 Marginal Hazards Models (WLW variant)

Suppose there are \( M \) survival endpoints of interest and \( P \) covariates of interest and suppose we are interested in distinct, unordered events, as described in section 4.2.1 (although this method can be applied also to recurrent events). Let \( m = 1, \ldots, M \) (e.g., time until heart attack, time until stroke, etc.), and let \( p = 1, \ldots, P \) (e.g., treatment group and other covariates of interest). If we consider a Cox proportional hazards model, per Wei et al (1989) we can model the hazard function for each of the \( M \) endpoints as:

\[
\lambda_{mi}(t) = \lambda_{m0}(t) \exp\{\beta_0' Z_{mi}(t)\}, \ t \geq 0
\]
where $\lambda_{m0}(t)$ is the baseline hazard for the $m^{th}$ endpoint, $\beta_0$ is the regression coefficient (or vector of coefficients), and $Z_{mi}(t)$ is the covariate (or vector of covariates) for the $m^{th}$ endpoint and $i^{th}$ subject. Note that since this is a stratified Cox model, there is a separate underlying hazard function for each endpoint (Wei et al 1989, Spiekerman and Lin 1998).

Let $\hat{\beta}$ be the vector of maximum partial likelihood estimates ($\hat{\beta}_1 ... \hat{\beta}_P$) for ($\beta_1 ... \beta_P$), where $P$ is the number of covariates in the model. Note that if we are interested in only one predictor (e.g., treatment group), then $P=1$ and $\hat{\beta}$ reduces to a scalar, $\hat{\beta}$.

$\hat{\beta}$ can then be estimated from the following partial likelihood function (Cox 1972; 1975, Wei et al 1989):

$$L_m(\beta) = \prod_{i=1}^{n} \left[ \frac{\exp(\beta'Z_{mi}(X_{mi})) \Delta_{mi}}{\sum_{\ell \in S_{mi}(X_{mi})} \exp(\beta'Z_{mi}(X_{mi}))} \right]$$

by setting $\partial \log L_m(\beta) / \partial \beta$ equal to zero and solving for $\beta$, where $m$ denotes the endpoint as before, $i=1, ..., n$ is the subject indicator, $\Delta_{mi}$ indicates whether subject $i$ experienced the $m^{th}$ event, the summation over $\ell$ denotes the set of subjects at risk prior to time $t$ for event $m$, and $X_{mi}$ denotes the failure time of the $i^{th}$ subject for event $m$. While $\hat{\beta}$ cannot be expressed in closed form, a numerical estimate can be obtained readily using popular software packages (SAS, Splus, etc.).

Then $\hat{\beta}$ is asymptotically multivariate $N(\beta, Q)$, where $Q$ is the covariance matrix (per Wei et al 1989). The asymptotic covariance matrix can be estimated by:
where the $D(\beta_i, \beta_j)$ terms are the dfbeta statistics (the standardized difference in the parameter estimate due to deleting the observation, or the change in the estimate of $\beta_j$ if observation $i$ is deleted). In the multiple endpoint per subject situation, we delete one subject at a time rather than one observation at a time (per Therneau and Grambsch, 2000). This expression for the covariance matrix is also referred to as a robust sandwich covariance matrix, since it can be rewritten in the “ABA” sandwich form (where A is the variance and B is a correction term) as:

$$
\hat{Q} = \left( J - \bar{J} \right) (J - \bar{J}) = D' D = I^{-1} (U' U) I^{-1},
$$

where $J$ is composed of jackknife influence values (i.e., $\hat{\beta}_{(i)} - \hat{\beta}$), $I$ is the identity matrix, and $U$ is a matrix of score residuals. Now we can form a single test statistic $W$ over all $M$ events and all $P$ covariates, i.e., a test of the global null hypothesis:

$$
W = \left( \hat{\beta}_1, \ldots, \hat{\beta}_p \right) \hat{Q}^{-1} \cdot \begin{pmatrix}
\hat{\beta}_1 \\
\vdots \\
\hat{\beta}_p
\end{pmatrix} \sim \chi^2_{(P)}
$$

Note that $W$, which is a scalar, is also a quadratic form (i.e., a homogeneous 2-degree polynomial, Hogg & Craig, 1995). However, if we are interested only in one element of the $P$-variate regression parameter $\beta$, for example, the treatment group covariate, we can form a single degree-of-freedom test statistic $W_p$ by selecting the $p^{th}$ diagonal element from the robust sandwich covariance matrix $\hat{Q}$ and the $p^{th}$ element of $\hat{\beta}$:
(4.6) \[ W_p = \left( \frac{\hat{\beta}}{SE_{robust}} \right)^2 \sim \chi^2(1) \]

where \( SE_{robust} \) is the \( p \)th element of the robust sandwich covariance matrix, as noted above. Since \( W_p \) is a one degree-of-freedom test, we can form a \( z \)-statistic:

(4.7) \[ z_p = \frac{\hat{\beta}}{SE_{robust}} \sim N(\beta, 1) \]

Note that \( z_p \), which is Wald’s test, is a test of \( H_0: \beta_p = 0 \) vs \( H_a: \beta_p > 0 \) across all \( M \) strata or endpoints for the \( p \)th covariate.

Recall that since our larger goal is to utilize this test statistic in a group sequential setting, we will have a separate test statistic at each interim analysis \( k \), i.e., we will have \( z_1, \ldots, z_K \).

The sequence of interim estimates \( \hat{\beta}^{(1)}, \ldots, \hat{\beta}^{(K)} \) (where \( K \) denotes the total number of interim analyses, as before) is asymptotically multivariate normal, as shown in Chapter 3 section 3.1 using a canonical joint distribution:

(4.8) \[ \hat{\beta}^{(k)} \sim N(\beta, I_{k}^{-1}) \text{ for each } k, \]

and \( \text{Cov}(\hat{\beta}^{(k_1)}, \hat{\beta}^{(k_2)}) = (I_{k_2})^{-1} \) for \( 1 \leq k_1 \leq k_2 \leq K \),

where \( I_k = \text{se}(\beta^{(k)})^{-1} \).

Hence the sequence of estimates \( (I_1)^{1/2} \hat{\beta}^{(1)}, \ldots, (I_K)^{1/2} \hat{\beta}^{(K)} \) is also multivariate normal.

Let \( Z_k = (I_k)^{1/2} \hat{\beta}^{(k)} \) (i.e., the Wald test). Note that this is equivalent to the expression above in 4.6. Then:

\[ Z_k \sim N((I_k)^{1/2}\beta, 1) \text{ for each } k, \]

and \( \text{Cov}(Z_{k_1}, Z_{k_2}) = \sqrt{(I_{k_1} / I_{k_2})} \) for \( 1 \leq k_1 \leq k_2 \leq K \).
Note as before that $Z_1, \ldots, Z_K$ are Markov, which greatly simplifies the group sequential stopping boundary computations.

This multivariate normal property will enable us to use the marginal proportional hazards model using a variety of typical group sequential procedures, as we shall illustrate in Chapter 5 using data examples and in Chapter 6 using simulations.

4.3.2 Accelerated Failure Time (AFT) Models

We consider the class of AFT models as an alternative to the marginal proportional hazards model when the proportionality assumption does not hold. To recall the discussion in Chapter 2, we are regressing the log-transformed survival times rather than the actual survival times. When multiple endpoints are of interest, the model can be written as:

$$\log_{10} T_{mi} = \beta^T\mu X_{mi} + e_{mi}, \quad i = 1, \ldots, n \tag{4.9}$$

where $T_{mi}$ denotes the failure time for the $m^{th}$ endpoint for patient $i$, $\beta_m$ is a vector of regression coefficients for the $m^{th}$ endpoint, $X_{mi}$ is a vector of covariates (such as treatment group) for the $m^{th}$ endpoint for patient $i$, and the error distribution $e_{mi}$ is i.i.d. and can be from a variety of parametric models, such as gamma or normal, or nonparametric.

Lin and Wei (1992) and Lin et al (1998) discuss this method in the context of recurrent, rather than distinct, multivariate failure times. One advantage of this method is that it does not present the problem of neglected covariates. But according to Wei and Glidden (1997), “it is not clear how this multivariate AFT model can handle the case when there
are competing risks.” We shall describe in this section how the AFT model can be used in the group sequential setting for recurrent events, and then discuss whether it can be extended also for the case where a patient may have distinct, multiple events of interest.

Let $\hat{\beta}$ be the vector of maximum partial likelihood estimates $(\hat{\beta}_1 \ldots \hat{\beta}_M)$ for $(\beta_1 \ldots \beta_M)$, where $M$ is the number of multiple or recurrent events of interest in the model.

For model 4.9, $\hat{\beta}$ can then be estimated using rank statistics as follows (Lin and Wei, 1992). Let $S_{mn}(\beta)$ be a weighted score function for the $m^{th}$ failure which corresponds to the Cox partial likelihood function or to the Peto-Prentice Wilcoxon statistic (depending on the score function $\phi$):

\begin{equation}
S_{mn}(\beta) = n^{-1} \sum_{i=1}^{n} \Delta_{mi} \cdot \phi(\hat{F}_m(e_m(\beta))) \cdot (x_{mi} - \bar{x}_{mi}) , \ i = 1, \ldots, n
\end{equation}

where $e_{mi}(\beta) = \log T_{mi} - \beta'X_{mi}$, $\hat{F}_m$ is the Kaplan-Meier estimate of the distribution of the $e_{mi}(\beta)$'s, $\Delta_{mi}$ is the censoring indicator, and:

\begin{equation}
\bar{x}_{mi} = \frac{\sum_{j=1}^{n} x_{mj} \cdot I(e_{mj}(\beta) \geq e_{mi}(\beta))}{\sum_{j=1}^{n} I(e_{mj}(\beta) \geq e_{mi}(\beta))}
\end{equation}

for indicator function $I(\cdot)$. Note that the score function $\phi$ in (4.10) is the Cox function when $\phi(\cdot)=1$, and is the Peto-Prentice Wilcoxon statistic when $\phi(u)=1-u$. We can estimate $\beta$ by setting $S_{mn}(\beta) = 0$ and solving for $\beta$. While this cannot be expressed in closed form, the usual numerical estimation methods (using SAS Proc Lifereg, Splus function survreg,
etc.) allow us to estimate $\hat{\beta}$. Further details of the estimation and asymptotic distributions are also presented in the appendix of Lin and Wei (1992).

Lin and Wei show that $n^{1/2}(\hat{\beta}_m - \beta_m)$, $m = 1, \ldots, M$ is asymptotically $N(0, \Sigma)$, where $
abla = \nabla' \nabla$, $\Lambda$ is a $pM \times pM$ matrix, and $\Gamma = \text{diag}(\Gamma_m)$. They also show in detail that a consistent estimator for $\Lambda$ is $\lambda_{kk}(\hat{\beta}_k, \hat{\beta}_l)$.

Since we are typically interested in testing the treatment group portion of the covariate vector at each interim analysis, we can form single degree-of-freedom Wald chi-square tests as illustrated above in previous sections. As above in WLW, we can form a $z$-statistic (4.7) at each interim analysis $k$, i.e., we will have $z_1, \ldots, z_K$. The usual argument can be made using a canonical joint distribution and using the fact that this sequence is Markov (as discussed in Chapter 3, section 3.1). We are then able to use the AFT model with a variety of typical group sequential procedures, as we shall illustrate using data examples in Chapter 5.

### 4.3.3 Frailty Models

We consider the class of frailty models as an alternative to the marginal proportional hazards model when we wish to assume a shared frailty for a “family” or “cluster” of observations. The frailty can be conceived as an unobserved random effect which is common to all members of a family. In this section we shall consider the gamma frailty and the gaussian frailty models, but a wide variety of methods of modelling the frailty are possible.
For example, we may assume that family members share a similar risk of experiencing a type of cancer due to shared environment or heritability. As another example, we can suppose a given patient is at risk of multiple, distinct cardiac events (e.g., angina, stroke, myocardial infarction, etc.). In this case, the patient himself becomes the “family” or “cluster”, and the distinct types of events are members of that “family”. The frailty then becomes a way of modelling the intrasubject correlation, since the distinct events are not independent of one another. When multiple endpoints are of interest, the model can be written as:

\[(4.12) \lambda_{ij}(t) = \lambda_0(t) u_i \exp(\beta Z_{ij}), i = 1, \ldots, G; j = 1, \ldots, n_i\]

where \(\lambda\) denotes the hazard, \(i\) denotes a subject with \(n_i\) events (or an individual member of a subgroup, such as littermates), \(u_i\) denotes the random frailty effect within each subject, \(G\) is the number of subjects, and \(j\) denotes the event indicator within subject. Essentially this form of the model is just like the Cox model, but with frailty \(u_i\) incorporated multiplicatively into the hazard.

Therneau and Grambsch (2000) propose to write the model instead as:

\[(4.13) \lambda_i(t) = \lambda_0(t) \exp(Z_i\beta + X_i\omega), i = 1, \ldots, G\]

where \(\lambda_i\) denotes the hazard for “family” member \(i\), for a family containing members \(i = 1, \ldots, G\), and \(Z\) is a matrix of fixed effect covariates, \(X\) is a design matrix, and \(\omega\) is a vector of frailties. We assume that subjects are conditionally independent given \(\omega\), where \(\omega\) is a random variable from a known distribution (e.g., gamma or normal). We can define the vector \(\omega\) of frailties using a penalty function \(p(\omega) = (1/\theta) \sum [\omega_i - \exp(\omega_i)]\) (for a gamma frailty model, where \(\omega_i\) are distributed as the log of iid gamma random variables),
or as \( p(\omega) = (1 / 2\theta) \sum [\omega_i^2] \) (for a gaussian frailty model). In both cases, \( \theta = \text{Var} (\omega) \).

Essentially, we can fit a penalized Cox model with penalty function defined according to the distribution of \( \omega \).

We must first test the significance of the frailty effect. Our null hypothesis is that the frailty \( \omega = 0 \) versus an alternative hypothesis that \( \omega \neq 0 \). This is tested using a score test developed by Commenges and Andersen (1995):

\[
(4.14) \quad z = \frac{T}{\sqrt{V}} \sim \text{N} (0, 1),
\]

where \( T \) is the score test statistic and \( V \) is the estimated variance (with details provided in Commenges and Andersen). If frailty is found to be significant, it must be retained in the model; if not, the frailty term is dropped and we are left with the usual proportional hazards model. In Splus, this can be accomplished with the following code: coxph (Surv(time, status), trt + covariates(s) + frailty(id), data = datasetname), where “id” denotes the random frailty effect in the model.

Next, to test the treatment effect, we find again that we can form a single degree-of-freedom Wald chi-square test:

\[
(4.15) \quad z = \frac{\hat{\beta}}{SE} \sim \text{N} (\beta, 1):
\]

where the treatment parameter \( \hat{\beta} \) can be estimated by maximizing the log likelihood, and \( SE \) is the associated standard error. If we assume a gamma frailty model with \( \Gamma(1/\theta) \), the likelihood to be maximized is (Klein and Moeschberger, 1997):

\[
(4.16) \quad L(\theta, \beta) = \sum_{i=1}^{G} D_i \ln \theta - \Gamma(1/\theta) - \Gamma(1/\theta + D_i) \ln[1 + \theta \sum_{j=1}^{n_i} H_0(T_{ij}) \exp(\beta^t Z_{ij})]
\]
\[ + \sum_{j=1}^{n_i} \delta_{ij} \{ \beta' Z_{ij} + \ln[h_0(T_{ij})] \} \]

where \( i = 1, \ldots, G \), \( j = 1, \ldots, n_i \), \( D_i \) denotes the number of events in group \( i \), and \( \theta \) denotes the variance. If \( h_0 \) is assumed to be parametric, \( \beta \) and its variance can be estimated using maximum likelihood estimates. If \( h_0 \) is not parametric, the EM algorithm can be used. The Splus code provided above can be modified to specify the frailty distribution. As above in WLW, we can form a z-statistic (4.15) at each interim analysis \( k \), i.e., we will have \( z_1, \ldots, z_K \). The usual argument can be made using a canonical joint distribution and using the fact that this sequence is Markov (as discussed in Chapter 3, section 3.1). We are then able to use the frailty model with a variety of typical group sequential procedures.

4.3.4 Closed Testing Methods

Considering the multivariate survival methods discussed in sections 4.3.1 – 4.3.5, we now discuss the application of multiple comparison strategies to the group sequential setting. We can now demonstrate that Tang and Geller’s (1999) closed testing method (described in Chapter 3) can be applied to multivariate survival outcomes using the marginal models approach (3) described above in sections 4.3.1 through 4.3.3. Suppose that we are comparing two treatment groups \((i = 1, 2)\) on \( M \) survival outcomes of interest, for \( m = 1, 2, \ldots, M \). Also, we are interested in conducting \( K \) interim analyses for \( k = 1, 2, \ldots, K \).

Our hypothesis of interest at each interim analysis \( k \) is that there is no difference between the two treatment groups, namely, \( H_0: \beta_k = 0 \) across all \( M \) strata or endpoints, for \( m = 1, 2, \ldots, M \), against a one-sided alternative \( H_a: \beta_k > 0 \), where \( \beta_k \) denotes the regression
parameter at interim analysis k. We can apply either variant of the closed testing procedure while preserving strong control of the type I error.

For each hypothesis, we can calculate a single Wald chi-squared test statistic using a proportional hazards marginal model by stratifying on each distinct survival outcome type and collapsing on the subject.

A robust sandwich estimate \( \hat{S} \) of the covariance matrix can be calculated as described in Wei, Lin, and Weissfeld (1989) for a situation with k multivariate outcomes:

\[
\hat{S} = \begin{pmatrix}
V_{11} & V_{12} & \cdots & V_{1k} \\
V_{21} & V_{22} & \cdots & V_{2k} \\
\vdots & \vdots & \ddots & \vdots \\
V_{k1} & V_{k2} & \cdots & V_{kk}
\end{pmatrix}, \quad \text{where } V_{ij} = \hat{A}_i (R_i^\prime R_j) \hat{A}_i,
\]

\( \hat{A}_i \) denotes the covariance matrix obtained by inverting the information matrix, and \( R_i \) denotes the matrix of score residuals. This matrix is easily computable using SAS PHREG with option COVSANDWICH or using S-plus macros such as those provided by Terry Therneau of Mayo’s statistics department (Therneau and Grambsch, 2000).

Since we are using a single test statistic, we can establish critical values at each interim analysis based on a wide variety of group sequential testing methods. We then proceed to step through the closed testing method as described in Chapter 3.

4.3.5 Bonferroni-Holm and Hochberg Methods

We now demonstrate how the Bonferroni-Holm (1979) and Hochberg (1988) sequentially rejective methods (described in Chapter 3) can be applied to multivariate survival
outcomes, a new application for these methods. As in the previous section, suppose that we are comparing two treatment groups \((i = 1, 2)\) on \(M\) survival outcomes of interest, for \(m = 1, 2, \ldots, M\). To simplify the illustration, we may suppose that one treatment is active and the other is placebo. As usual, we are interested in conducting \(K\) interim analyses for \(k = 1, 2, \ldots, K\). Recall that these methods control the familywise error rate without requiring the assumption that the multivariate endpoints are independent. Recall also that Hochberg’s method is more powerful than the Bonferroni-Holm method, although it does not always guarantee FWE control. We can modify Follmann et al’s (1994) version as follows:

(1) Our hypothesis of interest at each interim analysis \(k\) is \(H_0: \beta = 0\) for \(m = 1, \ldots, M\) versus a two-sided alternative; i.e., for each endpoint, there is no treatment difference in time to event. We follow the logic of Follmann et al, proposing to stop the study at interim analysis \(k\) if either of the following occurs:

(a) For any endpoint, the active treatment group is superior to placebo, OR

(b) For all endpoints, the active treatment group is inferior to placebo.

(2) At interim analysis \(k\), we perform the following three steps:

(a) Re-order the \(M\) pairwise p-values from smallest to largest, such that:

\[
p_1(1) \leq p_1(2) \leq \cdots \leq p_1(M).
\]

Note that since no composite hypotheses are required, we can calculate \(p_1(1), \ldots, p_1(M)\) using a simple Cox proportional hazards model rather than the marginal proportional hazards model utilized in the previous section. (For Hochberg’s method, we instead order from largest to smallest, as described in Chapter 3.)
(b) Compare \( p_{(1)} \) against \( \alpha^* / M \), \( p_{(2)} \) against \( \alpha^* / M-1 \), \ldots, \( p_{(M)} \) against \( \alpha^* \), where \( \alpha^* \) is the nominal alpha-level calculated for interim analysis \( k \). As before, \( \alpha^* \) may vary at each interim analysis, depending on which group sequential procedure is selected. And as described previously, we must add an additional step to ensure that the ordering of the adjusted comparisons is the same as the ordering of the unadjusted comparisons.

(c) If \( p_{(1)} \) is significant and if it indicates active is superior for that endpoint, we stop the trial, per (1a) above. If \( p_{(1)} \) is significant and if it indicates active is inferior to placebo, then we proceed to check \( p_{(2)} \). If placebo is found to be superior to active for all endpoints (i.e., all pairwise comparisons), then we stop the trial (for lack of efficacy), per (1b) above. If \( p_{(1)} \) is non-significant, then all comparisons are non-significant, we proceed to interim analysis \( k+1 \), and repeat steps (2a) through (2c). Note that for Hochberg’s method, we instead begin with \( p_{(M-1)} \), the largest p-value. If this is significant, then we know immediately that all \( M-1 \) comparisons are significant.

4.3.6 Fisher’s LSD Analogue

In the discussion section of Follmann et al’s (1994) paper, it was mentioned that a referee had suggested global-based monitoring using an analogue to Fisher’s protected least significant difference (LSD) procedure (Fisher, 1935), wherein a composite hypothesis would be tested before proceeding to test all pairwise comparisons. One caveat to this procedure is that Hayter (1986) showed that the probability of at least one incorrect assertion exceeds \( \alpha \) in the one-way ANOVA model. It is suitable for testing the composite hypothesis, but offers only weak control of the familywise error rate for
subsequent testing of individual hypotheses (Westfall et al, 1999). With this caveat in mind, in this section we show how this analogue might be extended to test M multiple endpoints rather than M multiple treatment groups. We shall then illustrate the application of this method in Chapter 5, and shall investigate using data simulations how well it attains the type-1 error rate.

As before, we compare two treatment groups (i = 1, 2) on M survival outcomes of interest (m = 1, 2, … , M) at each of K interim analyses (k = 1, 2, ..., ,K). The procedure is briefly outlined below:

1. **Global test (analogous to the F-test in Fisher’s protected LSD):** Utilizing the marginal proportional hazards model discussed above in the context of Tang and Geller’s (1999) closed testing procedure, we first test equality of treatment groups across all M endpoints, i.e., we calculate a single Wald chi-squared test statistic by stratifying on each distinct survival outcome type and collapsing on the subject.

2. If this global hypothesis is **significant** at $\alpha^*$, we **stop the trial** and test all pairwise comparisons, i.e., we test equality of active to placebo for each of the M endpoints at the same level $\alpha^*$.

3. If this global hypothesis is **non-significant** at $\alpha^*$, we **do not stop the trial** and do not examine any pairwise comparisons. Instead, the trial continues to the next interim analysis.

Again, $\alpha^*$ is calculated for each interim analysis using any of the group sequential methods discussed earlier, and the spacing and timing of the interim analyses can be flexible if we utilize a Lan-DeMets type error spending function.
4.4 Summary of Methods Proposed

In this chapter, we have outlined several survival data techniques that can be extended to the group sequential setting with multiple distinct or recurrent events. In all cases, after applying the survival methodology, a group sequential technique must then be selected in order to control the type I error rate, as discussed in Chapter 3. Survival methods discussed in this chapter include marginal proportional hazards models, the accelerated failure time model, and frailty models. We then discussed three multiple comparison procedures that can be used in combination with multivariate survival analysis in order to allow one to make inferences about multiple endpoints. We shall illustrate in the next chapter how to apply these methods to a real data example involving multiple, distinct endpoints. In Chapter 6 we then examine these methods in the context of simulation studies.
CHAPTER V
DATA EXAMPLES

5.1 Introduction

In this section we shall introduce the real dataset to be studied using the methods outlined in Chapter 4. In section 5.2 we shall apply the closed testing method to this dataset, using a variety of multivariate survival procedures and group sequential procedures. In section 5.3 we shall apply the Bonferroni-Holm method, in section 5.4 we shall apply the Hochberg method, and in section 5.5 we shall apply the Fisher’s LSD analogue, in each case using a variety of multivariate survival and group sequential procedures. Finally, in section 5.5 we review the results and discuss the benefits and limitations of these methods. SAS and Splus code is provided in Appendix C to apply these procedures. These methods will be studied later in Chapter 6 in data simulations.

As an example of a dataset containing multiple, distinct survival endpoints of interest, we shall analyze a primary biliary cirrhosis (PBC) data set discussed by Lindor et al (1994) and by Therneau and Grambsch (2000). This study, which was conducted by the Mayo Clinic, involved 180 subjects treated over a four-year period in a double-blind fashion with the active drug UDCA (ursodeoxycholic acid) or placebo. The data set (see Appendix B) includes 170 of those subjects and consists of times to ten adverse events of interest. To simplify our illustration, we shall confine our attention to four of these endpoints: death, transplant, histologic progression, and development of varices. When modeling the data, we shall also adjust for two baseline covariates: high histologic stage indicator (1 or 0), and log-transformed bilirubin levels. Note that each subject was at risk for each event and that the events could have occurred in a different order for different
subjects. While some subjects experienced more than one distinct type of event, other subjects experienced no events, and no subject experienced the same type of event more than once. The number of subjects experiencing each event is shown below in Table 5.1.

While the actual study did not involve group sequential monitoring, for the purposes of illustration we can pretend that there were two interim analyses (arbitrarily selected at year one and year three) and one final analysis (after year four). One complicating factor is that patient entry times were staggered, but we shall simplify the analyses for now by ignoring this complication and simply looking at the data in terms of time to event or censoring for each subject and for each endpoint of interest.

<table>
<thead>
<tr>
<th>Event</th>
<th>Timepoint</th>
<th>Cumulative Counts</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>UDCA N = 86</td>
<td>Placebo N = 84</td>
<td>Overall N = 170</td>
<td></td>
</tr>
<tr>
<td>Death</td>
<td>Month 12</td>
<td>0 (0.0)</td>
<td>3 (3.6)</td>
<td>3 (1.8)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Month 36</td>
<td>3 (3.5)</td>
<td>7 (8.3)</td>
<td>10 (5.9)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Final</td>
<td>6 (7.0)</td>
<td>10 (11.9)</td>
<td>16 (9.4)</td>
<td></td>
</tr>
<tr>
<td>Transplant</td>
<td>Month 12</td>
<td>1 (1.2)</td>
<td>1 (1.2)</td>
<td>2 (1.2)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Month 36</td>
<td>3 (3.5)</td>
<td>5 (6.0)</td>
<td>8 (4.7)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Final</td>
<td>6 (7.0)</td>
<td>6 (7.1)</td>
<td>12 (7.1)</td>
<td></td>
</tr>
<tr>
<td>Histologic Progression</td>
<td>Month 12</td>
<td>0 (0.0)</td>
<td>0 (0.0)</td>
<td>0 (0.0)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Month 36</td>
<td>7 (8.1)</td>
<td>10 (11.9)</td>
<td>17 (10.0)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Final</td>
<td>8 (9.3)</td>
<td>12 (14.3)</td>
<td>20 (11.8)</td>
<td></td>
</tr>
<tr>
<td>Development of Varices</td>
<td>Month 12</td>
<td>0 (0.0)</td>
<td>0 (0.0)</td>
<td>0 (0.0)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Month 36</td>
<td>7 (8.1)</td>
<td>13 (15.5)</td>
<td>20 (11.8)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Final</td>
<td>8 (9.3)</td>
<td>17 (20.2)</td>
<td>25 (14.7)</td>
<td></td>
</tr>
<tr>
<td>At Least One Event</td>
<td>Month 12</td>
<td>1 (1.2)</td>
<td>4 (4.8)</td>
<td>5 (2.9)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Month 36</td>
<td>17 (19.8)</td>
<td>31 (36.9)</td>
<td>48 (28.2)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Final</td>
<td>23 (26.7)</td>
<td>37 (44.0)</td>
<td>60 (35.3)</td>
<td></td>
</tr>
</tbody>
</table>

5.2 Analysis Results using Closed Testing Method (CTM)

We first illustrate the application of the Closed Testing method, which was first used in a group sequential setting by Tang et al (1999) and was described in the previous chapter.
Note that this multiple comparison method dictates the process and order of testing events but does not limit us to any particular group sequential or survival method. Hence, it is possible to combine CTM with a variety of group sequential and survival methods to see which performs best or is most illustrative for our data. Throughout section 5.2 we shall utilize a flexible Lan-DeMets error spending function for our group sequential method, and we select the marginal proportional hazards model for our survival method.

5.2.1 Lan-DeMets Error Spending Function Simulating O’Brien-Fleming Shape with WLW Method

Using a Lan-DeMets error spending function of \( f(t) = 2 - 2\Phi\left(\frac{z_{\alpha/2}}{\sqrt{t}}\right) \) to simulate the O’Brien-Fleming method with \( \alpha=0.05 \), we obtain approximate alpha-level boundaries and one-sided \( z \)-critical values for each analysis as indicated in Table 5.2:

<table>
<thead>
<tr>
<th>Analysis</th>
<th>Timepoint (Months)</th>
<th>( \alpha )-level</th>
<th>Cumulative Alpha</th>
<th>Z-critical Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>12</td>
<td>&lt;0.0001</td>
<td>&lt;0.0001</td>
<td>3.750</td>
</tr>
<tr>
<td>2</td>
<td>36</td>
<td>0.0235</td>
<td>0.0236</td>
<td>1.985</td>
</tr>
<tr>
<td>Final</td>
<td>48</td>
<td>0.0264</td>
<td>0.0500</td>
<td>1.715</td>
</tr>
</tbody>
</table>

Recall that one-sided critical values are required by the closed testing method for reasons explained in the previous chapter. Our four endpoints of interest are:

Endpoint 1: Death  
Endpoint 2: Transplant  
Endpoint 3: Histologic Progression  
Endpoint 4: Development of Varices

At each interim analysis, we must test the hypotheses in the order dictated by the flowchart in Chapter 3 (Figure 3.3), so our first hypothesis to be tested is:

\[ H_0: \beta_1 = 0 \cup \beta_2 = 0 \cup \beta_3 = 0 \cup \beta_4 = 0, \]
i.e., for each endpoint there is no treatment difference. Our one-sided alternative hypothesis is that Active is superior to Placebo. More precisely, for each endpoint we test $H_0: \beta_i = 0$ (where $\beta_i$ is the regression coefficient for the treatment group effect) versus $H_a: \beta_i > 0$, with at least one inequality (as in Tang and Geller, 1999). If we define the hazard ratio as $HR = \exp(\beta) = \lambda_P / \lambda_A$, i.e., a regression coefficient greater than zero implies $HR > 1$, which implies that the hazard for placebo $\lambda_P$ exceeds the hazard for active $\lambda_A$.

At interim analysis one, we apply a WLW marginal proportional hazards model to our data using the robust sandwich covariance matrix estimate, obtaining a single test statistic for all four endpoints of $Z_{\{1,2,3,4\}} = 1.22$. Since this does not exceed the critical value 3.750, we would continue the trial, since there is insufficient evidence of a treatment difference in survival across all four events at month 12.

At the second interim analysis we calculate $Z_{\{1,2,3,4\}} = 2.51$, which exceeds the critical value of 1.985. At this point we would stop the study and apply the closed testing procedure to test the sub-hypotheses using the same critical value 1.985. At this stage we calculate the following: $Z_{\{1,2,3\}} = 2.00$, $Z_{\{1,2,4\}} = 2.39$, $Z_{\{1,3,4\}} = 2.45$, and $Z_{\{2,3,4\}} = 2.09$. Since all of these $Z$ statistics exceed the critical value 1.985, we can test the subhypotheses that are implied by the rejected tests, namely, all 2-level subhypotheses. The results of these two-endpoint subhypotheses are all non-significant ($z$-statistics greater than 1.985), so we cannot proceed to test any single-endpoint hypotheses.
So, we are able to conclude that there is a global treatment difference in these four endpoints and we can stop the trial early while strongly controlling the type-1 error rate, although interpretation of the subhypotheses is not entirely straightforward.

5.2.2 Lan-DeMets Error Spending Function Simulating Pocock Shape with WLW Method

Using a Lan-DeMets error spending function of \( f(t) = \alpha \log(1+(e-1) t ) \) to simulate the Pocock method with \( \alpha=0.05 \), we will obtain approximate alpha-level boundaries and one-sided z-critical values for each analysis as indicated in Table 5.3.

<table>
<thead>
<tr>
<th>Analysis</th>
<th>Timepoint (Months)</th>
<th>( \alpha )-level</th>
<th>Cumulative Alpha</th>
<th>Z-critical Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>12</td>
<td>0.01787</td>
<td>0.01787</td>
<td>2.100</td>
</tr>
<tr>
<td>2</td>
<td>36</td>
<td>0.02353</td>
<td>0.04140</td>
<td>1.902</td>
</tr>
<tr>
<td>Final</td>
<td>48</td>
<td>0.00860</td>
<td>0.05</td>
<td>1.999</td>
</tr>
</tbody>
</table>

At interim analysis one, we apply a WLW marginal proportional hazards model to our data using the robust sandwich covariance matrix estimate, obtaining a test statistic of all four endpoints of \( Z_{\{1,2,3,4\}} = 1.22 \). Since this does not exceed the critical value 2.10, we would continue the trial, since there is insufficient evidence of a treatment difference in survival across all four events at month 12.

At the second interim analysis we calculate \( Z_{\{1,2,3,4\}} = 2.51 \), which exceeds the critical value of 2.10. At this point we would stop the study and apply the closed testing procedure to test the sub-hypotheses using the same critical value 2.10. At this stage we calculate the following: \( Z_{\{1,2,3\}} = 2.00 \), \( Z_{\{1,2,4\}} = 2.39 \), \( Z_{\{1,3,4\}} = 2.45 \), and \( Z_{\{2,3,4\}} = 2.09 \). Since two of these sub-hypotheses (\( Z_{\{1,2,3\}} \) and \( Z_{\{2,3,4\}} \)) do not exceed the critical value 2.10, we cannot test any lower level hypotheses that they imply, namely, any sub-
hypothesis involving endpoints 1, 2, 3 or 4. So our conclusions are much the same as those reached using the Lan-DeMets error spending function simulating the O’Brien-Fleming shape in the previous section.

At this point, it should be clear that the use of different spending functions may allow us to stop the trial earlier or later, depending on how quickly the type-I error rate is spent. However, it will not necessarily give us any clearer interpretation of the results if we are not allowed to proceed to the single-endpoint testing stage.

### 5.2.3 Lan-DeMets Error Spending Function Simulating O’Brien-Fleming Shape with AG Method

Using a Lan-DeMets error spending function to simulate the O’Brien-Fleming method with \( \alpha = 0.05 \), we can use the same alpha-level boundaries and one-sided z-critical values for each analysis as indicated above in Table 5.2. As discussed in Chapter 4, Andersen and Gill’s (AG) method was designed for recurrent data rather than distinct, non-recurrent events. Can the AG method be applied anyway to our PBC dataset?

Consider how we might structure the PBC data to prepare it for analysis. With recurrent data using the counting process style of input, we are interested in time to first recurrent event, then time to second recurrent event, etc. So the dataset is structured as:

<table>
<thead>
<tr>
<th>Patient</th>
<th>Start</th>
<th>Stop</th>
<th>Censored</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>t1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>t1</td>
<td>t2</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>t2</td>
<td>t3</td>
<td>0</td>
<td>3</td>
</tr>
</tbody>
</table>

Consider patient 6 in our PBC dataset, who experiences esophogeal varices (endpoint 4) on day 769, death (endpoint 1) on day 1347, and does not experience liver transplant (endpoint 2) or histologic progression (endpoint 3). Suppose we structure the data as:
But another patient such as patient 7 is censored for all four events. The AG method does not allow us to include more than one censored record for a patient, so suppose we structure the data as:

<table>
<thead>
<tr>
<th>Patient</th>
<th>Start</th>
<th>Stop</th>
<th>Censored</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>0</td>
<td>1870</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

But clearly this structure will not work since it fails to account for the fact that the patient was censored for events 2, 3, and 4 as well as event 1. There is no way to properly structure distinct endpoints for the AG method. Hence, for the remainder of this chapter we shall omit discussion of the AG method.

### 5.2.4 Lan-DeMets Error Spending Function Simulating O’Brien-Fleming Shape with PWP Method

We have illustrated in section 5.2.3 that it is not possible to correctly apply the AG method to non-recurrent data. That limitation will apply also to the PWP method, which also utilizes a counting-process style of input. However, this method and the AG method could potentially be of use were we to analyze recurrent survival data, a problem which is outside the scope of this dissertation.

### 5.2.5 Lan-DeMets Error Spending Function Simulating O’Brien-Fleming Shape with AFT Method

Using a Lan-DeMets error spending function to simulate the O’Brien-Fleming method with $\alpha=0.05$, we can use the same alpha-level boundaries and one-sided z-critical values for each analysis as indicated above in Table 5.2. We shall apply the AFT (accelerated
failure time) model to our data using the Splus `survreg` function, which allows us to stratify by event type (event =1 to 4, as numbered above in section 5.2.1) and cluster by patient to account for intra-subject correlation. Throughout this chapter we use the default Weibull distribution, but note that the software allows us to model the error term using other distributions such as gamma or normal.

At interim analysis one, we apply an AFT model to our data using the robust sandwich covariance matrix estimate, obtaining a test statistic of all four endpoints of $Z_{\{1,2,3,4\}} = 1.13$. Since this does not exceed the critical value 2.10, we would continue the trial, since there is insufficient evidence of a treatment difference in survival across all four events at month 12.

At the second interim analysis we calculate $Z_{\{1,2,3,4\}} = 2.82$, which exceeds the critical value of 1.9. At this point we would stop the study and apply the closed testing procedure to test the sub-hypotheses using the same critical value 1.9. At this stage we calculate the following: $Z_{\{1,2,3\}} = 2.21$, $Z_{\{1,2,4\}} = 2.47$, $Z_{\{1,3,4\}} = 2.61$, and $Z_{\{2,3,4\}} = 2.47$. Since all of these sub-hypotheses exceed the critical value 1.9, we can proceed to test all two-level subhypotheses: $Z_{\{1,2\}} = 1.78$, $Z_{\{1,3\}} = 1.93$, $Z_{\{1,4\}} = 2.23$, $Z_{\{2,3\}} = 1.73$, $Z_{\{2,4\}} = 2.05$. Two of these subhypotheses ($Z_{\{1,2\}}$ and $Z_{\{2,3\}}$) do not exceed 1.9, so we can test single-level hypotheses for endpoint 4 only: $Z_4 = 1.8$. So, we can “accept” the hypothesis $H_0: \beta_4 = 0$, i.e., there is no evidence of difference in the risk of development of varices for active versus placebo at the time of the second interim analysis. However, we have still rejected the compound hypothesis, i.e., there is evidence of difference in the
risk for active versus placebo at the time of the second interim analysis, when combining evidence across all four endpoints.

So our conclusions are much the same as those reached using the WLW method, although the reader may note that many of the Z-statistics were larger, which could make it possible to reject earlier under various circumstances, such as if a different interim point had been selected.

5.2.6 Lan-DeMets Error Spending Function Simulating O’Brien-Fleming Shape with Frailty Method

Using a Lan-DeMets error spending function to simulate the O’Brien-Fleming method with $\alpha=0.05$, we can use the same alpha-level boundaries and one-sided z-critical values for each analysis as indicated above in Table 5.2. We shall apply a gamma frailty model to our data using the Splus `coxph` function, which allows us to stratify by event type (event =1 to 4, as numbered above in section 5.2.1), assigning a separate frailty to each patient. Throughout this chapter we use the default gamma frailty, but note that the software allows us to model the frailty using other distributions such as normal.

At interim analysis one, we apply the model to our data, obtaining a test statistic of all four endpoints of $Z_{\{1,2,3,4\}} = 1.24$. If we exclude the frailty term, our Z-statistic will be 1.25. In either case this will not exceed the critical value 2.10, so we would continue the trial, since there is insufficient evidence of a treatment difference in survival across all four events at month 12.

At the second interim analysis we calculate $Z_{\{1,2,3,4\}} = 2.66$, which exceeds the critical value of 1.9. At this point we would stop the study and apply the closed testing
procedure to test the sub-hypotheses using the same critical value 1.9. Our test for frailty is non-significant \((p = 0.29)\), so we refit without the frailty. At this stage we calculate the following: \(Z_{\{1,2,3\}} = 2.24\), \(Z_{\{1,2,4\}} = 2.55\), \(Z_{\{1,3,4\}} = 2.62\), and \(Z_{\{2,3,4\}} = 2.46\). Since all of these sub-hypotheses exceed the critical value 1.9, we can proceed to test all two-level subhypotheses: \(Z_{\{1,2\}} = 1.83\), \(Z_{\{1,3\}} = 1.93\), \(Z_{\{1,4\}} = 2.29\), \(Z_{\{2,3\}} = 1.73\), \(Z_{\{2,4\}} = 2.09\). Two of these subhypotheses \((Z_{\{1,2\}}\) and \(Z_{\{2,3\}}\)) do not exceed 1.9, so we can test single-level hypotheses for endpoint 4 only: \(Z_4 = 1.76\). So, we can “accept” the hypothesis \(\beta_4 = 0\), i.e., there is no evidence of difference in the risk of development of varices for active versus placebo at the time of the second interim analysis. However, we have still rejected the compound hypothesis, i.e., there is evidence of difference in the risk for active versus placebo at the time of the second interim analysis, when combining evidence across all four endpoints.

So our conclusions are much the same as those reached using the WLW and AFT methods, although the reader may note that many of the Z-statistics were larger than the AFT Z-statistics (which were in turn larger than the WLW Z-statistics), which could make it possible to reject earlier under various circumstances, such as if a different interim point had been selected.

5.3 Analysis Results using Bonferroni-Holm Method (BHM)

We next illustrate the application of the Bonferroni-Holm (1979) method described in the previous chapter. Again, this multiple comparison methods dictates the process and order of testing events but does not limit us to any particular group sequential or survival method. Hence we shall combine BHM with a variety of group sequential and survival methods to see which performs best or is most illustrative for our data.
5.3.1 Lan-DeMets Error Spending Function Simulating O'Brien-Fleming Shape with WLW Method

As in section 5.2.1, we use a Lan-DeMets error spending function of \( f(t) = 2 - 2\Phi(z_{\alpha/2}/\sqrt{t}) \) to simulate the O'Brien-Fleming method. At each interim analysis, we shall use the alpha levels and z-critical values listed previously in Table 5.2.

At interim analysis one, we apply a separate proportional hazards model to our data for each endpoint, obtaining the following results: \( p_1 = 0.9971 \) (for Death), \( p_2 = 0.9460 \) (for Transplant), \( p_3 = p_4 > 0.9999 \) for both Histologic Progression and Varices. Using the Bonferroni-Holm procedure, we would place these unadjusted p-values in ascending order, but clearly we will not meet the nominal alpha-level of <0.0001 for this analysis, since there were so few events at the first interim analysis. Therefore we cannot stop the trial at the first interim analysis, and must proceed to the second interim analysis.

At interim analysis two, we shall use a nominal alpha level of 0.0235, but note that this alpha level is divided by the number of endpoints for comparison with the smallest p-value as described in Chapter 4. We obtain the following results: \( p_1 = 0.1469 \) (for Death), \( p_2 = 0.2580 \) (for Transplant), \( p_3 = 0.1899 \) (for Histologic Progression), and \( p_4 = 0.0781 \) (for Varices). Again, there is insufficient evidence of efficacy at this interim analysis, since the smallest p-value \( p_4 \) would have to be compared against a nominal alpha-level of \( 0.0235 / 4 = 0.0059 \), and the second smallest p-value \( p_1 \) would be compared with \( 0.0235 / 3 = 0.0078 \).

So, for this particular example, the Bonferroni-Holm Method does not allow us to stop the trial early, unlike the result we saw using the Closed Testing Method. Since we are
testing each endpoint individually, we sacrifice the additional power we gained by combining endpoints with the CTM. We are forced to use a smaller alpha-level, and our p-values are larger as well.

5.3.2 Lan-DeMets Error Spending Function Simulating Pocock Shape with WLW Method

Using a Lan-DeMets error spending function of \( f(t) = \alpha \log(1+(e-1) t) \) to simulate the Pocock method with \( \alpha = 0.05 \), we will obtain approximate alpha-level boundaries and one-sided z-critical values for each analysis as indicated in Table 5.3.

At interim analysis one, we apply a separate proportional hazards model to our data for each endpoint, obtaining the following results: \( p_1 = 0.9971 \) (for Death), \( p_2 = 0.9460 \) (for Transplant), \( p_3 = p_4 > 0.9999 \) for both Histologic Progression and Varices. Using the Bonferroni-Holm procedure, we would place these unadjusted p-values in ascending order, but clearly we will not meet the nominal alpha-level of 0.01787 for this analysis, since there were so few events at the first interim analysis.

At interim analysis two, we shall use a nominal alpha level of 0.0235, but note that this alpha level is divided by the number of endpoints for comparison with the smallest p-value as described in Chapter 4. We obtain the following results: \( p_1 = 0.1469 \) (for Death), \( p_2 = 0.2580 \) (for Transplant), \( p_3 = 0.1899 \) (for Histologic Progression), and \( p_4 = 0.0781 \) (for Varices). Again, there is insufficient evidence of efficacy at this interim analysis, since the smallest p-value \( p_4 \) would have to be compared against a nominal alpha-level of \( 0.0414 / 4 = 0.0104 \), and the second smallest p-value \( p_1 \) would be compared with \( 0.0414 / 3 = 0.0138 \).
Again, the Bonferroni-Holm Method does not allow us to stop the trial early, unlike the result we saw using the Closed Testing Method. We are allowed slightly larger alpha levels using Pocock rather than O’Brien-Fleming, but this is insufficient to allow us to reject the null hypotheses at any interim stage.

5.3.3 Lan-DeMets Error Spending Function Simulating O’Brien-Fleming Shape with AFT Method

As in section 5.2.1, we use a Lan-DeMets error spending function of \( f(t) = 2 - 2\Phi\left(\frac{z_{\alpha/2}}{\sqrt{t}}\right) \) to simulate the O’Brien-Fleming method. At each interim analysis, we shall use the alpha levels and z-critical values listed previously in Table 5.2.

At interim analysis one, we apply a separate proportional hazards model to our data for each endpoint, obtaining the following results: \( p_1 > 0.9999 \) (for Death), \( p_2 = 0.9385 \) (for Transplant), \( p_3 = p_4 > 0.9999 \) for both Histologic Progression and Varices. Using the Bonferroni-Holm procedure, we would place these unadjusted p-values in ascending order, but clearly we will not meet the nominal alpha-level of <0.0001 for this analysis, since there were so few events at the first interim analysis.

At interim analysis two, we shall use a nominal alpha level of 0.0235, but note that this alpha level is divided by the number of endpoints for comparison with the smallest p-value as described in Chapter 4. We obtain the following results: \( p_1 = 0.1609 \) (for Death), \( p_2 = 0.2579 \) (for Transplant), \( p_3 = 0.1695 \) (for Histologic Progression), and \( p_4 = 0.0719 \) (for Varices). Again, there is insufficient evidence of efficacy at this interim analysis, since the smallest p-value \( p_4 \) would have to be compared against a nominal
alpha-level of $0.0235 / 4 = 0.0059$, and the second smallest p-value $p_1$ would be compared with $0.0235 / 3 = 0.0078$.

Again, the Bonferroni-Holm Method does not allow us to stop the trial early, unlike the result we saw using the Closed Testing Method, and we see that the AFT p-values are not much different from the p-values obtained using the proportional hazard model.

### 5.3.4 Lan-DeMets Error Spending Function Simulating O’Brien-Fleming Shape with Frailty Method

As in section 5.2.1, we use a Lan-DeMets error spending function of $f(t) = 2 - 2\Phi(z_{\alpha/2}/\sqrt{t})$ to simulate the O’Brien-Fleming method. At each interim analysis, we shall use the alpha levels and z-critical values listed previously in Table 5.2.

At interim analysis one, we apply a separate frailty model to our data for each endpoint, obtaining the following results: $p_1 > 0.9999$ (for Death), $p_2 = 0.95$ (for Transplant), $p_3 = p_4 > 0.9999$ for both Histologic Progression and Varices. Using the Bonferroni-Holm procedure, we would place these unadjusted p-values in ascending order, but clearly we will not meet the nominal alpha-level of $<0.0001$ for this analysis, since there were so few events at the first interim analysis.

At interim analysis two, we shall use a nominal alpha level of 0.0235, but note that this alpha level is divided by the number of endpoints for comparison with the smallest p-value as described in Chapter 4. We obtain the following results: $p_1 = 0.150$ (for Death), $p_2 = 0.260$ (for Transplant), $p_3 = 0.19$ (for Histologic Progression), and $p_4 = 0.078$ (for Varices). Again, there is insufficient evidence of efficacy at this interim analysis.
Again, the Bonferroni-Holm Method does not allow us to stop the trial early, unlike the result we saw using the Closed Testing Method, and we see that the AFT p-values are not much different from the p-values obtained using the proportional hazard model.

5.4 Analysis Results using Hochberg Method

We next illustrate the application of the Hochberg (1988) step-up method described in the previous chapter. Again, this multiple comparison methods dictates the process and order of testing events but does not limit us to any particular group sequential or survival method. Hence we shall combine Hochberg with a variety of group sequential and survival methods to see which performs best or is most illustrative for our data.

5.4.1 Lan-DeMets Error Spending Function Simulating O’Brien-Fleming Shape with WLW Method

As in section 5.2.1, we use a Lan-DeMets error spending function of \( f(t) = 2 - 2\Phi(z_{\alpha/2}/\sqrt{t}) \) to simulate the O’Brien-Fleming method. At each interim analysis, we shall use the alpha levels and z-critical values listed previously in Table 5.2.

At interim analysis one, we apply a separate proportional hazards model to our data for each endpoint, obtaining the following results: \( p_1 = 0.9971 \) (for Death), \( p_2 = 0.9460 \) (for Transplant), \( p_3 = p_4 > 0.9999 \) for both Histologic Progression and Varices. Using the Hochberg procedure, we would place these unadjusted p-values in descending order, but clearly we will not meet the nominal alpha-level of <0.0001 for this analysis, since there were so few events at the first interim analysis. Therefore we cannot stop the trial at the first interim analysis, and must proceed to the second interim analysis.
At interim analysis two, we shall use a nominal alpha level of 0.0235, but note that this alpha level is divided by the number of endpoints for comparison with the smallest p-value as described in Chapter 4. We obtain the following results: \( p_1 = 0.1469 \) (for Death), \( p_2 = 0.2580 \) (for Transplant), \( p_3 = 0.1899 \) (for Histologic Progression), and \( p_4 = 0.0781 \) (for Varices). Although it is clear that even unadjusted, we would not meet the nominal alpha level, as an exercise (as described in Chapter 3), we could then adjust these four p-values as follows:

\[
\begin{align*}
\hat{p}^{*}(4) &= 0.2580 \\
\hat{p}^{*}(3) &= \min (\hat{p}^{*}(4), 2 \cdot p(3)) = \min (0.2580, 2 \cdot 0.1899) = 0.2580 \\
\hat{p}^{*}(2) &= \min (\hat{p}^{*}(3), 3 \cdot p(2)) = \min (0.2580, 3 \cdot 0.1469) = 0.2580 \\
\hat{p}^{*}(1) &= \min (\hat{p}^{*}(2), 4 \cdot p(1)) = \min (0.2580, 4 \cdot 0.0781) = 0.2580
\end{align*}
\]

Note that these calculations can be automated by feeding the raw p-values into SAS Proc MultTest (code provided in Appendix C). Each of these four adjusted p-values is then compared directly with the nominal alpha level of 0.0235. Again, there is insufficient evidence of efficacy at this interim analysis, since all four p-values are clearly much larger than the nominal alpha level.

By comparison, if we had chosen to calculate adjusted p-values using the Bonferroni-Holm method, we would have calculated the following adjusted p-values (all of which are larger than the Hochberg adjusted p-values, demonstrating that the Hochberg is often more powerful):

\[
\begin{align*}
\hat{p}^{*}(1) &= 4 \cdot p(1) = 4 \cdot 0.0781 = 0.3124 \\
\hat{p}^{*}(2) &= \max (\hat{p}^{*}(1), 3 \cdot p(2)) = \max (0.3124, 3 \cdot 0.1469) = 0.4407 \\
\hat{p}^{*}(3) &= \max (\hat{p}^{*}(2), 2 \cdot p(3)) = \max (0.4407, 2 \cdot 0.1899) = 0.4407 \\
\hat{p}^{*}(4) &= \max (\hat{p}^{*}(3), p(4)) = \max (0.4407, 0.2580) = 0.4407
\end{align*}
\]

Again, for this particular example, the Hochberg Method does not allow us to stop the trial early, unlike the result we saw using the Closed Testing Method. Since we are
testing each endpoint individually, we sacrifice the additional power we gained by combining endpoints with the CTM. We are forced to use a smaller alpha-level, and our p-values are larger as well, just as we saw with the Bonferroni-Holm Method in the previous section.

5.4.2 Lan-DeMets Error Spending Function Simulating Pocock Shape with WLW Method

Using a Lan-DeMets error spending function of \( f(t) = \alpha \log(1+(e-1) t) \) to simulate the Pocock method with \( \alpha=0.05 \), we will obtain approximate alpha-level boundaries and one-sided z-critical values for each analysis as indicated in Table 5.3.

At interim analysis one, we apply a separate proportional hazards model to our data for each endpoint, obtaining the following results: \( p_1 = 0.9971 \) (for Death), \( p_2 = 0.9460 \) (for Transplant), \( p_3 = p_4 > 0.9999 \) for both Histologic Progression and Varices. Using the Hochberg procedure, we would place these unadjusted p-values in descending order, but clearly we will not meet the nominal alpha-level of 0.01787 for this analysis, since there were so few events at the first interim analysis.

At interim analysis two, we shall use a nominal alpha level of 0.0235, but note that this alpha level is divided by the number of endpoints for comparison with the smallest p-value as described in Chapter 4. We obtain the following results: \( p_1 = 0.1469 \) (for Death), \( p_2 = 0.2580 \) (for Transplant), \( p_3 = 0.1899 \) (for Histologic Progression), and \( p_4 = 0.0781 \) (for Varices). Again, it is clear that we will be unable to meet the nominal alpha level.
Again, the Hochberg Method does not allow us to stop the trial early, unlike the result we saw using the Closed Testing Method. We are allowed slightly larger alpha levels using Pocock rather than O’Brien-Fleming, but this is insufficient to allow us to reject the null hypotheses at any interim stage.

5.4.3 Lan-DeMets Error Spending Function Simulating O’Brien-Fleming Shape with AFT Method

As in section 5.2.1, we use a Lan-DeMets error spending function of $f(t) = 2 - 2\Phi(z_{\alpha/2}/\sqrt{t})$ to simulate the O’Brien-Fleming method. At each interim analysis, we shall use the alpha levels and $z$-critical values listed previously in Table 5.2.

At interim analysis one, we apply a separate model to our data for each endpoint, obtaining the following results: $p_1 > 0.9999$ (for Death), $p_2 = 0.9385$ (for Transplant), $p_3 = p_4 > 0.9999$ for both Histologic Progression and Varices. Using the Hochberg procedure, we would place these unadjusted $p$-values in descending order, but clearly we will not meet the nominal alpha-level of <0.0001 for this analysis, since there were so few events at the first interim analysis.

At interim analysis two, we shall use a nominal alpha level of 0.0235, but note that this alpha level is divided by the number of endpoints for comparison with the smallest $p$-value as described in Chapter 4. We obtain the following results: $p_1 = 0.1609$ (for Death), $p_2 = 0.2579$ (for Transplant), $p_3 = 0.1695$ (for Histologic Progression), and $p_4 = 0.0719$ (for Varices). Again, there is insufficient evidence of efficacy at this interim analysis.
Again, the Hochberg Method does not allow us to stop the trial early, unlike the result we saw using the Closed Testing Method, and we see that the AFT p-values are not much different from the p-values obtained using the proportional hazard model.

5.4.4 Lan-DeMets Error Spending Function Simulating O’Brien-Fleming Shape with Frailty Method

As in section 5.2.1, we use a Lan-DeMets error spending function of \( f(t) = 2 - 2\Phi(z_{\alpha/2}/\sqrt{t}) \) to simulate the O’Brien-Fleming method. At each interim analysis, we shall use the alpha levels and z-critical values listed previously in Table 5.2.

At interim analysis one, we apply a separate frailty model to our data for each endpoint, obtaining the following results: \( p_1 > 0.9999 \) (for Death), \( p_2 = 0.95 \) (for Transplant), \( p_3 = p_4 > 0.9999 \) for both Histologic Progression and Varices. Using the Hochberg procedure, we would place these unadjusted p-values in descending order, but clearly we will not meet the nominal alpha-level of <0.0001 for this analysis, since there were so few events at the first interim analysis.

At interim analysis two, we shall use a nominal alpha level of 0.0235, but note that this alpha level is divided by the number of endpoints for comparison with the smallest p-value as described in Chapter 4. We obtain the following results: \( p_1 = 0.15 \) (for Death), \( p_2 = 0.26 \) (for Transplant), \( p_3 = 0.19 \) (for Histologic Progression), and \( p_4 = 0.078 \) (for Varices). Again, there is insufficient evidence of efficacy at this interim analysis.

Again, the Hochberg Method does not allow us to stop the trial early, unlike the result we saw using the Closed Testing Method, and we see that the AFT p-values are not much different from the p-values obtained using the proportional hazard model.
5.5 Analysis Results using Fisher’s LSD Analogue

We next illustrate the application of the Fisher’s LSD analogue procedure described in the previous chapter. Again, this multiple comparison method dictates the process and order of testing events but does not limit us to any particular group sequential or survival method. Hence we shall combine Fisher’s LSD with a variety of group sequential and survival methods to see which performs best or is most illustrative for our data.

5.5.1 Lan-DeMets Error Spending Function Simulating O’Brien-Fleming Shape with WLW Method

As before, we select a Lan-DeMets error spending function that will simulate the O’Brien-Fleming method with $\alpha=0.05$. The tabled alpha-levels and critical values earlier in this chapter (Table 5.2) will be utilized here as well. Results are presented in Table 5.4:

<table>
<thead>
<tr>
<th>Analysis</th>
<th>Timepoint (Months)</th>
<th>$\alpha$-level</th>
<th>Overall Z-test (Across 4 Endpoints)</th>
<th>Pairwise Test Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>12</td>
<td>&lt;0.0001</td>
<td>1.22 &lt; 3.75 =&gt; do not stop trial</td>
<td>endpoint 1: NA</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>endpoint 2: NA</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>endpoint 3: NA</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>endpoint 4: NA</td>
</tr>
<tr>
<td>2</td>
<td>36</td>
<td>0.0235</td>
<td>2.51 &gt; 1.985 =&gt; stop trial, test individual endpoints at same $\alpha$-level.</td>
<td>endpoint 1: z=1.45</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>endpoint 2: z=1.13</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>endpoint 3: z=1.37</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>endpoint 4: z=1.80</td>
</tr>
<tr>
<td>Final</td>
<td>48</td>
<td>0.0264</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Clearly these results are more informative than those obtained using the closed testing method and the Bonferroni-Holm method described earlier in this chapter. As before, we stop the trial at interim analysis two, but now we are able to make single-endpoint inferences immediately. Unfortunately, none of the single-endpoint tests exceed the
critical value $z = 1.985$, so we cannot conclude that active is superior to placebo for any of the individual endpoints.

### 5.5.2 Lan-DeMets Error Spending Function Simulating Pocock Shape with WLW Method

Using a Lan-DeMets error spending function of $f(t) = \alpha \log(1+(e-1) t)$ to simulate the Pocock method with $\alpha=0.05$, we will obtain approximate alpha-level boundaries and one-sided $z$-critical values for each analysis as indicated in Table 5.3. Results are presented in Table 5.5:

<table>
<thead>
<tr>
<th>Analysis</th>
<th>Timepoint (Months)</th>
<th>$\alpha$-level</th>
<th>Overall Z-test (Across 4 endpoints)</th>
<th>Pairwise Test Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>12</td>
<td>0.01787</td>
<td>$1.22 &lt; 2.100$ =&gt; do not stop trial</td>
<td>endpoint 1: NA</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>endpoint 2: NA</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>endpoint 3: NA</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>endpoint 4: NA</td>
</tr>
<tr>
<td>2</td>
<td>36</td>
<td>0.02353</td>
<td>$2.51 &gt; 1.902$ =&gt; stop trial, test individual endpoints at same $\alpha$-level.</td>
<td>endpoint 1: $z=1.45$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>endpoint 2: $z=1.13$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>endpoint 3: $z=1.37$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>endpoint 4: $z=1.80$</td>
</tr>
<tr>
<td>Final</td>
<td>48</td>
<td>0.00860</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

As before, we stop the trial at interim analysis two, but now we are able to make single-endpoint inferences immediately. Unfortunately, none of the single-endpoint tests exceed the critical value $z = 1.902$, so we cannot conclude that active is superior to placebo for any of the individual endpoints.

### 5.5.3 Lan-DeMets Error Spending Function Simulating O’Brien-Fleming with AFT Method

As before, we select a Lan-DeMets error spending function that will simulate the O’Brien-Fleming method with $\alpha=0.05$. The tabled alpha-levels and critical values earlier
in this chapter (Table 5.2) will be utilized here as well. Results are presented in Table 5.6:

Table 5.6. *Fisher’s LSD Analogue Results for AFT Method for PBC Data, using Lan-DeMets Spending Function with O’Brien-Fleming Shape*

<table>
<thead>
<tr>
<th>Analysis</th>
<th>Timepoint (Months)</th>
<th>α-level</th>
<th>Overall Z-test (Across 4 Endpoints)</th>
<th>Pairwise Test Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>12</td>
<td>&lt;0.0001</td>
<td>1.13 &lt; 3.75 =&gt; do not stop trial</td>
<td>endpoint 1: NA endpoint 2: NA endpoint 3: NA endpoint 4: NA</td>
</tr>
<tr>
<td>2</td>
<td>36</td>
<td>0.0235</td>
<td>2.82 &gt; 1.985 =&gt; stop trial, test individual endpoints at same α-level. endpoint 1: z=1.40 endpoint 2: z=1.13 endpoint 3: z=1.37 endpoint 4: z=1.80</td>
<td></td>
</tr>
<tr>
<td>Final</td>
<td>48</td>
<td>0.0264</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

As before, we stop the trial at interim analysis two, but now we are able to make single-endpoint inferences immediately. Unfortunately, none of the single-endpoint tests exceed the critical value $z = 1.985$, so we cannot conclude that active is superior to placebo for any of the individual endpoints.

5.5.4 Lan-DeMets Error Spending Function Simulating O’Brien-Fleming with Frailty Method

As before, we select a Lan-DeMets error spending function that will simulate the O’Brien-Fleming method with $\alpha=0.05$. The tabled alpha-levels and critical values earlier in this chapter (Table 5.2) will be utilized here as well. Results are presented below in Table 5.7.

As before, we stop the trial at interim analysis two, but now we are able to make single-endpoint inferences immediately. Unfortunately, none of the single-endpoint tests exceed the critical value $z = 1.985$, so we cannot conclude that active is superior to placebo for any of the individual endpoints.
Table 5.7. Fisher's LSD Analogue Results for Frailty Method for PBC Data, using Lan-DeMets Spending Function with O’Brien-Fleming Shape

<table>
<thead>
<tr>
<th>Analysis</th>
<th>Timepoint (Months)</th>
<th>α-level</th>
<th>Overall Z-test (Across 4 Endpoints)</th>
<th>Pairwise Test Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>12</td>
<td>&lt;0.0001</td>
<td>1.24 &lt; 3.75 =&gt; do not stop trial</td>
<td>endpoint 1: NA</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>endpoint 2: NA</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>endpoint 3: NA</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>endpoint 4: NA</td>
</tr>
<tr>
<td>2</td>
<td>36</td>
<td>0.0235</td>
<td>2.66 &gt; 1.985 =&gt; stop trial, test individual endpoints at same α-level.</td>
<td>endpoint 1: z=1.45</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>endpoint 2: z=1.13</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>endpoint 3: z=1.31</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>endpoint 4: z=1.76</td>
</tr>
<tr>
<td>Final</td>
<td>48</td>
<td>0.0264</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

5.6 Discussion and Summary of Results

In reviewing the above results, we find that the Closed Testing Method (CTM) and Fisher’s LSD Analogue Procedure allow us a better chance of stopping the trial early compared with the Bonferroni-Holm Method (BHM) and Hochberg Methods, since CTM and Fisher involve a global test, while BHM and Hochberg do not. If our goal was not early stopping, but rather the option to make inferences about individual endpoints, then we might prefer to utilize the BHM or Hochberg procedure. A summary of testing results is provided below in Table 5.8 for the PBC data. Note that the AG and PWP methods are missing in Table 5.8 since these methods cannot be applied to distinct multivariate data, as demonstrated earlier in this chapter. We omit summary of the Hochberg method from the tables below since results agreed with the BHM for our data example.

Since our goal may go beyond merely stopping the trial early, we compare these methods more closely to see which allows us the most inference about individual endpoints. Results are below in Table 5.9. In this table, NS indicates the test is not significant at the level dictated by the GS procedure, SIG indicates that the test is significant at that level,
Table 5.8. *Summary of Results using PBC Data*

<table>
<thead>
<tr>
<th>Analysis</th>
<th>Multiple Comparison Procedure</th>
<th>WLW – OF</th>
<th>WLW – Pocock</th>
<th>AG</th>
<th>PWP</th>
<th>AFT</th>
<th>Frailty</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>CTM</td>
<td>NS</td>
<td>NS</td>
<td>N/A</td>
<td>N/A</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td></td>
<td>BHM</td>
<td>NS</td>
<td>NS</td>
<td>N/A</td>
<td>N/A</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td></td>
<td>Fisher</td>
<td>NS</td>
<td>NS</td>
<td>N/A</td>
<td>N/A</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td>2</td>
<td>CTM</td>
<td>SIG</td>
<td>SIG</td>
<td>N/A</td>
<td>N/A</td>
<td>SIG</td>
<td>SIG</td>
</tr>
<tr>
<td></td>
<td>BHM</td>
<td>SIG</td>
<td>SIG</td>
<td>N/A</td>
<td>N/A</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td></td>
<td>Fisher</td>
<td>SIG</td>
<td>SIG</td>
<td>N/A</td>
<td>N/A</td>
<td>SIG</td>
<td>SIG</td>
</tr>
<tr>
<td>Final</td>
<td>CTM</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td>BHM</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td>Fisher</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
</tbody>
</table>

NS = not significant, do not stop the trial; SIG = significant at level required by GS procedure; NA = analysis not performed.

and N/A indicates that the test could not be performed due to rules of the testing procedure.

Table 5.9 shows that the AFT and frailty models in the CTM context performs the best, while BHM performs the worst. In the next chapter we shall explore some of these methods using data simulations in order to investigate power, robustness, adherence to type-I error rates, etc.
### Table 5.9. Summary of Results at Second Interim Analysis for each Hypothesis using PBC Data

<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>Multiple Comparison Procedure</th>
<th>Multivariate Survival Method</th>
<th>WLW – OF</th>
<th>WLW - Pocock</th>
<th>AFT</th>
<th>Frailty</th>
</tr>
</thead>
<tbody>
<tr>
<td>1234</td>
<td>CTM</td>
<td>SIG</td>
<td>SIG</td>
<td>SIG</td>
<td></td>
<td>SIG</td>
</tr>
<tr>
<td></td>
<td>BHM</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td></td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td>Fisher</td>
<td>SIG</td>
<td>SIG</td>
<td>SIG</td>
<td></td>
<td>SIG</td>
</tr>
<tr>
<td>123</td>
<td>CTM</td>
<td>SIG</td>
<td>NS</td>
<td>SIG</td>
<td></td>
<td>SIG</td>
</tr>
<tr>
<td></td>
<td>BHM</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td></td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td>Fisher</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td></td>
<td>N/A</td>
</tr>
<tr>
<td>124</td>
<td>CTM</td>
<td>SIG</td>
<td>SIG</td>
<td>SIG</td>
<td></td>
<td>SIG</td>
</tr>
<tr>
<td></td>
<td>BHM</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td></td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td>Fisher</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td></td>
<td>N/A</td>
</tr>
<tr>
<td>134</td>
<td>CTM</td>
<td>SIG</td>
<td>SIG</td>
<td>SIG</td>
<td></td>
<td>SIG</td>
</tr>
<tr>
<td></td>
<td>BHM</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td></td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td>Fisher</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td></td>
<td>N/A</td>
</tr>
<tr>
<td>234</td>
<td>CTM</td>
<td>SIG</td>
<td>NS</td>
<td>SIG</td>
<td></td>
<td>SIG</td>
</tr>
<tr>
<td></td>
<td>BHM</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td></td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td>Fisher</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td></td>
<td>N/A</td>
</tr>
<tr>
<td>12</td>
<td>CTM</td>
<td>NS</td>
<td>N/A</td>
<td>NS</td>
<td></td>
<td>NS</td>
</tr>
<tr>
<td></td>
<td>BHM</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td></td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td>Fisher</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td></td>
<td>N/A</td>
</tr>
<tr>
<td>13</td>
<td>CTM</td>
<td>NS</td>
<td>N/A</td>
<td>NS</td>
<td></td>
<td>NS</td>
</tr>
<tr>
<td></td>
<td>BHM</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td></td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td>Fisher</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td></td>
<td>N/A</td>
</tr>
<tr>
<td>14</td>
<td>CTM</td>
<td>NS</td>
<td>N/A</td>
<td>SIG</td>
<td>SIG</td>
<td>SIG</td>
</tr>
<tr>
<td></td>
<td>BHM</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td>Fisher</td>
<td>N/A</td>
<td>N/A</td>
<td>SIG</td>
<td></td>
<td>SIG</td>
</tr>
<tr>
<td>23</td>
<td>CTM</td>
<td>NS</td>
<td>N/A</td>
<td>NS</td>
<td></td>
<td>NS</td>
</tr>
<tr>
<td></td>
<td>BHM</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td></td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td>Fisher</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td></td>
<td>N/A</td>
</tr>
<tr>
<td>24</td>
<td>CTM</td>
<td>NS</td>
<td>N/A</td>
<td>SIG</td>
<td>SIG</td>
<td>SIG</td>
</tr>
<tr>
<td></td>
<td>BHM</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td>Fisher</td>
<td>N/A</td>
<td>N/A</td>
<td>SIG</td>
<td></td>
<td>SIG</td>
</tr>
<tr>
<td>34</td>
<td>CTM</td>
<td>NS</td>
<td>N/A</td>
<td>SIG</td>
<td>SIG</td>
<td>SIG</td>
</tr>
<tr>
<td></td>
<td>BHM</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td>Fisher</td>
<td>N/A</td>
<td>N/A</td>
<td>SIG</td>
<td></td>
<td>SIG</td>
</tr>
<tr>
<td>1</td>
<td>CTM</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td></td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td>BHM</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
<td></td>
<td>NS</td>
</tr>
<tr>
<td></td>
<td>Fisher</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
<td></td>
<td>NS</td>
</tr>
<tr>
<td>2</td>
<td>CTM</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td></td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td>BHM</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
<td></td>
<td>NS</td>
</tr>
<tr>
<td></td>
<td>Fisher</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
<td></td>
<td>NS</td>
</tr>
<tr>
<td>3</td>
<td>CTM</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td></td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td>BHM</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
<td></td>
<td>NS</td>
</tr>
<tr>
<td></td>
<td>Fisher</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
<td></td>
<td>NS</td>
</tr>
<tr>
<td>4</td>
<td>CTM</td>
<td>N/A</td>
<td>N/A</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td></td>
<td>BHM</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
<td></td>
<td>NS</td>
</tr>
<tr>
<td></td>
<td>Fisher</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
<td></td>
<td>NS</td>
</tr>
</tbody>
</table>
6.1 General Description of Data Simulation for PBC Data

Several of the procedures which were illustrated in Chapter 5 using the Primary Biliary Cirrhosis (PBC) dataset are explored further in simulation studies. To recall, our original dataset contained 170 patients (86 active + 84 placebo) with time-to-event data for the four outcomes of interest (death, transplant, histologic progression, varices) and a censoring variable for each outcome. We also have a dichotomous predictor (1 = high histologic stage, 0 = absent or low). We are interested in taking interim looks at the data at months 12 and 36, and then at the end of the study (month 48).

Our hypothesis of interest at each interim analysis k is that there is no difference between the two treatment groups across all four endpoints (m = 1,…,4), namely: $H_{0,k}: \beta_{m,k} = 0$ against the alternative hypothesis $H_{a,k}: \beta_{m,k} > 0$, with at least one inequality (as in Tang and Geller, 1999). Recall that the hazard ratio is defined as $HR = \exp(\beta) = \lambda_P / \lambda_A$, i.e., a regression coefficient greater than zero implies $HR > 1$, which implies that the hazard for placebo $\lambda_P$ exceeds the hazard for active $\lambda_A$. Recall also that these hypotheses are tested using a global test statistic (4.7), in the order dictated by the flowchart in Chapter 3 (Figure 3.3).

Because some of the events were too sparse, particularly at month 12 (by which point only five patients had an event), we create simulated datasets rather than sample with replacement from the original dataset. We also choose to simulate data in order to explore the behavior of our procedures under different patterns and levels of censoring.
and correlation. Each dataset was structured much like the original dataset, with 85 patients on each treatment group, four distinct outcomes of interest, a censoring variable, and a single predictor. For each simulated dataset, 85 patients were assigned to placebo and 85 to UDCA. The time-to-event variable was determined as follows:

1. Let event times \( Z_i = \{ z_1, z_2, z_3, z_4 \} \sim \text{MVN} (\mu_1, \Sigma) \) for four distinct events of interest. Using the SAS Institute-supplied IML macro MVN, generate multivariate normal data for \( i = 1 \) to \( B \) iterations with desired correlation \( \rho \); i.e., \( \text{Corr} (z_j, z_k) = \rho \) for \( 1 \leq j,k \leq 4 \), where \( j \neq k \). (See Appendix D for detailed SAS code, including the MVN macro.)

2. Repeat step 1 to generate censoring times \( c_1, c_2, c_3, c_4 \) distributed as \( \text{MVN} (\mu_2, \Sigma) \).

Note that \( c_j \) and \( z_j \) are uncorrelated. We designate that event \( j \) has occurred if time \( z_j \) occurs before censoring time \( c_j \). To achieve a desired overall censoring level, we select \( \mu_1 \) and \( \mu_2 \) so that \( P(z_j < c_j) = e = 1 - c \), where \( c \) is the desired censoring level and \( e \) is the event rate. The censoring level can be calculated as follows:

For notational simplicity, let \( X = z_j \) and \( Y = c_j \) and find \( P(X < Y) = e = 1 - c \), where \( c \) is the censoring level and \( e \) is the event rate. Suppose \( X \sim N(\mu_x, \sigma_x^2) \) and \( Y \sim N(\mu_y, \sigma_y^2) \). Recall that \( X \) and \( Y \) are independent. Let \( W = Y - X \). We want to find \( P(X < Y) = P(W > 0) \)

\[
P(W > 0) = P\left( \frac{W - E(W)}{\sqrt{\text{Var}(W)}} > \frac{0 - E(W)}{\sqrt{\text{Var}(W)}} \right).
\]

Suppose \( \mu_x = 5 \), \( \sigma_x^2 = 1 \), \( \mu_y = 7 \), and \( \sigma_y^2 = 1 \). Then \( E(W) = E(Y - X) = E(Y) - E(X) = 7 - 5 = 2 \), and \( \text{Var}(W) = \text{Var}(Y - X) = \text{Var}(Y) + \text{Var}(X) = 2 \). Then we have:

\[
P(W > 0) = P\left( \frac{W - 2}{\sqrt{2}} > \frac{0 - 2}{\sqrt{2}} \right) = P(W > -1.41) = 1 - \Phi(-1.41) = 0.92.
\]

85
We can also confirm this empirically using the SAS multivariate normal macro MVN. We selected $\mu_x = \{5,1,1,1\}$ and $\mu_y = \{7,1,1,1\}$ with variances of one and $\rho=0.5$, and generated a total of 50,000 observations. This yielded an event rate of 92.3% using $N=50,000$, i.e., we found $p(x_1<y_1)=0.923$. As expected, we also found $p(x_j<y_j)\approx 0.5$ for $j = 2,3,4$ in this example.

3. Let $e_1, \ldots, e_4$ be event indicators corresponding to event times $z_1, \ldots, z_4$. If $z_j<c_j$ then let $e_j=1$, else let $e_j=0$, where $e_j=1$ denotes an event and $e_j=0$ denotes censoring for event $j$.

4. Let $t_j = \exp(z_j)$ for $j=1,\ldots,4$. Now $t_j$ is distributed lognormally with $E(t_j)=\exp(\mu_{tj}+\sigma_{tj}^2/2)$ and $\text{Var}(t_j)=\exp(2\mu_{tj}+\sigma_{tj}^2)(\exp(\sigma_{tj}^2)-1)$. Note that $t_j$ and $t_k$ are now correlated as $\text{Corr}(t_j,t_k) = \rho_{jk}$ which is not equal to the correlation selected in step 1. We can adjust the correlation to achieve as desired level as follows:

Let $\text{Corr}(z_j,z_k)=\rho_{jk}$, as described above, where $Z_i = \{ z_1, z_2, z_3, z_4 \} \sim \text{MVN (} \mu_1, \Sigma \)$. Let $X=\exp(z_j)$ and $Y=\exp(z_k)$. When $z_j$ and $z_k$ are bivariate normal (instead of MVN, as in our case), we have this result from Mostafa and Mahmoud (1964):

$$\text{Corr}(X,Y) = \frac{\exp(\rho_{jk}\sigma_j\sigma_k)-1}{\sqrt{(\exp(\sigma_j^2)-1)(\exp(\sigma_k^2)-1)}}.$$ 

To revisit our example above, suppose $\mu_x = \{5,1,1,1\}$, $\mu_y = \{7,1,1,1\}$, $\sigma_j^2=1$, and $\rho_{jk} = 0.5$ for $1\leq j,k \leq 4$ and $j \neq k$. Applying Mostafa and Mahmoud’s formula, we calculate $\text{Corr (}X,Y\text{)} = 0.378$. This is very close to the simulated result using the MVN macro with $B=50000$ below. If we substitute $\rho_{jk} = 0.62$ instead of $\rho_{jk} = 0.5$ into the equation above, we calculate $\text{Corr (}X,Y\text{)} \approx 0.50$. 

86
A literature search did not reveal an expression for the correlation coefficient for multivariate lognormal data, but this simulation suggests that we can work backwards by testing different correlation levels on the MVN data to achieve desired correlation levels in the transformed lognormal data. This should be sufficient to design a simulation to test the behaviors of our survival and group sequential procedures at different censoring and correlation levels.

| Pearson Correlation Coefficients, N = 50000 | Prob > |r| under H0: Rho=0 |
|---------------------------------------------|-----------------------------------------------|
| lntime1                                    | lntime2 | lntime3 | lntime4 |
| lntime1 | 1.00000 | 0.36940 | 0.37737 | 0.36202 |
|        | <.0001 | <.0001 | <.0001 | <.0001 |
| lntime2 | 0.36940 | 1.00000 | 0.36088 | 0.36872 |
|        | <.0001 | <.0001 | <.0001 | <.0001 |
| lntime3 | 0.37737 | 0.36088 | 1.00000 | 0.37213 |
|        | <.0001 | <.0001 | <.0001 | <.0001 |
| lntime4 | 0.36202 | 0.36872 | 0.37213 | 1.00000 |
|        | <.0001 | <.0001 | <.0001 | <.0001 |

5. For simulation under the null hypothesis (Ho: $\beta_k = 0$):

- We wish to create a total of “B” simulated datasets, where each dataset contains 170 subjects, with 85 assigned to active group and 85 assigned to placebo.
- For each iteration (i=1 to B), we assign the first 85 subjects to active and the second 85 subjects to placebo.
- Regardless of treatment, we sample from N($\mu_i$, $\Sigma$), following steps 1-4 above.
- Note that event times are still correlated, as desired.
For simulation under the alternative hypothesis \( (H_a: \beta_k > 0) \):

- We wish to create a total of “B” simulated datasets, where each dataset contains 170 subjects, with 85 assigned to active group and 85 assigned to placebo.
- For subjects assigned to active, we sample from \( N(\mu_a, \Sigma) \). For subjects assigned to placebo, we sample from \( N(\mu_p, \Sigma) \).
- Note that event times are still correlated, as desired.

Our simulation was then conducted as follows:

Let \( B \) = Number of iterations for the simulation procedure (e.g., \( B = 5000 \))

Let \( R_1 \) = Number of rejections at interim look 1 (month 12), \( R_2 \) = rejections at interim look 2 (month 36), and \( R_3 \) = rejections at final analysis.

Let \( c_1 \) = critical value at interim 1, \( c_2 \) = critical value at interim 2, \( c_3 \) = critical value at final analysis. These are determined based on our choice of alpha-spending functions.

For iteration \( i \) (\( i = 1 \) to \( B \)):

Step 1: To sample under the null hypothesis, note that each patient has equal probability of being assigned active or placebo for each record in sample \( i \).

Step 2: For interim look 1 (month 12), we consider only events that occur at or before month 12. Events after month 12 are coded as censored in sample \( i \). We then run our marginal proportional hazards model using the Wei et al robust estimate of variance to obtain a global test statistic \( Z_1 \) for the composite hypothesis. If \( Z_1 > c_1 \) then increment \( R_1 \) by one.

Step 3: If we reject in step 2 …

Stop the study for sample \( i \), and proceed to iteration \( i+1 \).

Sample \( i \) will not be included in the calculations of \( R_2 \) and \( R_3 \).
If we fail to reject in step 2 …

Proceed to interim look 2 (month 36).

Calculate $Z_2$ as in step 2, comparing with critical value $c_2$.

If $Z_2 > c_2$ then increment $R_2$ by one.

Step 4: If we reject in step 3 …

Stop the study for sample i, and proceed to iteration $i+1$.

Sample i will not be included in the calculation of $R_3$.

If we fail to reject in step 3 …

Proceed to final analysis (month 48)

Calculate $Z_3$ as in step 2, comparing with critical value $c_3$.

If $Z_3 > c_3$ then increment $R_3$ by one.

Step 5: If $i < B$, then increment $i$ by one, and return to step 1.

If $i = B$, stop and calculate error rates.

For each simulated dataset $j$ (for $j=1$ to 5000), a z-statistic $Z_{j,t}$ \( \{1,2,3,4\} \) was calculated for interim analyses $t = 1, 2$, and the final analysis using the same marginal proportional model (Wei et al, 1989) described in Chapter 4 and illustrated in Chapter 5. We note that this same test statistic is utilized in stage one of the Fisher’s LSD Analogue Procedure discussed later in this chapter.

As noted above, censoring times were generated using a multivariate normal distribution, just as were the event times, to generate a random censoring pattern. This was selected in order to be as general as possible to apply to different types of data. The most typical cause of censoring is if a subject drops out of the study, but it is common to see separate censoring on different outcomes, sometimes seemingly at random due to poor data
collection or data management procedures, and sometimes for legitimate clinical reasons. As an example, suppose a subject has just experienced outcomes A and B, and then drops out of the study. He is then censored for the other outcomes under study (C and D). As another example, suppose we are considering time to event for a kidney tumor and time to event for kidney transplant or failure within the same subject. If the organ fails or a transplant occurs, the subject may no longer be susceptible to the tumor. We might then consider him censored at the time of transplant or organ failure/removal, since we cannot know whether or not the tumor would have occurred. Another option for data simulation would be to consider censoring of a whole subject at a time, but this is a special case of the above and is expected to be less common.

6.2 Data Simulation Results for PBC Data under Null Hypothesis

6.2.1 O’Brien-Fleming and Pocock Shape with Various Censoring Levels

Critical values were used for each interim timepoint as displayed previously in Table 5.2 from Chapter 5; i.e., using the Lan-DeMets error spending function of \( f(t) = \min \left( 2 - 2\Phi(z_{\alpha/2}/\sqrt{t}), \alpha \right) \) to simulate the O’Brien-Fleming method with \( \alpha=0.05 \). To simulate the Pocock method, critical values were used for each interim timepoint as displayed in Table 5.3; i.e., using the Lan-DeMets error spending function of \( f(t) = \min \left( \alpha \log(1+(e-1) t ), \alpha \right) \) with \( \alpha=0.05 \). After simulating 5000 datasets, the proportion of times that \( Z_{jt \in \{1,2,3,4\}} \) exceeded the critical value was calculated for each interim analysis. Figure 6.1 below illustrates the distribution of event times under zero correlation for one of the four distinct events. As can be seen, the event times are lognormally distributed with noticeable right skew. Simulations were run for different levels of correlation and censoring.
Figure 6.1. *Histogram of Lognormally Distributed Data for One Event Time*

N=5000 simulations, multivariate lognormal data with \( \mu \approx (23, 23, 23, 23) \) and \( \Sigma \approx \sigma^2 \cdot I \), where \( \sigma^2 \approx 42 \) and I is a 4x4 identity matrix

Results of the first simulation are displayed below in Table 6.1. Since the data were simulated under the null hypothesis, we should expect to see an overall type I error rate close to the nominal 5% level. For both the O'Brien-Fleming and Pocock, we see that the cumulative rejection by the final analysis stays very close to the 0.05 level for low/moderate correlation levels until one exceeds 60% censoring; across the correlation levels, one sees very similar rejection rates at the low/moderate correlation levels and consistently higher rejection rates when the correlation level is higher (rho=0.75). Cumulative rejection rates for the Pocock method are consistently slightly smaller than for the O’Brien-Fleming method, as shown in Figures 6.2 and 6.3.
Table 6.1. *Simulation under Null Hypothesis for WLW Method with Correlation rho = 0.0*

<table>
<thead>
<tr>
<th>Censoring Proportion</th>
<th>Analysis</th>
<th>O’Brien-Fleming Shape</th>
<th>Pocock Shape</th>
<th>Median Observed Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>0%</td>
<td>1</td>
<td>0.0048</td>
<td>0.0048</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.0231</td>
<td>0.0279</td>
<td>0.0236</td>
</tr>
<tr>
<td></td>
<td>Final</td>
<td>0.0230</td>
<td>0.0509</td>
<td>0.0500</td>
</tr>
<tr>
<td>20%</td>
<td>1</td>
<td>0.0066</td>
<td>0.0066</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.0244</td>
<td>0.0310</td>
<td>0.0236</td>
</tr>
<tr>
<td></td>
<td>Final</td>
<td>0.0190</td>
<td>0.0500</td>
<td>0.0500</td>
</tr>
<tr>
<td>40%</td>
<td>1</td>
<td>0.0070</td>
<td>0.0070</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.0224</td>
<td>0.0294</td>
<td>0.0236</td>
</tr>
<tr>
<td></td>
<td>Final</td>
<td>0.0200</td>
<td>0.0494</td>
<td>0.0500</td>
</tr>
<tr>
<td>60%</td>
<td>1</td>
<td>0.0070</td>
<td>0.0070</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.0242</td>
<td>0.0312</td>
<td>0.0236</td>
</tr>
<tr>
<td></td>
<td>Final</td>
<td>0.0188</td>
<td>0.0500</td>
<td>0.0500</td>
</tr>
<tr>
<td>80%</td>
<td>1</td>
<td>0.0110</td>
<td>0.0110</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.0247</td>
<td>0.0357</td>
<td>0.0236</td>
</tr>
<tr>
<td></td>
<td>Final</td>
<td>0.0182</td>
<td>0.0539</td>
<td>0.0500</td>
</tr>
<tr>
<td>90%</td>
<td>1</td>
<td>0.0288</td>
<td>0.0288</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.0220</td>
<td>0.0508</td>
<td>0.0236</td>
</tr>
<tr>
<td></td>
<td>Final</td>
<td>0.0179</td>
<td>0.0687</td>
<td>0.0500</td>
</tr>
<tr>
<td>95%</td>
<td>1</td>
<td>0.0755</td>
<td>0.0755</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.0201</td>
<td>0.0956</td>
<td>0.0236</td>
</tr>
<tr>
<td></td>
<td>Final</td>
<td>0.0179</td>
<td>0.1135</td>
<td>0.0500</td>
</tr>
</tbody>
</table>

N=5000 simulations, multiv. lognormal data with \( \mu \approx (23, 23, 23, 23) \) and \( \Sigma \approx \sigma^2 \cdot I \), where \( \sigma^2 \approx 42 \) and I is a 4x4 identity matrix.
Table 6.2. *Simulation under Null Hypothesis for WLW Method with Correlation rho = 0.25*

<table>
<thead>
<tr>
<th>Censoring Proportion</th>
<th>O'Brien-Fleming Shape</th>
<th>Pocock Shape</th>
<th>Median Observed Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>0%</td>
<td>0.0066</td>
<td>0.0066</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td></td>
<td>0.0213</td>
<td>0.0279</td>
<td>0.0236</td>
</tr>
<tr>
<td></td>
<td>0.0224</td>
<td>0.0503</td>
<td>0.0500</td>
</tr>
<tr>
<td>20%</td>
<td>0.0088</td>
<td>0.0088</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td></td>
<td>0.0212</td>
<td>0.0300</td>
<td>0.0236</td>
</tr>
<tr>
<td></td>
<td>0.0198</td>
<td>0.0498</td>
<td>0.0500</td>
</tr>
<tr>
<td>40%</td>
<td>0.0104</td>
<td>0.0104</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td></td>
<td>0.0200</td>
<td>0.0304</td>
<td>0.0236</td>
</tr>
<tr>
<td></td>
<td>0.0181</td>
<td>0.0485</td>
<td>0.0500</td>
</tr>
<tr>
<td>60%</td>
<td>0.0106</td>
<td>0.0106</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td></td>
<td>0.0224</td>
<td>0.0330</td>
<td>0.0236</td>
</tr>
<tr>
<td></td>
<td>0.0176</td>
<td>0.0506</td>
<td>0.0500</td>
</tr>
<tr>
<td>80%</td>
<td>0.0152</td>
<td>0.0152</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td></td>
<td>0.0252</td>
<td>0.0404</td>
<td>0.0236</td>
</tr>
<tr>
<td></td>
<td>0.0165</td>
<td>0.0569</td>
<td>0.0500</td>
</tr>
<tr>
<td>90%</td>
<td>0.0314</td>
<td>0.0314</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td></td>
<td>0.0221</td>
<td>0.0535</td>
<td>0.0236</td>
</tr>
<tr>
<td></td>
<td>0.0188</td>
<td>0.0723</td>
<td>0.0500</td>
</tr>
<tr>
<td>95%</td>
<td>0.0741</td>
<td>0.0741</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td></td>
<td>0.0246</td>
<td>0.0987</td>
<td>0.0236</td>
</tr>
<tr>
<td></td>
<td>0.0181</td>
<td>0.1168</td>
<td>0.0500</td>
</tr>
</tbody>
</table>

N=5000 simulations, multiv. lognormal data with $\mu \approx (23, 23, 23, 23)$ and $\Sigma$ is a 4x4 covariance matrix with $\rho=0.25$ and $\sigma=6.5$. 
Table 6.3. *Simulation under Null Hypothesis for WLW Method with Correlation rho = 0.50*

<table>
<thead>
<tr>
<th>Censoring Proportion</th>
<th>Analysis</th>
<th>O’Brien-Fleming Shape</th>
<th>Pocock Shape</th>
<th>Median Observed Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>0%</td>
<td>1</td>
<td>0.0114</td>
<td>0.0114</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.0235</td>
<td>0.0349</td>
<td>0.0236</td>
</tr>
<tr>
<td></td>
<td>Final</td>
<td>0.0186</td>
<td>0.0535</td>
<td>0.0500</td>
</tr>
<tr>
<td>20%</td>
<td>1</td>
<td>0.0144</td>
<td>0.0144</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.0203</td>
<td>0.0347</td>
<td>0.0236</td>
</tr>
<tr>
<td></td>
<td>Final</td>
<td>0.0186</td>
<td>0.0533</td>
<td>0.0500</td>
</tr>
<tr>
<td>40%</td>
<td>1</td>
<td>0.0162</td>
<td>0.0162</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.0215</td>
<td>0.0377</td>
<td>0.0236</td>
</tr>
<tr>
<td></td>
<td>Final</td>
<td>0.0141</td>
<td>0.0518</td>
<td>0.0500</td>
</tr>
<tr>
<td>60%</td>
<td>1</td>
<td>0.0168</td>
<td>0.0168</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.0218</td>
<td>0.0386</td>
<td>0.0236</td>
</tr>
<tr>
<td></td>
<td>Final</td>
<td>0.0185</td>
<td>0.0571</td>
<td>0.0500</td>
</tr>
<tr>
<td>80%</td>
<td>1</td>
<td>0.0232</td>
<td>0.0232</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.0223</td>
<td>0.0455</td>
<td>0.0236</td>
</tr>
<tr>
<td></td>
<td>Final</td>
<td>0.0182</td>
<td>0.0637</td>
<td>0.0500</td>
</tr>
<tr>
<td>90%</td>
<td>1</td>
<td>0.0412</td>
<td>0.0412</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.0242</td>
<td>0.0654</td>
<td>0.0236</td>
</tr>
<tr>
<td></td>
<td>Final</td>
<td>0.0195</td>
<td>0.0849</td>
<td>0.0500</td>
</tr>
<tr>
<td>95%</td>
<td>1</td>
<td>0.0839</td>
<td>0.0839</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.0207</td>
<td>0.1046</td>
<td>0.0236</td>
</tr>
<tr>
<td></td>
<td>Final</td>
<td>0.0185</td>
<td>0.1231</td>
<td>0.0500</td>
</tr>
</tbody>
</table>

N=5000 simulations, multiv. lognormal data with \( \mu \approx (23, 23, 23, 23) \) and \( \Sigma \) is a 4x4 covariance matrix with \( \rho=0.50 \) and \( \sigma=6.5 \).
Table 6.4. Simulation under Null Hypothesis for WLW Method with Correlation $\rho = 0.75$

<table>
<thead>
<tr>
<th>Censoring Proportion</th>
<th>Analysis</th>
<th>O'Brien-Fleming Shape</th>
<th>Pocock Shape</th>
<th>Median Observed Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>0%</td>
<td>1</td>
<td>0.0254</td>
<td>0.0254</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.0236</td>
<td>0.0490</td>
<td>0.0236</td>
</tr>
<tr>
<td></td>
<td>Final</td>
<td>0.0200</td>
<td>0.0690</td>
<td>0.0500</td>
</tr>
<tr>
<td>20%</td>
<td>1</td>
<td>0.0322</td>
<td>0.0322</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.0182</td>
<td>0.0504</td>
<td>0.0236</td>
</tr>
<tr>
<td></td>
<td>Final</td>
<td>0.0164</td>
<td>0.0668</td>
<td>0.0500</td>
</tr>
<tr>
<td>40%</td>
<td>1</td>
<td>0.0336</td>
<td>0.0336</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.0197</td>
<td>0.0533</td>
<td>0.0236</td>
</tr>
<tr>
<td></td>
<td>Final</td>
<td>0.0135</td>
<td>0.0668</td>
<td>0.0500</td>
</tr>
<tr>
<td>60%</td>
<td>1</td>
<td>0.0365</td>
<td>0.0365</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.0212</td>
<td>0.0577</td>
<td>0.0236</td>
</tr>
<tr>
<td></td>
<td>Final</td>
<td>0.0178</td>
<td>0.0755</td>
<td>0.0500</td>
</tr>
<tr>
<td>80%</td>
<td>1</td>
<td>0.0459</td>
<td>0.0459</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.0203</td>
<td>0.0662</td>
<td>0.0236</td>
</tr>
<tr>
<td></td>
<td>Final</td>
<td>0.0182</td>
<td>0.0844</td>
<td>0.0500</td>
</tr>
<tr>
<td>90%</td>
<td>1</td>
<td>0.0748</td>
<td>0.0748</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.0192</td>
<td>0.0940</td>
<td>0.0236</td>
</tr>
<tr>
<td></td>
<td>Final</td>
<td>0.0165</td>
<td>0.1105</td>
<td>0.0500</td>
</tr>
<tr>
<td>95%</td>
<td>1</td>
<td>0.1282</td>
<td>0.1282</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.0185</td>
<td>0.1467</td>
<td>0.0236</td>
</tr>
<tr>
<td></td>
<td>Final</td>
<td>0.0135</td>
<td>0.1602</td>
<td>0.0500</td>
</tr>
</tbody>
</table>

N=5000 simulations, multiv. lognormal data with $\mu \approx (23, 23, 23, 23)$ and $\Sigma$ is a 4x4 covariance matrix with $\rho=0.75$ and $\sigma \approx 6.5$. 
Figure 6.2. Cumulative Rejection Rate under Null Hypothesis with O'Brien-Fleming Shape
Figure 6.3. Cumulative Rejection Rate under Null Hypothesis with Pocock Shape
6.3 Data Simulation Results for PBC Data under Null Hypothesis with Non-Constant Correlation Structures

As in section 6.2, critical values were used for each interim timepoint as displayed previously in Tables 5.2 and 5.3 from Chapter 5. Instead of simulating four distinct endpoints with constant correlation levels, each simulated dataset contained time to event data for just three distinct events for each subject. This was chosen in order to reduce the number of combinations of correlation structures of interest. Seven different non-constant correlation structures were selected: LMH, LLM, LLH, MML, MMH, HHL, and HHM, where L, M, and H denote low (rho=0.25), medium (rho=0.50), and high (rho=0.75) levels of correlation, respectively. More specifically, LMH indicates that events 1 & 2 have rho=0.25, events 1 & 3 have rho=0.50, and events 2 & 3 have rho=0.75. The goal in selecting these different structures was to try to simulate a more realistic or unusual set of correlations.

After simulating 5000 datasets, the proportion of times that \( Z_{jt\{1,2,3\}} \) exceeded the critical value was calculated for each interim analysis. Simulations were run for the seven different non-constant levels of correlation and for four different levels of censoring (0% to 95%).

Results of the these simulations are displayed below in Table 6.5 (for the O'Brien-Fleming) and in Table 6.6 (for the Pocock method). Since the data were simulated under the null hypothesis, we should expect to see a overall type I error rate close to the nominal 5% level. We see that for both the O'Brien-Fleming and Pocock shapes, the cumulative rejection by the final analysis is higher than it was in the earlier, constant-rho simulations. At 0% censoring for the O'Brien-Fleming, for instance, the cumulative
rejection rate ranges from 6.5% (MML) to 8.0% (HHL) compared with the lower levels of 5% to 6.9% seen under the constant correlation structures. The same pattern is seen for the Pocock method, and as before the rejection rate increases monotonically as the censoring level increases. Since the patterns for non-constant correlation are so similar to what was seen in the previous section (albeit at higher levels of rejection), we do not focus additional attention on unusual or non-constant correlation structures. These results are also illustrated below in Figures 6.4 and 6.5 for the O’Brien-Fleming and Pocock methods, respectively.

Figure 6.4. Cumulative Rejection Rate under Null Hypothesis with O’Brien-Fleming Shape and Non-Constant Correlation Structure
Figure 6.5. *Cumulative Rejection Rate under Null Hypothesis with Pocock Shape and Non-Constant Correlation Structure*
Table 6.5. **Simulation under Null Hypothesis for WLW Method with O’Brien-Fleming Shape and Non-Constant Correlation Structures**

<table>
<thead>
<tr>
<th>Correlation Structure</th>
<th>Analysis</th>
<th>Expected Cumulative Rate</th>
<th>Obs Rej. Rate</th>
<th>Cumul. Rej. Rate</th>
<th>0 %</th>
<th>40 %</th>
<th>80 %</th>
<th>95 %</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Obs Rej. Rate</td>
<td>Cumul. Rej. Rate</td>
<td>Obs Rej. Rate</td>
<td>Cumul. Rej. Rate</td>
</tr>
<tr>
<td>LMH</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Obs Rej. Rate</td>
<td>Cumul. Rej. Rate</td>
<td>Obs Rej. Rate</td>
<td>Cumul. Rej. Rate</td>
</tr>
<tr>
<td>1</td>
<td>&lt;0.0001</td>
<td>0.0322</td>
<td>0.0322</td>
<td></td>
<td>0.0411</td>
<td>0.0411</td>
<td>0.0524</td>
<td>0.0524</td>
</tr>
<tr>
<td>2</td>
<td>0.0236</td>
<td>0.0205</td>
<td>0.0527</td>
<td></td>
<td>0.0206</td>
<td>0.0617</td>
<td>0.0236</td>
<td>0.0760</td>
</tr>
<tr>
<td>Final</td>
<td>0.0500</td>
<td>0.0230</td>
<td>0.0757</td>
<td></td>
<td>0.0164</td>
<td>0.0781</td>
<td>0.0171</td>
<td>0.0931</td>
</tr>
<tr>
<td>LLM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Obs Rej. Rate</td>
<td>Cumul. Rej. Rate</td>
<td>Obs Rej. Rate</td>
<td>Cumul. Rej. Rate</td>
</tr>
<tr>
<td>1</td>
<td>&lt;0.0001</td>
<td>0.0248</td>
<td>0.0248</td>
<td></td>
<td>0.0318</td>
<td>0.0318</td>
<td>0.0419</td>
<td>0.0419</td>
</tr>
<tr>
<td>2</td>
<td>0.0236</td>
<td>0.0226</td>
<td>0.0474</td>
<td></td>
<td>0.0192</td>
<td>0.0510</td>
<td>0.0238</td>
<td>0.0657</td>
</tr>
<tr>
<td>Final</td>
<td>0.0500</td>
<td>0.0225</td>
<td>0.0699</td>
<td></td>
<td>0.0187</td>
<td>0.0697</td>
<td>0.0197</td>
<td>0.0854</td>
</tr>
<tr>
<td>LLH</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Obs Rej. Rate</td>
<td>Cumul. Rej. Rate</td>
<td>Obs Rej. Rate</td>
<td>Cumul. Rej. Rate</td>
</tr>
<tr>
<td>1</td>
<td>&lt;0.0001</td>
<td>0.0272</td>
<td>0.0272</td>
<td></td>
<td>0.0349</td>
<td>0.0349</td>
<td>0.0459</td>
<td>0.0459</td>
</tr>
<tr>
<td>2</td>
<td>0.0236</td>
<td>0.0228</td>
<td>0.0500</td>
<td></td>
<td>0.0218</td>
<td>0.0567</td>
<td>0.0249</td>
<td>0.0708</td>
</tr>
<tr>
<td>Final</td>
<td>0.0500</td>
<td>0.0210</td>
<td>0.0710</td>
<td></td>
<td>0.0165</td>
<td>0.0732</td>
<td>0.0198</td>
<td>0.0906</td>
</tr>
<tr>
<td>MML</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Obs Rej. Rate</td>
<td>Cumul. Rej. Rate</td>
<td>Obs Rej. Rate</td>
<td>Cumul. Rej. Rate</td>
</tr>
<tr>
<td>1</td>
<td>&lt;0.0001</td>
<td>0.0224</td>
<td>0.0224</td>
<td></td>
<td>0.0298</td>
<td>0.0298</td>
<td>0.0417</td>
<td>0.0417</td>
</tr>
<tr>
<td>2</td>
<td>0.0236</td>
<td>0.0237</td>
<td>0.0461</td>
<td></td>
<td>0.0194</td>
<td>0.0492</td>
<td>0.0202</td>
<td>0.0619</td>
</tr>
<tr>
<td>Final</td>
<td>0.0500</td>
<td>0.0187</td>
<td>0.0648</td>
<td></td>
<td>0.0196</td>
<td>0.0688</td>
<td>0.0190</td>
<td>0.0809</td>
</tr>
<tr>
<td>MMH</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Obs Rej. Rate</td>
<td>Cumul. Rej. Rate</td>
<td>Obs Rej. Rate</td>
<td>Cumul. Rej. Rate</td>
</tr>
<tr>
<td>1</td>
<td>&lt;0.0001</td>
<td>0.0353</td>
<td>0.0353</td>
<td></td>
<td>0.0441</td>
<td>0.0441</td>
<td>0.0570</td>
<td>0.0570</td>
</tr>
<tr>
<td>2</td>
<td>0.0236</td>
<td>0.0222</td>
<td>0.0575</td>
<td></td>
<td>0.0218</td>
<td>0.0659</td>
<td>0.0235</td>
<td>0.0805</td>
</tr>
<tr>
<td>Final</td>
<td>0.0500</td>
<td>0.0191</td>
<td>0.0766</td>
<td></td>
<td>0.0141</td>
<td>0.0800</td>
<td>0.0156</td>
<td>0.0961</td>
</tr>
<tr>
<td>HHL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Obs Rej. Rate</td>
<td>Cumul. Rej. Rate</td>
<td>Obs Rej. Rate</td>
<td>Cumul. Rej. Rate</td>
</tr>
<tr>
<td>1</td>
<td>&lt;0.0001</td>
<td>0.0379</td>
<td>0.0379</td>
<td></td>
<td>0.0472</td>
<td>0.0472</td>
<td>0.0606</td>
<td>0.0606</td>
</tr>
<tr>
<td>2</td>
<td>0.0236</td>
<td>0.0235</td>
<td>0.0614</td>
<td></td>
<td>0.0222</td>
<td>0.0694</td>
<td>0.0204</td>
<td>0.0810</td>
</tr>
<tr>
<td>Final</td>
<td>0.0500</td>
<td>0.0187</td>
<td>0.0801</td>
<td></td>
<td>0.0157</td>
<td>0.0851</td>
<td>0.0154</td>
<td>0.0964</td>
</tr>
<tr>
<td>HHM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Obs Rej. Rate</td>
<td>Cumul. Rej. Rate</td>
<td>Obs Rej. Rate</td>
<td>Cumul. Rej. Rate</td>
</tr>
<tr>
<td>1</td>
<td>&lt;0.0001</td>
<td>0.0361</td>
<td>0.0361</td>
<td></td>
<td>0.0466</td>
<td>0.0466</td>
<td>0.0626</td>
<td>0.0626</td>
</tr>
<tr>
<td>2</td>
<td>0.0236</td>
<td>0.0212</td>
<td>0.0573</td>
<td></td>
<td>0.0208</td>
<td>0.0674</td>
<td>0.0220</td>
<td>0.0846</td>
</tr>
<tr>
<td>Final</td>
<td>0.0500</td>
<td>0.0199</td>
<td>0.0772</td>
<td></td>
<td>0.0165</td>
<td>0.0839</td>
<td>0.0142</td>
<td>0.0988</td>
</tr>
</tbody>
</table>

Note: L, M, H denote low (rho=0.25), medium (rho=0.50), and high (rho=0.75) correlations between events within a simulation. N=5000 simulations, multiv. lognormal data with μ ≈ (23, 23, 23) and Σ is a 3x3 covariance matrix with σ≈6.5 and rho as noted above.
Table 6.6. Simulation under Null Hypothesis for WLW Method with Pocock Shape and Non-Constant Correlation Structures

<table>
<thead>
<tr>
<th>Correlation Structure</th>
<th>Analysis</th>
<th>Expected Cumulative Rate</th>
<th>Obs Rej. Rate</th>
<th>Cumul. Rej. Rate</th>
<th>Censoring Level</th>
<th>Obs Rej. Rate</th>
<th>Cumul. Rej. Rate</th>
<th>Obs Rej. Rate</th>
<th>Cumul. Rej. Rate</th>
<th>Obs Rej. Rate</th>
<th>Cumul. Rej. Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>0 %</td>
<td>40 %</td>
<td>80 %</td>
<td>95 %</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LMH</td>
<td>1</td>
<td>0.0179</td>
<td>0.0344</td>
<td>0.0344</td>
<td>0.0433</td>
<td>0.0433</td>
<td>0.0542</td>
<td>0.0542</td>
<td>0.1494</td>
<td>0.1494</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.0414</td>
<td>0.0255</td>
<td>0.0599</td>
<td>0.0230</td>
<td>0.0663</td>
<td>0.0275</td>
<td>0.0817</td>
<td>0.0238</td>
<td>0.1732</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Final</td>
<td>0.0500</td>
<td>0.0015</td>
<td>0.0614</td>
<td>0.0000</td>
<td>0.0663</td>
<td>0.0000</td>
<td>0.0817</td>
<td>0.0000</td>
<td>0.1732</td>
<td></td>
</tr>
<tr>
<td>LLM</td>
<td>1</td>
<td>0.0179</td>
<td>0.0262</td>
<td>0.0262</td>
<td>0.0330</td>
<td>0.0330</td>
<td>0.0429</td>
<td>0.0429</td>
<td>0.1327</td>
<td>0.1327</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.0414</td>
<td>0.0275</td>
<td>0.0537</td>
<td>0.0242</td>
<td>0.0572</td>
<td>0.0311</td>
<td>0.0740</td>
<td>0.0276</td>
<td>0.1603</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Final</td>
<td>0.0500</td>
<td>0.0013</td>
<td>0.0550</td>
<td>0.0000</td>
<td>0.0572</td>
<td>0.0000</td>
<td>0.0740</td>
<td>0.0000</td>
<td>0.1603</td>
<td></td>
</tr>
<tr>
<td>LLH</td>
<td>1</td>
<td>0.0179</td>
<td>0.0296</td>
<td>0.0296</td>
<td>0.0367</td>
<td>0.0367</td>
<td>0.0472</td>
<td>0.0472</td>
<td>0.1452</td>
<td>0.1452</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.0414</td>
<td>0.0280</td>
<td>0.0576</td>
<td>0.0251</td>
<td>0.0618</td>
<td>0.0315</td>
<td>0.0787</td>
<td>0.0249</td>
<td>0.1701</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Final</td>
<td>0.0500</td>
<td>0.0008</td>
<td>0.0584</td>
<td>0.0000</td>
<td>0.0618</td>
<td>0.0002</td>
<td>0.0789</td>
<td>0.0000</td>
<td>0.1701</td>
<td></td>
</tr>
<tr>
<td>MML</td>
<td>1</td>
<td>0.0179</td>
<td>0.0242</td>
<td>0.0242</td>
<td>0.0321</td>
<td>0.0321</td>
<td>0.0437</td>
<td>0.0437</td>
<td>0.1271</td>
<td>0.1271</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.0414</td>
<td>0.0285</td>
<td>0.0527</td>
<td>0.0233</td>
<td>0.0554</td>
<td>0.0247</td>
<td>0.0684</td>
<td>0.0233</td>
<td>0.1504</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Final</td>
<td>0.0500</td>
<td>0.0004</td>
<td>0.0531</td>
<td>0.0000</td>
<td>0.0554</td>
<td>0.0000</td>
<td>0.0684</td>
<td>0.0000</td>
<td>0.1504</td>
<td></td>
</tr>
<tr>
<td>MMH</td>
<td>1</td>
<td>0.0179</td>
<td>0.0379</td>
<td>0.0379</td>
<td>0.0467</td>
<td>0.0467</td>
<td>0.0592</td>
<td>0.0592</td>
<td>0.1453</td>
<td>0.1453</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.0414</td>
<td>0.0270</td>
<td>0.0649</td>
<td>0.0252</td>
<td>0.0719</td>
<td>0.0272</td>
<td>0.0864</td>
<td>0.0235</td>
<td>0.1688</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Final</td>
<td>0.0500</td>
<td>0.0013</td>
<td>0.0662</td>
<td>0.0002</td>
<td>0.0721</td>
<td>0.0000</td>
<td>0.0864</td>
<td>0.0000</td>
<td>0.1688</td>
<td></td>
</tr>
<tr>
<td>HHL</td>
<td>1</td>
<td>0.0179</td>
<td>0.0421</td>
<td>0.0421</td>
<td>0.0498</td>
<td>0.0498</td>
<td>0.0624</td>
<td>0.0624</td>
<td>0.1504</td>
<td>0.1504</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.0414</td>
<td>0.0276</td>
<td>0.0697</td>
<td>0.0255</td>
<td>0.0753</td>
<td>0.0235</td>
<td>0.0859</td>
<td>0.0192</td>
<td>0.1696</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Final</td>
<td>0.0500</td>
<td>0.0009</td>
<td>0.0706</td>
<td>0.0000</td>
<td>0.0753</td>
<td>0.0000</td>
<td>0.0859</td>
<td>0.0000</td>
<td>0.1696</td>
<td></td>
</tr>
<tr>
<td>HHM</td>
<td>1</td>
<td>0.0179</td>
<td>0.0397</td>
<td>0.0397</td>
<td>0.0492</td>
<td>0.0492</td>
<td>0.0642</td>
<td>0.0642</td>
<td>0.1527</td>
<td>0.1527</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.0414</td>
<td>0.0271</td>
<td>0.0668</td>
<td>0.0240</td>
<td>0.0732</td>
<td>0.0248</td>
<td>0.0890</td>
<td>0.0199</td>
<td>0.1726</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Final</td>
<td>0.0500</td>
<td>0.0006</td>
<td>0.0674</td>
<td>0.0000</td>
<td>0.0732</td>
<td>0.0000</td>
<td>0.0890</td>
<td>0.0000</td>
<td>0.1726</td>
<td></td>
</tr>
</tbody>
</table>

Note: L, M, H denote low (rho=0.25), medium (rho=0.50), and high (rho=0.75) correlations between events within a simulation. N=5000 simulations, multiv. lognormal data with \( \mu \approx (23, 23, 23) \) and \( \Sigma \) is a 3x3 covariance matrix with \( \sigma \approx 6.5 \) and rho as noted above.
6.4 Simulations under Alternative Hypotheses

Using the same critical values for the O’Brien-Fleming and Pocock methods as used in sections 6.2 and 6.3, we simulate multivariate lognormally distributed data under three different alternative hypotheses, displaying the results alongside the null hypothesis results for the usual four levels of correlation (Tables 6.7 – 6.10, for $\rho=0.0$, 0.25, 0.50, and 0.75, respectively). Under each hypothesis, subjects assigned to the placebo group were simulated with mean time to event of 23 months as in sections 6.2 and 6.3 with a standard deviation of 6.5 months (i.e., $\mu_1=23$, $\sigma_1=6.5$), whereas subjects assigned to the active group were simulated with mean time to event as indicated below. The three alternative hypotheses selected were as follows:

1. $\Delta=a$: $\mu_1=23$, $\sigma_1=6.5$ vs. $\mu_2=17.1$, $\sigma_2=4.9$;
2. $\Delta=b$: $\mu_1=23$, $\sigma_1=6.5$ vs. $\mu_2=18.9$, $\sigma_2=5.4$; and
3. $\Delta=c$: $\mu_1=23$, $\sigma_1=6.5$ vs. $\mu_2=20.9$, $\sigma_2=6.0$.

After simulating 5000 datasets, the proportion of times that $Z_{jt\{1,2,3,4\}}$ exceeded the critical value was calculated for each interim analysis. Other than the choices of alternative hypotheses, the general outline described in section 6.1 was utilized for this simulation as before.

As shown in Figures 6.7 and 6.8 for the O’Brien-Fleming method (illustrated for $H_0$: $\Delta=a$), the power to reject the alternative hypothesis is highest at the lowest censoring and correlation levels. The cumulative rejection rate by the final analysis is 100% up until at least 40% censoring and for all correlation levels, regardless of whether the O’Brien-Fleming or Pocock method is selected. Rejection rates decrease monotonically as correlation and censoring levels increase, also regardless of which method is selected.
The tabled results in particular show that the Pocock method outperforms the O’Brien-Fleming in terms of total cumulative power to reject when both correlation levels and censoring levels increase (also illustrated by comparison of Figures 6.7 and 6.9 for $H_0$: $\Delta=a$. Since the O’Brien-Fleming begins with a much lower nominal alpha level, Figure 6.8 shows much higher rejection levels compared with Figure 6.6.

Table 6.7. *Simulation under Alternative Hypotheses for WLW Method with Correlation \( \rho = 0.0 \)*

<table>
<thead>
<tr>
<th>Proportion Censored</th>
<th>Analysis</th>
<th>$\Delta=0$</th>
<th>$\Delta=a$</th>
<th>$\Delta=b$</th>
<th>$\Delta=c$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>OF</td>
<td>Poc</td>
<td>OF</td>
<td>Poc</td>
</tr>
<tr>
<td>0%</td>
<td>1</td>
<td>0.48</td>
<td>1.06</td>
<td>96.8</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>2.79</td>
<td>4.08</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>Final</td>
<td>5.09</td>
<td>4.32</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>40%</td>
<td>1</td>
<td>0.70</td>
<td>1.24</td>
<td>91.9</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>2.92</td>
<td>3.94</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>Final</td>
<td>4.86</td>
<td>3.94</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>80%</td>
<td>1</td>
<td>1.10</td>
<td>1.62</td>
<td>60.9</td>
<td>99.8</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>3.54</td>
<td>4.50</td>
<td>94.6</td>
<td>99.9</td>
</tr>
<tr>
<td></td>
<td>Final</td>
<td>5.30</td>
<td>4.50</td>
<td>96.9</td>
<td>99.9</td>
</tr>
<tr>
<td>95%</td>
<td>1</td>
<td>7.55</td>
<td>7.63</td>
<td>7.98</td>
<td>75.9</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>9.34</td>
<td>9.82</td>
<td>12.4</td>
<td>75.9</td>
</tr>
<tr>
<td></td>
<td>Final</td>
<td>10.96</td>
<td>9.82</td>
<td>15.5</td>
<td>75.9</td>
</tr>
</tbody>
</table>

Note: OF = O’Brien-Fleming Shape, Poc = Pocock Shape

5000 simulations with 4 endpoints, multivariate lognormally distributed from 0 to 48 months with:
$\Delta=a$: $\mu_1=23.0$, $\mu_2=17.1$, $\sigma_1=6.5$, $\sigma_2=4.9$
$\Delta=b$: $\mu_1=23.0$, $\mu_2=18.9$, $\sigma_1=6.5$, $\sigma_2=5.4$
$\Delta=c$: $\mu_1=23.0$, $\mu_2=20.9$, $\sigma_1=6.5$, $\sigma_2=6.0$
Table 6.8. *Simulation under Alternative Hypotheses for WLW Method with Correlation rho = 0.25*

<table>
<thead>
<tr>
<th>Proportion Censored</th>
<th>Analysis</th>
<th>Cumulative Rejection rates under $H_a$: $\Delta=x$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$\Delta=0$</td>
</tr>
<tr>
<td>0%</td>
<td>1</td>
<td>0.66</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>2.78</td>
</tr>
<tr>
<td></td>
<td>Final</td>
<td>4.96</td>
</tr>
<tr>
<td>40%</td>
<td>1</td>
<td>1.04</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>3.02</td>
</tr>
<tr>
<td></td>
<td>Final</td>
<td>4.78</td>
</tr>
<tr>
<td>80%</td>
<td>1</td>
<td>1.52</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>4.00</td>
</tr>
<tr>
<td></td>
<td>Final</td>
<td>5.58</td>
</tr>
<tr>
<td>95%</td>
<td>1</td>
<td>7.41</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>9.64</td>
</tr>
<tr>
<td></td>
<td>Final</td>
<td>11.28</td>
</tr>
</tbody>
</table>

Note: OF = O’Brien-Fleming Shape, Poc = Pocock Shape
5000 simulations with 4 endpoints, multivariate lognormally distributed from 0 to 48 months with:

$\Delta=a$: $\mu_1=23.0$, $\mu_2=17.1$, $\sigma_1=6.5$, $\sigma_2=4.9$

$\Delta=b$: $\mu_1=23.0$, $\mu_2=18.9$, $\sigma_1=6.5$, $\sigma_2=5.4$

$\Delta=c$: $\mu_1=23.0$, $\mu_2=20.9$, $\sigma_1=6.5$, $\sigma_2=6.0$
Table 6.9. *Simulation under Alternative Hypotheses for WLW Method with Correlation \( \rho = 0.50 \)

<table>
<thead>
<tr>
<th>Proportion Censored</th>
<th>Analysis</th>
<th>( \Delta = 0 )</th>
<th>( \Delta = a )</th>
<th>( \Delta = b )</th>
<th>( \Delta = c )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>OF</td>
<td>Poc</td>
<td>OF</td>
<td>Poc</td>
</tr>
<tr>
<td>0%</td>
<td>1</td>
<td>1.14</td>
<td>1.90</td>
<td>70.1</td>
<td>99.7</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>3.46</td>
<td>4.52</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>Final</td>
<td>5.26</td>
<td>4.60</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>40%</td>
<td>1</td>
<td>1.62</td>
<td>2.28</td>
<td>61.2</td>
<td>99.5</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>3.74</td>
<td>4.60</td>
<td>99.3</td>
<td>99.9</td>
</tr>
<tr>
<td></td>
<td>Final</td>
<td>5.10</td>
<td>4.60</td>
<td>99.6</td>
<td>99.9</td>
</tr>
<tr>
<td>80%</td>
<td>1</td>
<td>2.32</td>
<td>2.88</td>
<td>27.2</td>
<td>95.7</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>4.50</td>
<td>5.48</td>
<td>62.3</td>
<td>96.6</td>
</tr>
<tr>
<td></td>
<td>Final</td>
<td>6.24</td>
<td>5.48</td>
<td>70.3</td>
<td>96.6</td>
</tr>
<tr>
<td>95%</td>
<td>1</td>
<td>8.39</td>
<td>8.47</td>
<td>9.6</td>
<td>56.2</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>10.22</td>
<td>10.84</td>
<td>14.0</td>
<td>56.4</td>
</tr>
<tr>
<td></td>
<td>Final</td>
<td>11.88</td>
<td>10.84</td>
<td>17.3</td>
<td>56.4</td>
</tr>
</tbody>
</table>

Note: OF = O’Brien-Fleming Shape, Poc = Pocock Shape

5000 simulations with 4 endpoints, multivariate lognormally distributed from 0 to 48 months with:

- \( \Delta = a: \mu_1 = 23.0, \mu_2 = 17.1, \sigma_1 = 6.5, \sigma_2 = 4.9 \)
- \( \Delta = b: \mu_1 = 23.0, \mu_2 = 18.9, \sigma_1 = 6.5, \sigma_2 = 5.4 \)
- \( \Delta = c: \mu_1 = 23.0, \mu_2 = 20.9, \sigma_1 = 6.5, \sigma_2 = 6.0 \)
Table 6.10. *Simulation under Alternative Hypotheses for WLW Method with Correlation rho = 0.75*

<table>
<thead>
<tr>
<th>Proportion Censored</th>
<th>Analysis</th>
<th>Δ=0 OF</th>
<th>Δ=a OF</th>
<th>Δ=b OF</th>
<th>Δ=c OF</th>
<th>Δ=0 Poc</th>
<th>Δ=a Poc</th>
<th>Δ=b Poc</th>
<th>Δ=c Poc</th>
</tr>
</thead>
<tbody>
<tr>
<td>0%</td>
<td>1</td>
<td>2.54</td>
<td>43.0</td>
<td>97.7</td>
<td>6.1</td>
<td>63.6</td>
<td>2.6</td>
<td>17.4</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>4.84</td>
<td>100</td>
<td>100</td>
<td>99.6</td>
<td>99.7</td>
<td>65.5</td>
<td>71.6</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Final</td>
<td>6.74</td>
<td>100</td>
<td>100</td>
<td>99.7</td>
<td>99.7</td>
<td>75.7</td>
<td>71.8</td>
<td></td>
</tr>
<tr>
<td>40%</td>
<td>1</td>
<td>3.36</td>
<td>37.0</td>
<td>96.7</td>
<td>5.8</td>
<td>60.8</td>
<td>3.5</td>
<td>17.3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>5.26</td>
<td>97.5</td>
<td>99.6</td>
<td>72.7</td>
<td>85.8</td>
<td>24.8</td>
<td>35.8</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Final</td>
<td>6.54</td>
<td>98.6</td>
<td>99.6</td>
<td>80.1</td>
<td>85.8</td>
<td>34.0</td>
<td>35.9</td>
<td></td>
</tr>
<tr>
<td>80%</td>
<td>1</td>
<td>4.59</td>
<td>15.9</td>
<td>85.1</td>
<td>5.3</td>
<td>45.9</td>
<td>4.9</td>
<td>13.7</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>6.52</td>
<td>47.8</td>
<td>87.8</td>
<td>23.8</td>
<td>52.7</td>
<td>11.6</td>
<td>19.6</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Final</td>
<td>8.22</td>
<td>57.9</td>
<td>87.8</td>
<td>31.7</td>
<td>52.7</td>
<td>16.9</td>
<td>19.6</td>
<td></td>
</tr>
<tr>
<td>95%</td>
<td>1</td>
<td>12.82</td>
<td>14.9</td>
<td>46.3</td>
<td>14.8</td>
<td>32.5</td>
<td>14.4</td>
<td>18.5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>14.20</td>
<td>18.3</td>
<td>46.8</td>
<td>16.7</td>
<td>33.1</td>
<td>15.5</td>
<td>19.4</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Final</td>
<td>15.36</td>
<td>21.1</td>
<td>46.8</td>
<td>18.8</td>
<td>33.1</td>
<td>17.0</td>
<td>19.4</td>
<td></td>
</tr>
</tbody>
</table>

Note: OF = O’Brien-Fleming Shape, Poc = Pocock Shape

5000 simulations with 4 endpoints, multivariate lognormally distributed from 0 to 48 months with:

Δ=a: μ₁=23.0, μ₂=17.1, σ₁=6.5, σ₂=4.9
Δ=b: μ₁=23.0, μ₂=18.9, σ₁=6.5, σ₂=5.4
Δ=c: μ₁=23.0, μ₂=20.9, σ₁=6.5, σ₂=6.0
Figure 6.6. Rejection Rate at Interim Analysis 1 under $H_a: \Delta = \alpha$ with O’Brien-Fleming Shape
Figure 6.7. Cumulative Rejection Rate at Final Analysis under $H_a: \Delta = \alpha$ with O’Brien-Fleming Shape
Figure 6.8. *Rejection Rate at Interim Analysis 1 under $H_a: \Delta = a$ with Pocock Shape*
Figure 6.9. *Cumulative Rejection Rate at Final Analysis under $H_a: \Delta = a$ with Pocock Shape*
6.5 Data Simulation Results for Fisher’s LSD Analogue

Next, we apply the Fisher’s LSD Analogue to these same simulated datasets, with results displayed below in Table 6.11 – 6.14 for the O’Brien-Fleming method for the usual four levels of correlation, under the null and alternative hypotheses. We skip the Pocock method for this procedure because the stage 1 results are already known; we instead use this section as a demonstration of the performance of the Fisher procedure, so the O’Brien-Fleming is sufficient for that purpose. As discussed in Chapter 4, our concern is that the probability of at least one incorrect assertion at stage 2 may exceed the nominal alpha-level for that interim analysis (Hayter, 1986). Per Westfall et al (1999), this procedure offers only weak control of the family-wise or experiment-wise error rate (EER). Recall that the EER is the probability that each least one of the single-endpoint hypotheses would lead to a type I error (i.e., falsely rejecting the single-endpoint hypothesis when the null is true for that hypothesis).

Recall that we can reach stage 2 (testing single endpoint hypotheses) only after the composite hypothesis is rejected at a given interim analysis, and indeed, the results below show that the nominal alpha level is never exceeded at the single endpoint level because of this “protection”. Recall also that we are conducting one-sided testing, so it should not be possible to make a false assertion based on the direction of the test statistic. As before, the results below show that the procedure is more powerful at lower censoring levels. For the lowest correlation level, at 0% censoring, the cumulative stage 2 rejection rate under the null is 2.3% (compared with a stage 1 observed rate of 5%), and at 95% censoring, the rejection rate is 8.2% (compared with a stage 1 observed rate of 11%). Similar results are seen in Table 6.7 for the O’Brien-Fleming.
Under the alternative hypothesis ($H_a: \Delta = a$, as in the previous section), the stage 1 results are of course identical to the results given in section 6.4; the probability of rejecting at least one endpoint (stage 2) is always slightly less than the probably of rejecting the null hypothesis. As illustrated in Figure 6.10 (for the first interim analysis), as usual the power decreases steadily as the censoring proportion increases, as well as a decrease that is seen as the correlation increases. Figure 6.11 (cumulative rejection at the final analysis) still shows a decrease in power with increase in censoring, but surprisingly and in contrast to the interim results shows lower power for lower correlation levels.

Table 6.11. *Simulation of Fisher’s for WLW Method with O’Brien-Fleming Shape and Correlation $\rho=0.0$*

<table>
<thead>
<tr>
<th>Censor Level</th>
<th>Analysis</th>
<th>Alternative Hypothesis: Cumulative Rejection Rates</th>
<th>Null hypothesis: Cumulative Rejection Rates</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Stage 1: Composite Hypothesis</td>
<td>Stage 2: Single Endpoint Rejection Rate</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\geq 1$</td>
<td>69.4</td>
</tr>
<tr>
<td>0%</td>
<td>1: $\alpha &lt; 0.0001$</td>
<td>96.8</td>
<td>all 4</td>
</tr>
<tr>
<td></td>
<td>2: $\alpha = 0.0235$</td>
<td>100</td>
<td>$\geq 1$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>all 4</td>
<td>3.7</td>
</tr>
<tr>
<td></td>
<td>Final: $\alpha = 0.0264$</td>
<td>100</td>
<td>$\geq 1$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>all 4</td>
<td>3.7</td>
</tr>
<tr>
<td>40%</td>
<td>1</td>
<td>91.9</td>
<td>$\geq 1$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>all 4</td>
<td>0.8</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>100</td>
<td>$\geq 1$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>all 4</td>
<td>4.0</td>
</tr>
<tr>
<td></td>
<td>Final</td>
<td>100</td>
<td>$\geq 1$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>all 4</td>
<td>4.0</td>
</tr>
<tr>
<td>80%</td>
<td>1</td>
<td>60.9</td>
<td>$\geq 1$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>all 4</td>
<td>1.2</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>94.6</td>
<td>$\geq 1$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>all 4</td>
<td>1.2</td>
</tr>
<tr>
<td></td>
<td>Final</td>
<td>96.9</td>
<td>$\geq 1$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>all 4</td>
<td>1.2</td>
</tr>
<tr>
<td>95%</td>
<td>1</td>
<td>8.0</td>
<td>$\geq 1$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>all 4</td>
<td>6.5</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>12.4</td>
<td>$\geq 1$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>all 4</td>
<td>6.5</td>
</tr>
<tr>
<td></td>
<td>Final</td>
<td>15.5</td>
<td>$\geq 1$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>all 4</td>
<td>6.5</td>
</tr>
</tbody>
</table>

5000 simulations with 4 endpoints, multivariate lognormally distributed from 0 to 48 months with:

$H_a$: $\mu_1=23.0, \mu_2=17.1, \sigma_1=6.5, \sigma_2=4.9$

$H_0$: $\mu_1=\mu_2=23.0, \sigma_1=\sigma_2=6.5$
Table 6.12. *Simulation of Fisher’s for WLW Method with O’Brien-Fleming Shape and Correlation rho=0.25*

<table>
<thead>
<tr>
<th>Censor Level</th>
<th>Analysis</th>
<th>Alternative Hypothesis: Cumulative Rejection Rates</th>
<th>Null Hypothesis: Cumulative Rejection Rates</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Stage 1: Composite Hypothesis</td>
<td>Stage 2: Single Endpoint Rejection Rate</td>
</tr>
<tr>
<td>0%</td>
<td>1: $\alpha &lt; 0.0001$</td>
<td>89.2</td>
<td>$\geq 1$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>all 4</td>
<td>0.6</td>
</tr>
<tr>
<td></td>
<td>2: $\alpha = 0.0235$</td>
<td>100</td>
<td>$\geq 1$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>all 4</td>
<td>11.3</td>
</tr>
<tr>
<td></td>
<td>Final: $\alpha = 0.0264$</td>
<td>100</td>
<td>$\geq 1$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>all 4</td>
<td>11.3</td>
</tr>
<tr>
<td>40%</td>
<td>1</td>
<td>81.8</td>
<td>$\geq 1$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>all 4</td>
<td>0.8</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>100</td>
<td>$\geq 1$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>all 4</td>
<td>8.8</td>
</tr>
<tr>
<td></td>
<td>Final</td>
<td>100</td>
<td>$\geq 1$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>all 4</td>
<td>8.8</td>
</tr>
<tr>
<td>80%</td>
<td>1</td>
<td>46.0</td>
<td>$\geq 1$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>all 4</td>
<td>1.3</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>80.8</td>
<td>$\geq 1$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>all 4</td>
<td>1.6</td>
</tr>
<tr>
<td></td>
<td>Final</td>
<td>86.6</td>
<td>$\geq 1$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>all 4</td>
<td>1.6</td>
</tr>
<tr>
<td>95%</td>
<td>1</td>
<td>7.8</td>
<td>$\geq 1$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>all 4</td>
<td>6.5</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>11.8</td>
<td>$\geq 1$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>all 4</td>
<td>6.5</td>
</tr>
<tr>
<td></td>
<td>Final</td>
<td>14.9</td>
<td>$\geq 1$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>all 4</td>
<td>6.5</td>
</tr>
</tbody>
</table>

5000 simulations with 4 endpoints, multivariate lognormally distributed from 0 to 48 months with:

$H_a: \mu_1=23.0, \mu_2=17.1, \sigma_1=6.5, \sigma_2=4.9$

$H_0: \mu_1=\mu_2=23.0, \sigma_1=\sigma_2=6.5$
Table 6.13. *Simulation of Fisher’s for WLW Method with O’Brien-Fleming Shape and Correlation rho=0.50*

<table>
<thead>
<tr>
<th>Censor Level</th>
<th>Analysis</th>
<th>Stage 1: Composite Hypothesis</th>
<th>Stage 2: Single Endpoint</th>
<th>Rejection Rate</th>
<th>Null hypothesis: Cumulative Rejection Rates</th>
</tr>
</thead>
<tbody>
<tr>
<td>0%</td>
<td>1: $\alpha &lt; 0.0001$</td>
<td>70.1</td>
<td>≥1</td>
<td>52.8</td>
<td>1.1</td>
</tr>
<tr>
<td></td>
<td>2: $\alpha = 0.0235$</td>
<td>100</td>
<td>≥1</td>
<td>82.7</td>
<td>3.5</td>
</tr>
<tr>
<td></td>
<td>Final: $\alpha = 0.0264$</td>
<td>100</td>
<td>≥1</td>
<td>82.7</td>
<td>5.3</td>
</tr>
<tr>
<td>40%</td>
<td>1</td>
<td>61.2</td>
<td>≥1</td>
<td>49.4</td>
<td>1.6</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>99.3</td>
<td>≥1</td>
<td>87.1</td>
<td>3.7</td>
</tr>
<tr>
<td></td>
<td>Final</td>
<td>99.6</td>
<td>≥1</td>
<td>87.3</td>
<td>5.1</td>
</tr>
<tr>
<td>80%</td>
<td>1</td>
<td>27.2</td>
<td>≥1</td>
<td>24.6</td>
<td>2.3</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>62.3</td>
<td>≥1</td>
<td>51.7</td>
<td>4.5</td>
</tr>
<tr>
<td></td>
<td>Final</td>
<td>70.3</td>
<td>≥1</td>
<td>54.7</td>
<td>6.2</td>
</tr>
<tr>
<td>95%</td>
<td>1</td>
<td>9.6</td>
<td>≥1</td>
<td>9.6</td>
<td>8.4</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>14.0</td>
<td>≥1</td>
<td>11.5</td>
<td>10.2</td>
</tr>
<tr>
<td></td>
<td>Final</td>
<td>17.3</td>
<td>≥1</td>
<td>12.1</td>
<td>11.9</td>
</tr>
</tbody>
</table>

5000 simulations with 4 endpoints, multivariate lognormally distributed from 0 to 48 months with:

$H_a: \mu_1=23.0, \mu_2=17.1, \sigma_1=6.5, \sigma_2=4.9$

$H_0: \mu_1=\mu_2=23.0, \sigma_1=\sigma_2=6.5$
Table 6.14. *Simulation of Fisher’s for WLW Method with O’Brien-Fleming Shape and Correlation rho=0.75*

<table>
<thead>
<tr>
<th>Censor Level</th>
<th>Analysis</th>
<th>Stage 1: Composite Hypothesis</th>
<th>Stage 2: Single Endpoint Rejection Rate</th>
<th>Stage 1: Composite Hypothesis</th>
<th>Stage 2: Single Endpoint Rejection Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>0%</td>
<td>1: $\alpha &lt; 0.0001$</td>
<td>43.0 ≥1 all 4 2.6</td>
<td>35.8</td>
<td>2.5 ≥1 all 4</td>
<td>2.5</td>
</tr>
<tr>
<td></td>
<td>2: $\alpha = 0.0235$</td>
<td>100 ≥1 all 4 59.6</td>
<td>92.8</td>
<td>4.8 ≥1 all 4</td>
<td>4.4</td>
</tr>
<tr>
<td></td>
<td>Final: $\alpha = 0.0264$</td>
<td>100 ≥1 all 4 59.6</td>
<td>92.8</td>
<td>6.7 ≥1 all 4</td>
<td>5.2</td>
</tr>
<tr>
<td>40%</td>
<td>1</td>
<td>37.0 ≥1 all 4 3.5</td>
<td>31.8</td>
<td>3.4 ≥1 all 4</td>
<td>3.4</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>97.5 ≥1 all 4 44.6</td>
<td>91.7</td>
<td>5.3 ≥1 all 4</td>
<td>4.9</td>
</tr>
<tr>
<td></td>
<td>Final</td>
<td>98.6 ≥1 all 4 44.6</td>
<td>92.2</td>
<td>6.5 ≥1 all 4</td>
<td>5.4</td>
</tr>
<tr>
<td>80%</td>
<td>1</td>
<td>15.9 ≥1 all 4 4.8</td>
<td>15.0</td>
<td>4.6 ≥1 all 4</td>
<td>4.6</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>47.8 ≥1 all 4 6.9</td>
<td>42.9</td>
<td>6.5 ≥1 all 4</td>
<td>6.0</td>
</tr>
<tr>
<td></td>
<td>Final</td>
<td>57.9 ≥1 all 4 6.9</td>
<td>47.2</td>
<td>8.2 ≥1 all 4</td>
<td>6.5</td>
</tr>
<tr>
<td>95%</td>
<td>1</td>
<td>14.9 ≥1 all 4 12.6</td>
<td>14.8</td>
<td>12.8 ≥1 all 4</td>
<td>12.6</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>18.3 ≥1 all 4 12.6</td>
<td>16.7</td>
<td>14.2 ≥1 all 4</td>
<td>13.4</td>
</tr>
<tr>
<td></td>
<td>Final</td>
<td>21.1 ≥1 all 4 12.6</td>
<td>17.5</td>
<td>15.4 ≥1 all 4</td>
<td>13.6</td>
</tr>
</tbody>
</table>

5000 simulations with 4 endpoints, multivariate lognormally distributed from 0 to 48 months with:

$H_a: \mu_1=23.0, \mu_2=17.1, \sigma_1=6.5, \sigma_2=4.9$

$H_0: \mu_1=\mu_2=23.0, \sigma_1=\sigma_2=6.5$
Figure 6.10. *Fisher’s LSD Rejection Rate of at Least One Endpoint at Interim Analysis 1 with O’Brien-Fleming Shape*
Figure 6.11.  *Fisher’s LSD Cumulative Rejection Rate of at Least One Endpoint at Final Analysis with O’Brien-Fleming Shape*
6.6 Data Simulation Results for BHM vs Hochberg Methods

Additional results using the same simulated datasets are displayed below in Table 6.15 – 6.18, comparing the Bonferroni-Holm method (BHM) with the Hochberg method, using the O’Brien-Fleming-type alpha spending function, under the alternative hypothesis $H_a$: $\mu_1 = 23.0$ vs. $\mu_2 = 17.1$, and $\sigma_1 = 6.5$ vs. $\sigma_2 = 4.9$, as above. As expected, the Hochberg method is consistently as powerful or occasionally slightly more powerful compared with BHM. This is true at all censoring levels and at all levels of correlation.

At interim analysis 1, power unexpectedly increases monotonically both as the censoring level increases and as the correlation level decreases, as illustrated in Figure 6.12. The reader will recall that much less information is available both because this is only the first interim analysis, and especially because we are utilizing test statistics on individual endpoints rather than a global test statistic across all four endpoints. In fact, we see that when the data are extremely sparse and heavily censored, it often occurs that only one event is observed in the active group and none in the control group; as a result, the estimated regression coefficient is extremely large (and hence the test statistic is also extremely large, with an associated p-value of less than 0.0001). As a result, even after adjustment using the BHM and Hochberg procedures (recall Chapter 3 sections 3.4.6 and 3.4.7), a p-value of 0.0001 would still be highly significant. Hence, it seems that the BHM and Hochberg procedures are unreliable under heavy censoring at interim analyses. In contrast, at the final analysis (Figure 6.13) where more events are observed, power decreases as the censoring level increases, with the exception of an upturn at the highest censoring level; as usual, power increases monotonically as the correlation decreases.
At the interim and final analyses, we note that the BHM and Hochberg methods are much more powerful than Fisher’s method across all censoring and correlation levels, i.e., the probability of rejecting at least one endpoint is greater under BHM/Hochberg than under Fisher’s method. The BHM and Hochberg methods are less powerful than the WLW closed testing method (section 6.4) in terms of early termination at lower censoring levels, but become more powerful as the censoring level increases.

Table 6.15. Simulation of BHM versus Hochberg for WLW Method with O’Brien-Fleming Shape and Correlation rho=0.0 under $H_a^*$: $\Delta=a$

<table>
<thead>
<tr>
<th>Censor Level</th>
<th>Analysis</th>
<th>Endpoint</th>
<th>BHM</th>
<th>Hochberg</th>
<th>Endpoint</th>
<th>BHM</th>
<th>Hochberg</th>
</tr>
</thead>
<tbody>
<tr>
<td>0%</td>
<td>1: $\alpha &lt; 0.0001$</td>
<td>reject ≥1</td>
<td>71.5</td>
<td>71.5</td>
<td>reject all 4</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td>2: $\alpha = 0.0235$</td>
<td>reject ≥1</td>
<td>100</td>
<td>100</td>
<td>reject all 4</td>
<td>29.0</td>
<td>29.0</td>
</tr>
<tr>
<td></td>
<td>Final: $\alpha = 0.0264$</td>
<td>reject ≥1</td>
<td>100</td>
<td>100</td>
<td>reject all 4</td>
<td>29.0</td>
<td>29.0</td>
</tr>
<tr>
<td>40%</td>
<td>1</td>
<td>reject ≥1</td>
<td>76.2</td>
<td>76.2</td>
<td>reject all 4</td>
<td>0.8</td>
<td>0.8</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>reject ≥1</td>
<td>99.7</td>
<td>99.7</td>
<td>reject all 4</td>
<td>11.2</td>
<td>11.7</td>
</tr>
<tr>
<td></td>
<td>Final</td>
<td>reject ≥1</td>
<td>99.7</td>
<td>99.8</td>
<td>reject all 4</td>
<td>11.2</td>
<td>11.7</td>
</tr>
<tr>
<td>80%</td>
<td>1</td>
<td>reject ≥1</td>
<td>80.0</td>
<td>80.0</td>
<td>reject all 4</td>
<td>1.2</td>
<td>1.2</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>reject ≥1</td>
<td>85.4</td>
<td>85.5</td>
<td>reject all 4</td>
<td>1.2</td>
<td>1.2</td>
</tr>
<tr>
<td></td>
<td>Final</td>
<td>reject ≥1</td>
<td>85.8</td>
<td>85.8</td>
<td>reject all 4</td>
<td>1.2</td>
<td>1.2</td>
</tr>
<tr>
<td>95%</td>
<td>1</td>
<td>reject ≥1</td>
<td>95.0</td>
<td>95.0</td>
<td>reject all 4</td>
<td>7.9</td>
<td>7.9</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>reject ≥1</td>
<td>95.0</td>
<td>95.0</td>
<td>reject all 4</td>
<td>7.9</td>
<td>7.9</td>
</tr>
<tr>
<td></td>
<td>Final</td>
<td>reject ≥1</td>
<td>95.0</td>
<td>95.0</td>
<td>reject all 4</td>
<td>7.9</td>
<td>7.9</td>
</tr>
</tbody>
</table>

5000 simulations with 4 endpoints, multivariate lognormally distributed from 0 to 48 months with:
* $H_a^*$: $\mu_1=23.0$, $\mu_2=17.1$, $\sigma_1=6.5$, $\sigma_2=4.9$
Table 6.16. Simulation of BHM versus Hochberg for WLW Method with O'Brien-Fleming Shape and Correlation rho=0.25 under $H_a^*$: $\Delta=a$

<table>
<thead>
<tr>
<th>Censor Level</th>
<th>Analysis</th>
<th>Endpoint BHM</th>
<th>Endpoint Hochberg</th>
<th>Censor Level</th>
<th>Analysis</th>
<th>Endpoint BHM</th>
<th>Endpoint Hochberg</th>
</tr>
</thead>
<tbody>
<tr>
<td>0%</td>
<td>1: $\alpha &lt; 0.0001$ reject $\geq 1$ 70.1</td>
<td>reject all 4 0.6</td>
<td>reject all 4 0.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2: $\alpha = 0.0235$ reject $\geq 1$ 100</td>
<td>reject all 4 30.5</td>
<td>reject all 4 30.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Final: $\alpha = 0.0264$ reject $\geq 1$ 100</td>
<td>reject all 4 30.5</td>
<td>reject all 4 30.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>40%</td>
<td>reject $\geq 1$ 73.9</td>
<td>reject all 4 0.8</td>
<td>reject all 4 0.8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>reject $\geq 1$ 98.9</td>
<td>reject all 4 13.8</td>
<td>reject all 4 14.4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Final</td>
<td>reject $\geq 1$ 99.0</td>
<td>reject all 4 13.8</td>
<td>reject all 4 14.4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>80%</td>
<td>reject $\geq 1$ 78.5</td>
<td>reject all 4 1.3</td>
<td>reject all 4 1.3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>reject $\geq 1$ 83.1</td>
<td>reject all 4 1.3</td>
<td>reject all 4 1.4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Final</td>
<td>reject $\geq 1$ 83.4</td>
<td>reject all 4 1.3</td>
<td>reject all 4 1.4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>95%</td>
<td>reject $\geq 1$ 93.3</td>
<td>reject all 4 7.7</td>
<td>reject all 4 7.7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>reject $\geq 1$ 93.3</td>
<td>reject all 4 7.7</td>
<td>reject all 4 7.7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Final</td>
<td>reject $\geq 1$ 93.3</td>
<td>reject all 4 7.7</td>
<td>reject all 4 7.7</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

5000 simulations with 4 endpoints, multivariate lognormally distributed from 0 to 48 months with:
$*H_a$: $\mu_1=23.0$, $\mu_2=17.1$, $\sigma_1=6.5$, $\sigma_2=4.9$
Table 6.17. *Simulation of BHM versus Hochberg for WLW Method with O'Brien-Fleming Shape and Correlation rho=0.50 under Hₐ*: Δ=α

<table>
<thead>
<tr>
<th>Censor Level</th>
<th>Analysis</th>
<th>Endpoint</th>
<th>BHM</th>
<th>Hochberg</th>
<th>Endpoint</th>
<th>BHM</th>
<th>Hochberg</th>
</tr>
</thead>
<tbody>
<tr>
<td>0%</td>
<td>1: ( \alpha &lt; 0.0001 ) reject ≥1</td>
<td>67.1</td>
<td>67.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2 : ( \alpha = 0.0235 ) reject ≥1</td>
<td>100</td>
<td>100</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Final: ( \alpha = 0.0264 ) reject ≥1</td>
<td>100</td>
<td>100</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>40%</td>
<td>1 reject ≥1</td>
<td>70.2</td>
<td>70.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2 reject ≥1</td>
<td>97.9</td>
<td>98.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Final reject ≥1</td>
<td>98.0</td>
<td>98.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>80%</td>
<td>1 reject ≥1</td>
<td>75.6</td>
<td>75.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2 reject ≥1</td>
<td>79.7</td>
<td>79.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Final reject ≥1</td>
<td>80.0</td>
<td>80.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>95%</td>
<td>1 reject ≥1</td>
<td>89.8</td>
<td>89.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2 reject ≥1</td>
<td>89.8</td>
<td>89.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Final reject ≥1</td>
<td>89.8</td>
<td>89.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

5000 simulations with 4 endpoints, multivariate lognormally distributed from 0 to 48 months with:
*\( Hₐ: \mu₁=23.0, \mu₂=17.1, \sigma₁=6.5, \sigma₂=4.9 \)
Table 6.18. *Simulation of BHM versus Hochberg for WLW Method with O'Brien-Fleming Shape and Correlation rho=0.75 under Ha*: $Δ=α$

<table>
<thead>
<tr>
<th>Censor Level</th>
<th>Analysis</th>
<th>Endpoint</th>
<th>BHM</th>
<th>Hochberg</th>
<th>Cumulative Rejection Rate of Rejecting at least One Endpoint</th>
<th>Cumulative Rejection Rate of Rejecting all Four Endpoints</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Endpoint</td>
<td>BHM</td>
</tr>
<tr>
<td>0%</td>
<td>1: $α &lt; 0.0001$</td>
<td>reject $≥1$</td>
<td>59.7</td>
<td>59.7</td>
<td></td>
<td>reject all 4</td>
</tr>
<tr>
<td></td>
<td>2: $α = 0.0235$</td>
<td>reject $≥1$</td>
<td>100</td>
<td>100</td>
<td></td>
<td>reject all 4</td>
</tr>
<tr>
<td></td>
<td>Final: $α = 0.0264$</td>
<td>reject $≥1$</td>
<td>100</td>
<td>100</td>
<td></td>
<td>reject all 4</td>
</tr>
<tr>
<td>40%</td>
<td>1</td>
<td>reject $≥1$</td>
<td>62.3</td>
<td>62.3</td>
<td></td>
<td>reject all 4</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>reject $≥1$</td>
<td>95.4</td>
<td>95.7</td>
<td></td>
<td>reject all 4</td>
</tr>
<tr>
<td></td>
<td>Final</td>
<td>reject $≥1$</td>
<td>95.4</td>
<td>95.9</td>
<td></td>
<td>reject all 4</td>
</tr>
<tr>
<td>80%</td>
<td>1</td>
<td>reject $≥1$</td>
<td>67.9</td>
<td>67.9</td>
<td></td>
<td>reject all 4</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>reject $≥1$</td>
<td>72.7</td>
<td>72.8</td>
<td></td>
<td>reject all 4</td>
</tr>
<tr>
<td></td>
<td>Final</td>
<td>reject $≥1$</td>
<td>72.9</td>
<td>73.1</td>
<td></td>
<td>reject all 4</td>
</tr>
<tr>
<td>95%</td>
<td>1</td>
<td>reject $≥1$</td>
<td>83.5</td>
<td>83.5</td>
<td></td>
<td>reject all 4</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>reject $≥1$</td>
<td>83.5</td>
<td>83.5</td>
<td></td>
<td>reject all 4</td>
</tr>
<tr>
<td></td>
<td>Final</td>
<td>reject $≥1$</td>
<td>83.5</td>
<td>83.5</td>
<td></td>
<td>reject all 4</td>
</tr>
</tbody>
</table>

5000 simulations with 4 endpoints, multivariate lognormally distributed from 0 to 48 months with:

$*H_a$: $μ_1=23.0$, $μ_2=17.1$, $σ_1=6.5$, $σ_2=4.9$
Figure 6.12. BHM vs Hochberg: Rejection Rate of at Least One Endpoint at Interim Analysis 1 with O’Brien-Fleming Shape
Figure 6.13: BHM vs Hochberg: Cumulative Rejection Rate of at Least One Endpoint at Final Analysis with O'Brien-Fleming Shape
CHAPTER VII
SUMMARY, CONCLUSIONS, AND DIRECTIONS FOR FUTURE RESEARCH

7.1 Summary of Methodology Developed in This Dissertation

In this dissertation, we have demonstrated that in studies involving distinct multiple time-to-event endpoints of interest, it may be beneficial or of interest to consider all events jointly rather than considering each event alone, particularly when counts within each treatment group may be too small to detect a treatment difference. In situations involving interim study monitoring, we would like to be able to stop the study early if sufficient evidence becomes available showing the superiority of one treatment group over the other. Additionally, we often want to be able to get more detailed information about the nature of the treatment differences than is available with a global hypothesis test.

We have considered several survival analysis procedures previously available only in the fixed study setting and extended their application to the group sequential setting, such as the marginal proportional hazards model of Wei et al (1989), accelerated failure time models, and several others. Each of these methods has been utilized to analyze multivariate survival data in the fixed study setting, but their application to the group sequential setting has not previously been studied in the literature. For this dissertation, we have examined these procedures using data examples involving only distinct multiple endpoints, although an examination of recurrrent multiple events could also be undertaken. For selected methods, we have also examined their properties using large data simulations.
By accounting for within-subject correlation in a marginal proportional hazards model, we can calculate robust estimates of variance and avoid underestimating the variance. Additionally, by considering multiple endpoints of interest rather than reducing the information to a single variable, we may in certain cases realize a reduction in variance, leading to greater power to detect treatment differences.

We have also considered several procedures available in the group sequential setting for non-survival type endpoints, and extended their application to multivariate survival data. In particular, we have taken Tang and Geller’s (1999) closed testing group sequential procedure, which was previously used for multiple continuous endpoints, and modified it for use with multiple survival endpoints. Tang and Geller’s method was itself an extension of work introduced by Marcus et al (1976) for a fixed sample study involving multiple treatment groups.

Finally, we have tested the application of several multiple comparison procedures to multiple survival endpoints in order to give us more flexibility in interpretation while adhering to the overall type I error rate. These multiple comparison procedures included the Bonferroni-Holm step-down method (1979), the Hochberg step-up method (1988), and a procedure analogous to Fisher’s protected least significant differences procedure (1935). This latter procedure was adapted at the suggestion of discussion in Follmann et al (1994).

7.2 Discussion of Results

In Chapter 5, we compared several multivariate group sequential survival procedures using a dataset involving distinct, multiple events (the PBC study of UDCA versus
Placebo). In this chapter we saw that certain survival methods cannot be applied to distinct multiple events (the Andersen-Gill and the Prentice-Williams-Petersen). We also saw that the choice of multiple comparison procedure had a bigger effect than the choice of survival procedure on whether or not we were able to stop the study early (as shown in summary tables 5.8 and 5.9). Specifically, procedures that involve an omnibus test (the closed testing method and the Fisher’s analogue) offered a better chance of stopping early compared with procedures that immediately allow testing of individual endpoints (the BHM and the Hochberg). However, once a study has been stopped, the frailty and AFT procedures proved to be significant more often than the WLW in testing individual endpoints.

In Chapter 6, we compared some of these methods using data simulations under a variety of null and alternative hypotheses, using different censoring and correlation levels. We found that in most cases, the procedures under study performed best under lower levels of censoring as well as under lower levels of correlation. Under the null hypothesis, we found that the type I error rate was maintained even under very heavy censoring, with significant breakdown occurring only above the 90% level. This was true across the range of correlation levels tested, although at correlation levels of 0.75 the error rate was slightly elevated across the range of censoring tested (Figures 6.2, 6.3). A similar pattern was seen when non-constant correlation levels were tested (Figure 6.4).

In comparing rejection rates in simulations under the alternative hypothesis, we found that power decreased monotonically as the censoring level was increased and as the correlation level increased (Figures 6.6 – 6.9). This finding held for the interim and final analysis, as well as for both the O’Brien-Fleming and Pocock shapes. In examining the
results of the Fisher-type procedure (Figures 6.10-6.11, we believe that the type I error rate has been adequately controlled, although it is not clear whether the EER has been controlled. We have avoided the problem of directional error, however, by conducting one-sided testing.

Additional simulations were done to compare the Bonferroni-Holm (BHM) and Hochberg stepwise procedures. As expected, it was found that the Hochberg was consistently as powerful or more powerful than the BHM. Additionally, these two procedures were found to be more powerful than the Fisher-type procedure at both the interim and final analyses. However, as explained in Chapter 6, it was found at the interim analysis that rejection rates increased rather than decreased as the censoring level increased (Figure 6.12). These two procedures therefore do not appear to be suitable for use in interim analyses when censoring is expected to be heavy or event rates are particularly low.

7.3 Conclusions and Directions for Further Research

In this paper we have considered only a limited number of multivariate survival methods, as noted above. Additional work is needed in order to determine which multivariate survival fixed-sample testing procedures will perform best in the group sequential setting. We have seen that there are different types of multivariate survival data, such as ordered and unordered outcomes, and distinct and recurrent outcomes. In this dissertation we have considered only distinct multiple outcomes. A variety of procedures have been developed for working with these different types of multivariate outcomes, and so more work is needed to see how they perform under different group sequential settings. Of
particular interest would be the different types of frailty models, of which we have considered only the gamma distribution as implemented using a penalized Cox regression model. Hougaard (2000) includes extensive discussion of many types of frailty models, including several varieties of instantaneous and short-term frailty models, competing risk models, marginal and copula models, and nonparametric varieties. At present the statistical software is lacking, as frailty models cannot be analyzed in SAS using built-in procedures. While it is possible to model a few types of frailty models in Splus, as demonstrated in Chapter 5, in this dissertation all simulations were done in SAS because of the greater flexibility available for manipulating data, which meant that frailty models could not be included.

In this dissertation, we have applied only two different error spending functions, the Pocock type and the O’Brien-Fleming type, to our data examples and simulations. The reader should recall from our discussion in Chapter 3 that many other variants are possible to preserve the type I error rate while allowing flexibility in spacing and the rate at which alpha is spent. It may be of interest to simulate data under different distributions than the few attempted in Chapter 6. For example, we have not specifically compared distributions where events are expected to occur relatively early in the study versus midway or later in the study. Different spending functions could be applied in this instance to compare performance.

Finally, we have also considered only a few very common multiple comparison procedures, including the closed testing method, the analogue to Fisher’s LSD, one stepup, and one stepdown procedure. The literature on multiple comparison procedures is rich, and much research is ongoing, with many new methods and refinements of
existing methods having been developed in recent years. For example, Westfall et al (1999) discuss eight different sequentially rejective procedures and devote a whole chapter to stepwise and closed testing procedures. Many of these procedures could be considered and possibly combined with the multivariate group sequential survival procedures developed in this dissertation to yield greater power, flexibility, or control of error rates.
A.1 Numerical Quadrature for Pocock’s Method

As mentioned in chapter 3, Pocock’s method (1977) keeps a constant alpha-level at each interim analysis, to maintain an overall type I error rate of 0.05. In a case where four interim looks are planned, the nominal alpha level at each stage would be $\alpha' = 0.0182$ (calculated using numerical quadrature as indicated in Armitage et al, 1969), implying a Z-critical value of 2.361, i.e.:

\begin{equation}
\alpha' = 2 \left( 1 - \Phi(z) \right) = 2 \left( 1 - \Phi(2.361) \right) = 0.0182
\end{equation}

The alpha levels are calculated using the joint distribution of $Z_1, ..., Z_K$, where $K$ denotes the number of interim looks, as usual. Recall that $Z_1, ..., Z_K$ are Markov, as discussed in chapters three and four. We can then say that:

\begin{equation}
Z_1 \sim N(\theta \sqrt{I_1}, 1)
\end{equation}

\begin{equation}
Z_k \sqrt{I_k} - Z_{k-1} \sqrt{I_{k-1}} \sim N(\theta(I_k - I_{k-1}), (I_k - I_{k-1})) \text{ for any subsequent } k \leq K.
\end{equation}

In practice the calculation would be done automatically using any of a variety of statistical packages, rather than by writing code to perform numerical integration. In Armitage et al (1969), numerical quadrature was applied using Simpson’s rule and Newton-Cotes formula to calculate the joint distribution using weighted sums in place of the integrals. In the general case, the integral is approximated as:
where the density to be computed in the sequential procedure is:

\[
(1.5) \quad f_n(s_n) = \begin{cases} 
\int_{y_{n-1}}^{y_n} f_{n-1}(u) \frac{1}{\sqrt{2\pi}} \exp\{-1/2(s_n - u)^2\} du, & -y_n \leq s_n \leq y_n \\
0, & \text{otherwise}.
\end{cases}
\]

which is then evaluated by Armitage et al (1969) at points on a grid of width h using a piecewise approach: Simpson’s three-point formula over a portion of the range and a second-order Newton-Cotes formula over the remainder of the range. Newton-Cotes formulas were used over only a portion of the range, presumably because the calculations are cumbersome and results are inaccurate over larger intervals.

SAS code is provided below in Appendix C.1 which will calculate (1.5) using numerical integration for Pocock’s method, maintaining a constant alpha-level at each interim analysis. Sample input and output are also provided for an example with four interim analyses, showing the critical alpha level is 0.0182, as stated earlier in this section.
Appendix B

Data Sources

B.1 UDCA Treatment in Patients with Primary Biliary Cirrhosis

Data Source:

Originally reported in Lindor et al (1994) based on a study conducted at Mayo. This dataset can be downloaded from the Mayo Biostatistics Web site:

http://mayoresearch.mayo.edu/mayo/research/biostat/upload/therneau_upload/urso.dat

Other datasets may be found by navigating from the main Mayo biostatistics department page: http://mayoresearch.mayo.edu/mayo/research/biostat/

Data Structure:

Column 1: Patient Identifier
Column 2: Study Drug (1=UDCA, 0=Placebo)
Column 3: Study Entry Date
Column 4: Date of Last Followup
Column 5: Histologic Stage Indicator
Column 6: bili
Column 7: risk
Column 8: death
Column 9: trans
Column 10: withdraw
Column 11: histprog
Column 12: varices
Column 13: ascites
Column 14: ence
Column 15: bili_2
Column 16: worse

data udca;
    infile 'c:\my documents\group sequential\udca dataset.txt';
    input id drug entry lastfu hi_stage bili risk;
    death trans withdraw histprog varices ascites ence bili_2 worse;
    informat entry lastfu death trans withdraw histprog varices ascites
        ence bili_2 worse date9.;
    format entry lastfu death trans withdraw histprog varices ascites
        ence bili_2 worse date9.;
run;

proc transpose data=udca out=udca2(rename=(col1=eventdt));
    by id drug entry lastfu hi_stage bili;
    var death trans withdraw histprog varices ascites ence bili_2 worse;
run;
B.2 CF Dataset: rhDNase for the Treatment of Cystic Fibrosis

Data Source:

Based on a study conducted by Genetech in 1992, and reported in Therneau and Hamilton (1997). Note that this is an example of recurrent, not distinct multiple event data. This dataset was used for testing some of the methods described in this dissertation, but results were not reported since our focus was on distinct event data. This dataset can be downloaded from the Mayo Biostatistics Web site:

http://mayoresearch.mayo.edu/mayo/research/biostat/upload/therneau_upload/dnase.dat

Data Structure:

Column 1: Patient Identifier
Column 2: Study Drug (0=placebo, 1=rhDNase)
Column 3: Study Entry Date
Column 4: Study End Date
Column 5: FEV (Forced Expiratory Volume), a measure of lung capacity
Column 6: Infection Start Date
Column 7: Infection Stop Date

Note that there may be multiple records per subject, corresponding to the number of infections that each patient had.
Appendix C

SAS and Splus Code and Output for Data Examples and Methodology

C.1 Numerical Integration for Pocock’s Method

Note: This SAS/IML code is extracted from code (Pocock portions only) on Barry Moser’s website at Duke University:
http://www.duke.edu/~bkmoser/FixedSample_Pock_OBF_critpts.pdf

Note that numerical integration is performed in this code using the “CALL SEQ” routine.

```
proc iml;
/* This program calculates critical points for the Pocock
one or two-sided procedure in the equal interim sample size case.

INPUTS designated by @@ are the responsibility of the user
alpha= the overall alpha level (default 0.05)
method = poc for Pocock
sided = 1 for a one sided test, =2 for a two sided test
n = the total number of interim analyses

OUTPUT
method = poc for Pocock
alpha = overall significance level of the test
sided = 1 for one-sided test, = 2 for two-sided test
n = the total number of interim analyses
critical = the critical points at each interim analysis stage
critical = the alpha level associated with each critical point
tau = the alpha spending function values */

alpha=.05;
method='poc';
sided=1;
n=4;

/* The following values are necessary for the operation of the program */
if sided=1 then alpha=2#alpha;
z=2;
critical=j(n,1,0); /* n = # rows, 1 = # columns, 0 = value to populate with */
minz=0;
maxz=10;
difflim=10##(-4);
m=j(2,n,1);
do k=1 to 30;
/* Calculations for Pocock procedure in this equal interim sample size
case the boundary points Zc(1)=...=Zc(n). The boundary point Yc(1),...,Yc(n)
in the SAS calculations must be scaled such that Yc(i')=Zc(i')*sqrt(i').
For a two sided test the alpha spending function is twice the sum from 1 to n
of lower limit probability */
if method='poc' then do i=1 to n;
m[1,i]=z*sqrt(i);
m[2,i]=z*sqrt(i);
```

136
end;
call seq(prob,m) eps = 1.e-8;
  tau=j(n,1,0); lastp=0; critalph=j(n,1,0);
  /* tau is the Pocock spending function */
  do j=1 to n;
    if j=1 then tau[j,1]=2*prob[1,j];
    if j>1 then tau[j,1]=lastp + 2*prob[1,j];
    lastp=tau[j,1];
  end;
alphdiff=alpha-lastp;
  if alphdiff>difflim then do;
    maxz=z;
    z=(minz + maxz)/2;
  end;
  if alphdiff<-difflim then do;
    minz=z;
    z=(minz + maxz)/2;
  end;
  if abs(alphdiff)<difflim then k=30;
end;
  /* Since the program already scales the Pocock by sqrt(i)
  for i=1,...,n the pocock critical point is simply z */
  if method='poc' then critical=j(n,1,z);
    if sided=1 then critalph[1]=1-probnorm(critical[1,1]);
    if sided=2 then critalph[1]=2#(1-probnorm(critical[1,1]));
  end;
print method alpha sided n critical critalph tau lastp;
run;
quit;

SAMPLE INPUT:
  alpha=.05;
  method='poc';
  sided=2;
  n=4;

SAMPLE OUTPUT:

<table>
<thead>
<tr>
<th>METHOD</th>
<th>ALPHA</th>
<th>SIDED</th>
<th>N</th>
<th>CRITICAL</th>
<th>CRITALPH</th>
<th>TAU</th>
<th>LASTP</th>
</tr>
</thead>
<tbody>
<tr>
<td>poc</td>
<td>0.05</td>
<td>2</td>
<td>4</td>
<td>2.3613281</td>
<td>0.0182096</td>
<td>0.0182096</td>
<td>0.0499962</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2.3613281</td>
<td>0.0182096</td>
<td>0.0315435</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2.3613281</td>
<td>0.0182096</td>
<td>0.0417516</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2.3613281</td>
<td>0.0182096</td>
<td>0.0499962</td>
<td></td>
</tr>
</tbody>
</table>

Note that critical alpha level of 0.0182 agrees with published values such as Jennison & Turnbull (2000), Geller and Pocock (1987), etc.
C.2 SAS Code for BHM and Hochberg

As discussed in Chapter 5 Section 5.4.1, SAS Proc Multtest (Westfall et al, 1999) can be used to convert raw p-values using the Hochberg, Bonferroni-Holm, and other multiple comparison methods. Manual calculation is illustrated in Chapter 3, sections 3.4.6 and 3.4.7. An example is given below using results of the PBC data:

*--- SAS code;

*---- enter raw p-values;

    data a;
    input Test$ Raw_P;
    datalines;
    test1 .2580
    test2 .1899
    test3 .1469
    test4 .0781
    ;

*--- use option "hoc" for Hochberg method, "holm" for Bonferroni-Holm;
*--- see SAS manual (Westfall et al, 1999) for more options;
    proc multtest pdata=a hoc holm;
    run;

*--- SAS output;

The Multtest Procedure

  p-Values

    | Stepdown | Raw | Bonferroni | Hochberg |
    |----------|-----|------------|----------|
    | Test     |     |            |          |
    | 1        | 0.2580 | 0.4407     | 0.2580   |
    | 2        | 0.1899 | 0.4407     | 0.2580   |
    | 3        | 0.1469 | 0.4407     | 0.2580   |
    | 4        | 0.0781 | 0.3124     | 0.2580   |
C.3 SAS Code and Output for PBC Data

C.3.1: Closed Testing Method

SAS Code:
This code is used in sections 5.2.1, 5.2.2. Note that the test statistics must be compared with the critical values for the appropriate method (Pocock or O’Brien-Fleming) at the appropriate interim analysis.

data temp;
  infile '"C:\windows\desktop\school\phd\stochastic curtailment\urso.dat"';
  input @1 id @6 rx @8 entry @16 lastfu @25 stage @27 bili @32 risk
    @37 death @45 trans @53 withdraw @61 histprog @69 varices @77 ascites @85 ence @93
    bili_2 @98 worse;
  informat entry lastfu death trans withdraw histprog varices ascites ence bili_2 worse   date7.;
  format entry lastfu death trans withdraw histprog varices ascites
    ence bili_2 worse   date7.;
  run;

data urso2; *time to all events;
  set temp;
  keep id rx stage bili risk futime status etype;
  array dt death trans /*withdraw*/ histprog varices ascites ence bili_2 worse;
  do i=1 to 8;
    etype =i;
    if (dt[i] =.) then do;
      futime = lastfu - entry;
      status =0;
      end;
    else do;
      futime = dt[i] - entry;
      status =1;
      end;
    output;
  end;
  lbili=log(bili);

data urso2; set urso2;
  if rx=0 then trt=1; else trt=2;
  trt1=trt*{etype=1};
  trt2=trt*{etype=2};
  trt3=trt*{etype=3};
  trt4=trt*{etype=4};
  trt5=trt*{etype=5};
  trt6=trt*{etype=6};
  trt7=trt*{etype=7};
  trt8=trt*{etype=8};
\[ bili = \log(bili); \]

if futime > 365*1 or futime = . then do;
    time_int1 = 365*1; status_i = 0;
end;
else do;
    time_int1 = futime; status_i = status;
end;

if futime > 365*3 or futime = . then do;
    time_int2 = 365*3; status_i2 = 0;
end;
else do;
    time_int2 = futime; status_i2 = status;
end;
run;

data urso2_int; set urso2;
if futime > 365 or futime = . then do;
    futime = 365; status = 0;
end;
run;

*--- global test at interim 1;
proc phreg data=urso2_int covsandwich(aggregate) covs;
model futime*status(0) = trt lbili stage;
   id id; strata etype;
   where etype in (1 2 3 4);
run;

*--- interim analysis 2;
data urso2_int2; set urso2;
if futime > 365*3 or futime = . then do;
    futime = 365*3; status = 0;
end;
run;

*--- global test at interim 2;
proc phreg data=urso2_int2 covsandwich(aggregate) covs;
model futime*status(0) = trt lbili stage;
   id id; strata etype;
   where etype in (1 2 3 4);
run;

*--- since we reject the global null, we can check subhypotheses (but not all);
%macro sub(list);
   proc phreg data=urso2_int2 covsandwich(aggregate) covs;
   model futime*status(0) = trt lbili stage;
   id id; strata etype;
   where etype in (&list);
run;
%mend;
%sub(1 2 3);
%sub(1 2 4);
%sub(1 3 4);
%sub(2 3 4);

%sub(1 4);
%sub(1 2);
%sub(1 3);
%sub(2 4);
%sub(2 3);
%sub(3 4);
Selected SAS Output

For stage 1, interim analysis one, as in section 5.2.1, 5.2.2:

Testing Global Null Hypothesis: BETA=0

<table>
<thead>
<tr>
<th>Test</th>
<th>Chi-Square</th>
<th>DF</th>
<th>Pr &gt; ChiSq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Likelihood Ratio</td>
<td>8.4071</td>
<td>3</td>
<td>0.0383</td>
</tr>
<tr>
<td>Score (Model-Based)</td>
<td>7.3351</td>
<td>3</td>
<td>0.0620</td>
</tr>
<tr>
<td>Score (Sandwich)</td>
<td>4.8698</td>
<td>3</td>
<td>0.1816</td>
</tr>
<tr>
<td>Wald (Model-Based)</td>
<td>4.2650</td>
<td>3</td>
<td>0.2342</td>
</tr>
</tbody>
</table>

The SAS System

The PHREG Procedure

Testing Global Null Hypothesis: BETA=0

<table>
<thead>
<tr>
<th>Test</th>
<th>Chi-Square</th>
<th>DF</th>
<th>Pr &gt; ChiSq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wald (Sandwich)</td>
<td>1328.9558</td>
<td>3</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>

Analysis of Maximum Likelihood Estimates

<table>
<thead>
<tr>
<th>Variable</th>
<th>Parameter Estimate</th>
<th>Standard Error</th>
<th>StdErr Ratio</th>
<th>Chi-Square</th>
<th>Pr &gt; ChiSq</th>
</tr>
</thead>
<tbody>
<tr>
<td>trt</td>
<td>-1.40090</td>
<td>1.14893</td>
<td>1.026</td>
<td>1.4867</td>
<td>0.2227</td>
</tr>
<tr>
<td>lbili</td>
<td>0.84663</td>
<td>0.60174</td>
<td>1.208</td>
<td>1.9795</td>
<td>0.1594</td>
</tr>
<tr>
<td>stage</td>
<td>16.00750</td>
<td>0.74469</td>
<td>0.000</td>
<td>462.0558</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>

Analysis of Maximum Likelihood Estimates

<table>
<thead>
<tr>
<th>Hazard Variable</th>
<th>Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>trt</td>
<td>0.246</td>
</tr>
<tr>
<td>lbili</td>
<td>2.332</td>
</tr>
<tr>
<td>stage</td>
<td>8952970</td>
</tr>
</tbody>
</table>
C.3.2 Hochberg and Bonferroni-Holm Methods

SAS Code:

This code is used in sections 5.3.1, 5.3.2, 5.4.1, and 5.4.2. Note that this produces raw p-values which must be converted either manually or using the code provided in Appendix C.2.

data temp;
    infile 'C:\windows\desktop\school\phd\stochastic curtailment\urso.dat';
    input @1 id @6 rx @8 entry @16 lastfu @25 stage @27 bili @32 risk
        @37 death @45 trans @53 withdraw @61 histprog @69 varices @77 ascites @85 ence @93
        bili_2 @101 worse;
    informat entry lastfu death trans withdraw histprog varices ascites
        ence bili_2 worse date7.;
    format entry lastfu death trans withdraw histprog varices ascites
        ence bili_2 worse date7.;
run;

data urso2; *time to all events;
    set temp;
    keep id rx stage bili risk futime status etype;
    array dt death trans /*withdraw*/ histprog varices ascites ence bili_2 worse;
    do i=1 to 8;
        etype = i;
        if (dt[i] = .) then do;
            futime = lastfu - entry;
            status = 0;
        end;
        else do;
            futime = dt[i] - entry;
            status = 1;
        end;
        output;
    end;
    lbili=log(bili);

data urso2; set urso2;
    if rx=0 then trt=1; else trt=2;
    trt1=trt*(etype=1);
    trt2=trt*(etype=2);
    trt3=trt*(etype=3);
    trt4=trt*(etype=4);
    trt5=trt*(etype=5);
    trt6=trt*(etype=6);
    trt7=trt*(etype=7);
    trt8=trt*(etype=8);
    lbili=log(bili);
    if futime>365*1 or futime=. then do; time_int1=365*1; status_i=0; end;
else do; time_int1=futime; status_i=status; end;

if futime>365*3 or futime=. then do; time_int2=365*3; status_i2=0; end;
else do; time_int2=futime; status_i2=status; end;
run;

data urso2_int; set urso2;
if futime>365 or futime=. then do;
futime=365; status=0; end;
run;
proc means data=urso2_int;
class trt; var futime; run;
proc freq data=urso2_int; table trt*status; run;
proc phreg data=urso2_int;
model futime*status(0)=trt lbili stage;
where etype=1;
run;
proc phreg data=urso2_int;
model futime*status(0)=trt lbili stage;
where etype=2;
run;
proc phreg data=urso2_int;
model futime*status(0)=trt lbili stage;
where etype=3;
run;
proc phreg data=urso2_int;
model futime*status(0)=trt lbili stage;
where etype=4;
run;

*--- interim analysis 2;
data urso2_int2; set urso2;
if futime>365*3 or futime=. then do;
futime=365*3; status=0; end;
run;
proc phreg data=urso2_int2;
model futime*status(0)=trt lbili stage;
where etype=1;
run;
proc phreg data=urso2_int2;
model futime*status(0)=trt lbili stage;
where etype=2;
run;
proc phreg data=urso2_int2;
model futime*status(0)=trt lbili stage;
where etype=3;
run;
proc phreg data=urso2_int2;
model futime*status(0)=trt lbili stage;
where etype=4;
run;
Sample SAS Output:

For Event 1, Interim Look 1:

The PHREG Procedure

Model Information

Data Set: WORK.URSO2_INT
Dependent Variable: time_int1
Censoring Variable: status_i
Censoring Value(s): 0
Ties Handling: BRESLOW

Summary of the Number of Event and Censored Values

<table>
<thead>
<tr>
<th>Total</th>
<th>Event</th>
<th>Censored</th>
<th>Percent Censored</th>
</tr>
</thead>
<tbody>
<tr>
<td>170</td>
<td>3</td>
<td>167</td>
<td>98.24</td>
</tr>
</tbody>
</table>

Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

With Covariates | Without Covariates |
-----------------|---------------------|
-2 LOG L         | 30.548              | 23.641            |
AIC              | 30.548              | 29.641            |
SBC              | 30.548              | 26.937            |

Testing Global Null Hypothesis: BETA=0

<table>
<thead>
<tr>
<th>Test</th>
<th>Chi-Square</th>
<th>DF</th>
<th>Pr &gt; ChiSq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Likelihood Ratio</td>
<td>6.9075</td>
<td>3</td>
<td>0.0749</td>
</tr>
<tr>
<td>Score</td>
<td>4.9005</td>
<td>3</td>
<td>0.1792</td>
</tr>
<tr>
<td>Wald</td>
<td>0.5200</td>
<td>3</td>
<td>0.9145</td>
</tr>
</tbody>
</table>

Analysis of Maximum Likelihood Estimates

<table>
<thead>
<tr>
<th>Variable</th>
<th>DF</th>
<th>Parameter Estimate</th>
<th>Standard Error</th>
<th>Chi-Square</th>
<th>Pr &gt; ChiSq</th>
<th>Hazard Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>trt</td>
<td>1</td>
<td>-18.16711</td>
<td>5042</td>
<td>0.0000</td>
<td>0.9971</td>
<td>0.000</td>
</tr>
<tr>
<td>lbili</td>
<td>1</td>
<td>0.51239</td>
<td>0.71058</td>
<td>0.5200</td>
<td>0.4709</td>
<td>1.669</td>
</tr>
<tr>
<td>stage</td>
<td>1</td>
<td>17.12798</td>
<td>5670</td>
<td>0.0000</td>
<td>0.9976</td>
<td>27452954</td>
</tr>
</tbody>
</table>

144
C.3.3 Fisher’s LSD Method for WLW Method

SAS Code:

This code is used in sections 5.5.1, 5.5.2.

```sas
data temp;
  infile 'C:\...\urso.dat';
  input @1 id @6 rx @8 entry @16 lastfu @25 stage @27 bili @32 risk
    @37 death @45 trans @53 withdraw @61 histprog @69 varices @77 ascites @85 ence @93
    bili_2 @101 worse;
  informat entry lastfu death trans withdraw histprog varices ascites
    ence bili_2 worse date7.;
  format entry lastfu death trans withdraw histprog varices ascites
    ence bili_2 worse date7.;
run;

data urso2; *time to all events;
  set temp;
  keep id rx stage bili risk futime status etype;
  array dt death trans /*withdraw*/ histprog varices ascites ence bili_2 worse;
  do i=1 to 8;
    etype = i;
    if (dt[i] =.) then do;
      futime = lastfu - entry;
      status = 0;
    end;
    else do;
      futime = dt[i] - entry;
      status = 1;
    end;
    output;
  end;
  lbili=log(bili);

data urso2; set urso2;
  if rx=0 then trt=1; else trt=2;
  trt1=trt*(etype=1);
  trt2=trt*(etype=2);
  trt3=trt*(etype=3);
  trt4=trt*(etype=4);
  trt5=trt*(etype=5);
  trt6=trt*(etype=6);
  trt7=trt*(etype=7);
  trt8=trt*(etype=8);
  lbili=log(bili);
  if futime>365*1 or futime=. then do; time_int1=365*1; status_i=0; end;
  else do; time_int1=futime; status_i=status; end;
  if futime>365*3 or futime=. then do; time_int2=365*3; status_i2=0; end;
  else do; time_int2=futime; status_i2=status; end;
run;
```

145
data urso2_int; set urso2;
if futime>365 or futime=. then do;
  futime=365; status=0; end;
run;
proc means data=urso2_int;
class trt; var futime; run;
proc freq data=urso2_int; table trt*status; run;
*--- FIRST THE STAGE 1, OVERALL TEST AT INTERIM 1;
proc phreg data=urso2_int covsandwich(aggregate) covs;
model futime*status(0)=trt lbili stage;
id id; strata etype;
where etype in (1 2 3 4);
run;
*--- INDIVIDUAL ENDPOINTS ARE EXAMINED ONLY IF STAGE 1 IS SIGNIFICANT AT THE DESIRED
ALPHA-LEVEL FOR INTERIM 1;
proc phreg data=urso2_int;
model futime*status(0)=trt lbili stage;
where etype=1;
run;
proc phreg data=urso2_int;
model futime*status(0)=trt lbili stage;
where etype=2;
run;
proc phreg data=urso2_int;
model futime*status(0)=trt lbili stage;
where etype=3;
run;
proc phreg data=urso2_int;
model futime*status(0)=trt lbili stage;
where etype=4;
run;

*--- interim analysis 2;
data urso2_int2; set urso2;
if futime>365*3 or futime=. then do;
  futime=365*3; status=0; end;
run;
*--- FIRST THE STAGE 1, OVERALL TEST AT INTERIM 1;
proc phreg data=urso2_int2 covsandwich(aggregate) covs;
model futime*status(0)=trt lbili stage;
id id; strata etype;
where etype in (1 2 3 4);
run;
*--- INDIVIDUAL ENDPOINTS ARE EXAMINED ONLY IF STAGE 1 IS SIGNIFICANT AT THE DESIRED
ALPHA-LEVEL FOR INTERIM 1;
proc phreg data=urso2_int2;
model futime*status(0)=trt lbili stage;
where etype=1;
run;
proc phreg data=urso2_int2;
model futime*status(0)=trt lbili stage;
where etype=2;
run;
proc phreg data=urso2_int2;
model futime*status(0)=trt lbili stage;
where etype=3;
run;
proc phreg data=urso2_int2;
model futime*status(0)=trt lbili stage;
where etype=3;
run;
proc phreg data=urso2_int2;
model futime*status(0)=trt lbili stage;
where etype=4;
run;
*--- repeat the above if needed for final analysis, with no truncating of timepoints:

**Selected SAS Output:**

For stage 1, interim analysis two, as in section 5.5.1, 5.5.2:

**Testing Global Null Hypothesis: BETA=0**

<table>
<thead>
<tr>
<th>Test</th>
<th>Chi-Square</th>
<th>DF</th>
<th>Pr &gt; ChiSq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Likelihood Ratio</td>
<td>19.6304</td>
<td>3</td>
<td>0.0002</td>
</tr>
<tr>
<td>Score (Model-Based)</td>
<td>19.2939</td>
<td>3</td>
<td>0.0002</td>
</tr>
<tr>
<td>Score (Sandwich)</td>
<td>14.5693</td>
<td>3</td>
<td>0.0022</td>
</tr>
<tr>
<td>Wald (Model-Based)</td>
<td>18.3539</td>
<td>3</td>
<td>0.0004</td>
</tr>
</tbody>
</table>

The SAS System

The PHREG Procedure

**Testing Global Null Hypothesis: BETA=0**

<table>
<thead>
<tr>
<th>Test</th>
<th>Chi-Square</th>
<th>DF</th>
<th>Pr &gt; ChiSq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wald (Sandwich)</td>
<td>17.0320</td>
<td>3</td>
<td>0.0007</td>
</tr>
</tbody>
</table>

**Analysis of Maximum Likelihood Estimates**

<table>
<thead>
<tr>
<th>Variable</th>
<th>DF</th>
<th>Parameter Estimate</th>
<th>Standard Error</th>
<th>StdErr Ratio</th>
<th>Chi-Square</th>
<th>Pr &gt; ChiSq</th>
</tr>
</thead>
<tbody>
<tr>
<td>trt</td>
<td>1</td>
<td>-0.82053</td>
<td>0.32704</td>
<td>1.137</td>
<td>6.2951</td>
<td>0.0121</td>
</tr>
<tr>
<td>lbili</td>
<td>1</td>
<td>0.49355</td>
<td>0.19457</td>
<td>1.137</td>
<td>6.4346</td>
<td>0.0112</td>
</tr>
<tr>
<td>stage</td>
<td>1</td>
<td>0.31440</td>
<td>0.40174</td>
<td>1.120</td>
<td>0.6125</td>
<td>0.4339</td>
</tr>
</tbody>
</table>

**Analysis of Maximum Likelihood Estimates**

<table>
<thead>
<tr>
<th>Hazard Variable</th>
<th>Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>trt</td>
<td>0.440</td>
</tr>
<tr>
<td>lbili</td>
<td>1.638</td>
</tr>
<tr>
<td>stage</td>
<td>1.369</td>
</tr>
</tbody>
</table>
C.3.4 AFT Method

--- Global hypothesis for CTM, or stage 1 for Fisher’s;

\[
\text{survreg(formula = Surv(time, status) ~ trt + strata(etype) + cluster(id) + lbili + stage, data = mydata)}
\]

Deviance Residuals:

<table>
<thead>
<tr>
<th></th>
<th>Min</th>
<th>1Q</th>
<th>Median</th>
<th>3Q</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Value</td>
<td>-3.3</td>
<td>0.000327</td>
<td>0.000766</td>
<td>0.108</td>
<td>0.531</td>
</tr>
</tbody>
</table>

Coefficients:

<table>
<thead>
<tr>
<th>Value</th>
<th>Std. Error</th>
<th>z value</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>20.451317</td>
<td>109.65638</td>
<td>0.1865</td>
</tr>
<tr>
<td>trt</td>
<td>1.049135</td>
<td>0.93143</td>
<td>1.1264</td>
</tr>
<tr>
<td>strata(etype)1</td>
<td>0.149510</td>
<td>0.34296</td>
<td>0.4359</td>
</tr>
<tr>
<td>strata(etype)2</td>
<td>3.074121</td>
<td>49.49366</td>
<td>0.0621</td>
</tr>
<tr>
<td>strata(etype)3</td>
<td>1.557431</td>
<td>41.12688</td>
<td>0.0379</td>
</tr>
<tr>
<td>cluster(id)</td>
<td>-0.000711</td>
<td>0.00717</td>
<td>-0.0991</td>
</tr>
<tr>
<td>lbili</td>
<td>-0.636721</td>
<td>0.44072</td>
<td>-1.4447</td>
</tr>
<tr>
<td>stage</td>
<td>-8.017102</td>
<td>95.32305</td>
<td>-0.0841</td>
</tr>
</tbody>
</table>

Extreme value distribution: Dispersion (scale) = 0.7374761

Degrees of Freedom: 680 Total; 671 Residual

-2*Log-Likelihood: 50.8

Number of Newton-Raphson Iterations: 18

Correlation of Coefficients:

<table>
<thead>
<tr>
<th></th>
<th>(Intercept)</th>
<th>trt</th>
<th>strata(etype)1</th>
<th>strata(etype)2</th>
<th>strata(etype)3</th>
</tr>
</thead>
<tbody>
<tr>
<td>trt</td>
<td>0.015</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>strata(etype)1</td>
<td>0.010</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>strata(etype)2</td>
<td>0.339</td>
<td>0.011</td>
<td>0.004</td>
<td></td>
<td></td>
</tr>
<tr>
<td>strata(etype)3</td>
<td>0.240</td>
<td>0.006</td>
<td>0.002</td>
<td>-0.300</td>
<td></td>
</tr>
<tr>
<td>cluster(id)</td>
<td>-0.009</td>
<td>-0.042</td>
<td>-0.011</td>
<td>-0.001</td>
<td>-0.001</td>
</tr>
<tr>
<td>lbili</td>
<td>-0.028</td>
<td>-0.293</td>
<td>-0.106</td>
<td>-0.013</td>
<td>-0.008</td>
</tr>
<tr>
<td>stage</td>
<td>-0.870</td>
<td>-0.013</td>
<td>-0.006</td>
<td>-0.001</td>
<td>0.000</td>
</tr>
</tbody>
</table>

0.014

*** INTERIM ANALYSIS 2;

> mydata<-read.table("clipboard",header=T)
> fit1<-survreg(Surv(time,status)~trt+strata(etype)+cluster(id)+lbili+stage, data=mydata)
> summary(fit1)

Call:
\[
\text{survreg(formula = Surv(time, status) ~ trt + strata(etype) + cluster(id) + lbili + stage, data = mydata)}
\]

Deviance Residuals:

<table>
<thead>
<tr>
<th></th>
<th>Min</th>
<th>1Q</th>
<th>Median</th>
<th>3Q</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Value</td>
<td>-3.97</td>
<td>0.22</td>
<td>0.329</td>
<td>0.454</td>
<td>0.874</td>
</tr>
</tbody>
</table>

Coefficients:

<table>
<thead>
<tr>
<th>Value</th>
<th>Std. Error</th>
<th>z value</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>20.451317</td>
<td>109.65638</td>
<td>0.1865</td>
</tr>
<tr>
<td>trt</td>
<td>1.049135</td>
<td>0.93143</td>
<td>1.1264</td>
</tr>
<tr>
<td>strata(etype)1</td>
<td>0.149510</td>
<td>0.34296</td>
<td>0.4359</td>
</tr>
<tr>
<td>strata(etype)2</td>
<td>3.074121</td>
<td>49.49366</td>
<td>0.0621</td>
</tr>
<tr>
<td>strata(etype)3</td>
<td>1.557431</td>
<td>41.12688</td>
<td>0.0379</td>
</tr>
<tr>
<td>cluster(id)</td>
<td>-0.000711</td>
<td>0.00717</td>
<td>-0.0991</td>
</tr>
<tr>
<td>lbili</td>
<td>-0.636721</td>
<td>0.44072</td>
<td>-1.4447</td>
</tr>
<tr>
<td>stage</td>
<td>-8.017102</td>
<td>95.32305</td>
<td>-0.0841</td>
</tr>
</tbody>
</table>
\begin{verbatim}
(Intercept)  7.638031    0.28775  26.544 3.04e-155
trt  0.400534    0.14217   2.817 4.84e-003
strata(etype)1  0.052753    0.11267   0.468 6.40e-001
strata(etype)2 -0.114578    0.05520  -2.076 3.79e-002
strata(etype)3 -0.076303    0.03502  -2.179 2.93e-002
cluster(id)  0.000751    0.00136   0.552 5.81e-001
lbili -0.240483    0.08437  -2.850 4.37e-003
stage -0.135967    0.17108  -0.795 4.27e-001

Extreme value distribution: Dispersion (scale) = 0.474292
Degrees of Freedom: 680 Total; 671 Residual
-2*Log-Likelihood: 378

Number of Newton-Raphson iterations: 5

Correlation of Coefficients:

<table>
<thead>
<tr>
<th>(Intercept)</th>
<th>trt</th>
<th>strata(etype)1</th>
<th>strata(etype)2</th>
<th>strata(etype)3</th>
</tr>
</thead>
<tbody>
<tr>
<td>trt</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>strata(etype)1</td>
<td>0.040</td>
<td>0.018</td>
<td></td>
<td></td>
</tr>
<tr>
<td>strata(etype)2</td>
<td>0.141</td>
<td>0.070</td>
<td>0.089</td>
<td></td>
</tr>
<tr>
<td>strata(etype)3</td>
<td>0.145</td>
<td>0.082</td>
<td>0.074</td>
<td>0.234</td>
</tr>
<tr>
<td>cluster(id)</td>
<td>0.358</td>
<td>0.008</td>
<td>0.006</td>
<td>0.028</td>
</tr>
<tr>
<td>lbili</td>
<td>0.052</td>
<td>-0.267</td>
<td>-0.017</td>
<td>0.077</td>
</tr>
<tr>
<td>stage</td>
<td>-0.603</td>
<td>0.105</td>
<td>0.006</td>
<td>0.024</td>
</tr>
</tbody>
</table>

-0.263
\end{verbatim}
C.3.5 Frailty Method

*--- Selected code for sections 5.2.6 and others utilizing stage 1 global test;

```r
> coxph(Surv(time,status)~trt+strata(etype)+frailty(id,d="gamma")+lbili+stage,data=udca1)
Call:
coxph(formula = Surv(time, status) ~ trt + strata(etype) + frailty(id,
    d = "gamma") + lbili + stage, data = udca1)

    coef   se(coef) se2      Chisq DF  p
trt    -1.401    1.120    1.120  1.560  1   0.210
frailty(id, d = "gamma")                          0.00 0  1   0.960
lbili   0.847    0.498    0.498  2.890  1   0.089
stage   15.007 1693.574 1693.574  0.000  1  0.990

Iterations: 6 outer, 27 Newton-Raphson
Variance of random effect= 5e-07   I-likelihood = -21.3
Degrees of freedom for terms= 1 0 1 1
Likelihood ratio test=8.41 on 3 df, p=0.0383  n= 680
> coxph(Surv(time,status)~trt+strata(etype)+lbili+stage,data=udca1)
Call:
coxph(formula = Surv(time, status) ~ trt + strata(etype) + lbili +
    stage, data = udca1)

    coef   se(coef) se2        z     p
trt    -1.401 0.246e+01 1.12e+00 -1.25055 0.210
lbili   0.847 2.33e+00 4.98e-01  1.69986 0.089
stage   19.007 1.80e+08 1.25e+04  0.00152 1.000

Likelihood ratio test=8.41 on 3 df, p=0.0383  n= 680
Warning message:
Loglik converged before variable 3; beta may be infinite.  in: fitter(X, Y, strats,
offset, init, control, weights = weights,
> coxph(Surv(time,status)~trt+strata(etype)+frailty(id,d="gamma")+lbili+stage,data=udca2)
Call:
coxph(formula = Surv(time, status) ~ trt + strata(etype) + frailty(id,
    d = "gamma") + lbili + stage, data = udca2)

    coef   se(coef) se2      Chisq DF  p
trt    -0.850 0.320    0.288  7.080  1   0.0078
frailty(id, d = "gamma")                      31.02 27.4  0.2900
lbili   0.509 0.193    0.172  6.910  1   0.0086
stage   0.355 0.393    0.361  0.820  1   0.3700

Iterations: 6 outer, 19 Newton-Raphson
Variance of random effect= 0.614   I-likelihood = -260.6
Degrees of freedom for terms= 0.8 27.4  0.8  0.8
Likelihood ratio test=70.5 on 29.9 df, p=3.96e-05  n= 680
> z123<-udca2[!(udca2$etype==4),]
> z124<-udca2[!(udca2$etype==3),]
> z134<-udca2[!(udca2$etype==2),]
> z234<-udca2[!(udca2$etype==1),]
> coxph(Surv(time,status)~trt+strata(etype)+lbili+stage,data=z123)
```

150
Call:
coxph(formula = Surv(time, status) ~ trt + strata(etype) + lbili +
       stage, data = z123)

    coef  exp(coef)  se(coef)      z      p
trt   -0.8071     0.446    0.361 -2.238 0.0250
lbili  0.6137     1.847    0.210  2.925 0.0034
stage  0.0538     1.055    0.429  0.125 0.9000

Likelihood ratio test=13.7  on 3 df, p=0.00339  n= 510

> coxph(Surv(time, status)~trt+strata(etype)+lbili+stage,data=z124)
Call:
coxph(formula = Surv(time, status) ~ trt + strata(etype) + lbili +
       stage, data = z124)

    coef  exp(coef)  se(coef)      z      p
trt   -0.894     0.409    0.351 -2.55 0.0110
lbili  0.534     1.706    0.204  2.62 0.0088
stage  0.778     2.177    0.502  1.55 0.1200

Likelihood ratio test=20.3  on 3 df, p=0.000147  n= 510

> coxph(Surv(time, status)~trt+strata(etype)+lbili+stage,data=z134)
Call:
coxph(formula = Surv(time, status) ~ trt + strata(etype) + lbili +
       stage, data = z134)

    coef  exp(coef)  se(coef)      z      p
trt   -0.813     0.443    0.310 -2.623 0.0087
lbili  0.402     1.495    0.189  2.127 0.0330
stage  0.305     1.357    0.383  0.801 0.4200

Likelihood ratio test=13.9  on 3 df, p=0.00297  n= 510

> coxph(Surv(time, status)~trt+strata(etype)+lbili+stage,data=z234)
Call:
coxph(formula = Surv(time, status) ~ trt + strata(etype) + lbili +
       stage, data = z234)

    coef  exp(coef)  se(coef)      z     p
trt   -0.777      0.46    0.316 -2.46 0.014
lbili  0.454      1.57    0.190  2.38 0.017
stage  0.218      1.24    0.383  0.57 0.570

Likelihood ratio test=13.3  on 3 df, p=0.00406  n= 510

> z12<-z123[!(z123$etype==3),]
> z23<-z123[!(z123$etype==1),]
> z13<-z123[!(z123$etype==2),]
> z14<-z124[!(z124$etype==2),]
> z24<-z124[!(z124$etype==1),]
> coxph(Surv(time, status)~trt+strata(etype)+lbili+stage,data=z12)
Call:
coxph(formula = Surv(time, status) ~ trt + strata(etype) + lbili +
       stage, data = z12)
```
coef exp(coef) se(coef)      z      p
trt   -0.943     0.389    0.516 -1.829 0.0670
lbili  0.797     2.218    0.285  2.792 0.0052
stage  0.763     2.145    0.777  0.982 0.3300

Likelihood ratio test=14.5  on 3 df, p=0.00227  n= 340
> coxph(Surv(time,status)~trt+strata(etype)+lbili+stage,data=z13)
Call:
  coxph(formula = Surv(time, status) ~ trt + strata(etype) + lbili +
         stage, data = z13)

  coef exp(coef) se(coef)     z     p
trt   -0.7897     0.454    0.408 -1.9344 0.053
lbili  0.4993     1.648    0.245  2.0359 0.042
stage -0.0286     0.972    0.472 -0.0605 0.950

Likelihood ratio test=7.74  on 3 df, p=0.0518  n= 340
> coxph(Surv(time,status)~trt+strata(etype)+lbili+stage,data=z14)
Call:
  coxph(formula = Surv(time, status) ~ trt + strata(etype) + lbili +
         stage, data = z14)

  coef exp(coef) se(coef)     z     p
trt   -0.904     0.405    0.395 -2.29 0.022
lbili  0.405     1.500    0.235  1.72 0.085
stage  0.858     2.359    0.563  1.52 0.130

Likelihood ratio test=14.2  on 3 df, p=0.00261  n= 340
> coxph(Surv(time,status)~trt+strata(etype)+lbili+stage,data=z23)
Call:
  coxph(formula = Surv(time, status) ~ trt + strata(etype) + lbili +
         stage, data = z23)

  coef exp(coef) se(coef)     z     p
trt   -0.730     0.482    0.423 -1.725 0.084
lbili  0.592     1.808    0.248  2.391 0.017
stage -0.189     0.828    0.478 -0.395 0.690

Likelihood ratio test=7.79  on 3 df, p=0.0505  n= 340
> coxph(Surv(time,status)~trt+strata(etype)+lbili+stage,data=z24)
Call:
  coxph(formula = Surv(time, status) ~ trt + strata(etype) + lbili +
         stage, data = z24)

  coef exp(coef) se(coef)     z     p
trt   -0.848     0.428    0.405 -2.09 0.036
lbili  0.487     1.628    0.239  2.04 0.041
stage  0.727     2.070    0.567  1.28 0.200

Likelihood ratio test=13.1  on 3 df, p=0.00436  n= 340
> z4<-z24[!(z24$etype==2),]
> coxph(Surv(time,status)~trt+strata(etype)+lbili+stage,data=z4)
```
Call:
coxph(formula = Surv(time, status) ~ trt + strata(etype) + lbili +
    stage, data = z4)

<table>
<thead>
<tr>
<th>coef</th>
<th>exp(coef)</th>
<th>se(coef)</th>
<th>z</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>trt</td>
<td>-0.841</td>
<td>0.431</td>
<td>0.477</td>
<td>-1.76</td>
</tr>
<tr>
<td>lbili</td>
<td>0.266</td>
<td>1.305</td>
<td>0.296</td>
<td>0.90</td>
</tr>
<tr>
<td>stage</td>
<td>0.838</td>
<td>2.311</td>
<td>0.660</td>
<td>1.27</td>
</tr>
</tbody>
</table>

Likelihood ratio test=7.55 on 3 df, p=0.0562 n= 170
Appendix D
SAS Code for Data Simulations

D.1 SAS/IML Multivariate Normal Macro

/************************************************************************
* MVN macro: Generating multivariate normal data
*
* DISCLAIMER:
* THIS INFORMATION IS PROVIDED BY SAS INSTITUTE INC. AS A SERVICE TO ITS USERS. IT IS PROVIDED "AS IS". THERE ARE NO WARRANTIES, EXPRESSED OR IMPLIED, AS TO MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE REGARDING THE ACCURACY OF THE MATERIALS OR CODE CONTAINED HEREIN.
*
* PURPOSE:
* The %MVN macro generates multivariate normal data using the Cholesky root of the variance-covariance matrix. Bivariate normal data can be generated using the DATA step code that follows the macro.
*
* REQUIRES:
* The %MVN macro requires Version 6.06 or later of SAS/IML software. The DATA step code for generating bivariate normal data requires only Version 6.06 Base SAS software.
*
* USAGE:
* The macro input/output parameters are:
*
* VARCOV= SAS data set that contains the variance-covariance (and only the variance covariance) matrix. The macro expects m variables and m observations in the data set, where m is the number of variables to generate.
*
* MEANS= SAS data set that contains the mean vector. The macro expects a single variable with m observations containing the m means for the variables generated.
*
* N= Number of observations to generate.
*
* SEED= Starting seed value for the random number generator. Default value is 0, which will use the system clock to generate a seed.
*
* SAMPLE= SAS data set name for the resulting multivariate normal data. The variable names will be COL1-COLm.
*
* LIMITATIONS:
* No error checking is done. The macro assumes that dataset names entered are valid, and exist in the case of the VARCOV= and MEANS= options.
EXAMPLE:

This example generates 20000 observations from a 3 variable multivariate normal distribution with specified mean vector and covariance matrix.

* Store the variance-covariance matrix in a data set ;
data varcov;
  input m1-m3;
cards:
  4 1.8  4
  1.8   9 3.6
  4 3.6 16;
;
* Store the mean vector in a data set ;
data means;
  input m1;
cards:
  10
  20
  30;
;
%mvn(varcov=varcov,
  means=means,
  n=20000,
  sample=test)

proc corr data=test noprob cov;
run;

************************************************************************/
%macro mvn(varcov=,       /* dataset for variance-covariance matrix */
  means=,        /* dataset for mean vector */
  n=,            /* sample size */
  seed=0,        /* seed for random number generator */
  sample=);      /* output dataset name */

/* Get initial seed value. If seed<=0, then generate seed from the system clock. */
data _null_;     
  if &seed le 0 then do;
    seed = int(time()); /* get clock time in integer seconds */
    put seed=;
    call symput('seed',seed); /* store seed as macro variable */
  end;
run;

/* Generate the multivariate normal data in SAS/IML */
proc iml worksize=100;
use &varcov;            /* read variance-covariance matrix */
read all into cov;
use &means;             /* read means */
read all into mu;
v=nrow(cov);            /* calculate number of variables */
n=&n;
seed = &seed;
l=t(root(cov));         /* calculate cholesky root of cov matrix */
z=normal(j(v,n,&seed)); /* generate nvars*samplesize normals */
x=l*z;                  /* premultiply by cholesky root */
x=repeat(mu,1,n)+x;    /* add in the means */
tx=t(x);
create &sample from tx; /* write out sample data to sas dataset */
append from tx;
quit;

%mend mvn;
D.2 SAS Code to Generate Simulations

*--- include SAS/IML multivariate normal macro;
%inc 'C:\Documents and Settings\vhasfcscherr\Desktop\misc\pgms\mvn.sas';

*--- location to store simulation results;
libname sim 'C:\...\simulation data';

*--- set up starting values for MVN data;

%macro generate(cens=, rho=, HA=0, MUa=, VARa=);
data varcov1 means1 means2 test1 test2 samples;
run;

%let censdec=%sysevalf(&cens / 100);
%let rhodec=%sysevalf(&rho / 100);
%put censdec = &censdec; %put rhodec = &rhodec;

* Store a default variance-covariance matrix;
data varcov1;
m1=1; m2=0.5; m3=0.5; m4=0.5; output;
m1=0.5; m2=1; m3=0.5; m4=0.5; output;
m1=0.5; m2=0.5; m3=1; m4=0.5; output;
m1=0.5; m2=0.5; m3=0.5; m4=1; output;

data varcov1o; set varcov1;
array ms m1-m4;
do over ms; if ms=0.5 then ms=&rhodec; ms=ms/12.5; end;
if _n_=1 then call symput('var',m1);
run;

* Store the mean vector in a data set (times on normal scale);
data means1;
m1=3.1; output; m1=3.1; output; m1=3.1; output; m1=3.1; output;
run;

* censoring times;
data means2(rename=(m2=m1)); set means1;
x = 1-&censdec; *--- 92% of the time event is censored, if censdec=0.92;
if x=0 then x=0.00000001;
if x=1 then x=0.999999;
var= 2*var;
m2 = -1* (probit(1-x) * sqrt(var) - m1);
keep m2;
run;

*--- simulation under the alternative hypothesis;

%if &ha=1 %then %do;
data means1_alt;
m1=&ha; output; m1=&ha; output; m1=&ha; output; m1=&ha; output;
run;

%if &Var1 ^= %then %do;
data varcov1a; set varcov1;
array ms m1-m4;
do over ms; if ms=0.5 then ms=&rhodec; ms=ms/12.5; end;
if _n_=1 then call symput('var',m1);
run;
%end;
%else %do;
data varcov1a; set varcov1o; run;
%end;

*--- generate first set of MVN survival times;
*--- we need N=170 pts, 5000 replications => 850,000 obs;
*--- first half (425000) on treatment 1, second half on treatment 2;
%mvn(varcov=varcov1o, means=means1, n=425000, seed=5000, sample=test1)
data test1; set test1;
rename col1=time1 col2=time2 col3=time3 col4=time4; run;

%mvn(varcov=varcov1a, means=means1_alt, n=425000, seed=6000, sample=test1_alt)
data test1; set test1(in=a)
test1_alt(rename=(col1=time1 col2=time2 col3=time3 col4=time4));
if a then trt=1;
else trt=2;
run;
data means2_alt(rename=(m2=m1)); set means1_alt;
x = 1 - &censdec; *---92% of the time event is censored, if censdec=0.92;
if x=0 then x=0.00000001;
if x=1 then x=0.999999;
var = 2*&var;
m2 = -1* (probit(1-x) * sqrt(var) - m1);
keep m2;
run;

*--- generate second set of MVN survival times, use to determine censoring;
%mvn(varcov=varcov1o, means=means2, n=425000, seed=1000, sample=test2)
data test2; set test2;
rename col1=c1 col2=c2 col3=c3 col4=c4;
run;

%mvn(varcov=varcov1a, means=means2_alt, n=425000, seed=2000, sample=test2)
data test2_alt; set test2_alt;
rename col1=c1 col2=c2 col3=c3 col4=c4;
run;
data test2; set test2 test2_alt;
run;
%end;

*--- simulation under the null hypothesis;
%else %do;

*--- generate first set of MVN survival times;
*--- we need N=170 pts, 5000 replications => 850,000 obs;
%mvn(varcov=varcov1o, means=means1, n=850000, seed=5000, sample=test1)
data test1; set test1;
rename col1=time1 col2=time2 col3=time3 col4=time4;
run;
*--- generate second set of MVN survival times, use to determine censoring;
%mvn(varcov=varcov1o, means=means2, n=850000, seed=1000, sample=test2)
data test2; set test2;
rename col1=c1 col2=c2 col3=c3 col4=c4;
run;
%end;

data samples; merge test1 test2;
array times time1-time4;
array cs c1-c4;
array events event1-event4;
*--- transform to lognormal:
array lntimes lntime1-lntime4;
do over times;
  if times<cs then events=1;
  else events=0;
  if &cens=0 then events=1;
  lntimes=exp(times);
*--- ensure truncated at 48 months (rarely needed);
  if lntimes>48 then lntimes=ranuni(5000)*48;
end;
patient=_n; 
keep patient time1-time4 event1-event4 lntime1-lntime4 c1-c4;
run;

*-- sampling under Ho or Ha;
data samples_rho&rho._cens&cens.;
set samples(keep=event1-event4 lntime1-lntime4 patient);
%if &HA=0 %then %do;
sample=ceil(patient/170);
  if mod(patient,2)=0 then trt=1;
  else trt=2;
%end;
%else %if &HA=1 %then %do;
sample=ceil(patient/85);
  if trt=2 then sample=sample-5000;
%end;
x=ranuni(50); *-- random covariate;
run;
%mend generate;

*--- now we have the data, so run the simulation;
%macro simulate(B=5000, cens=, rho=, type=pocock, HA=, MUa=, VARa=);
%generate(cens=&Cens, rho=&rho, HA=&HA, MUa=&MUa, VARa=&VARa);
%if &type=pocock %then %do;
  %let crit1=2.1; %let crit2=1.902; %let crit3=1.999;
%end;
%else %if &type=of %then %do;
  %let crit1=3.75; %let crit2=1.985; %let crit3=1.715;
%end;
data x1 x2 x3 x1a x2a x3a; run;
%let int1=0; %let int2=0; %let int3=0;
%let time1=%sysfunc(time(),time5.);
%put &time1;
data _null_; x=time(); format x time.; put x=; run;
%do loop=1 %to &b;
proc transpose data=samples_rho&rho._cens&cens out=one(rename=(col1=event));
WHERE SAMPLE=&loop;
by patient trt x; var event1-event4; run;
proc transpose data=samples_rho&rho._cens&cens out=two(rename=(col1=time));
WHERE SAMPLE=&loop;
by patient trt x; var lntime1-lntime4; run;
data temp; merge one two;
K=input(substr(_name_,7),best.);
%if &cens=0 %then %do;
  event=1;
%end;
*--- now construct times for interim 1 and interim 2;
if time>12 or time=. then do; time1=12; event1=0; end;
else do; time1=time; event1=event; end;
if time>36 or time=. then do; time2=36; event2=0; end;
else do; time2=time; event2=event; end;
run;
*--- run WLW cox model at interim analysis one;
%let k1=; %let k2=; %let k3=; %let k4=;
/*proc freq data=temp noprint;
table k / out=k;
where event1=1;
run;
data _null_; set k;
if k=1 then call symput('k1','1');
if k=2 then call symput('k2','2');
if k=3 then call symput('k3','3');
if k=4 then call symput('k4','4');
run;*/
ods output Phreg.ParameterEstimates=x1;
ods listing close;
proc phreg data=temp covsandwich(aggregate) ;
model time1*event1(0)=trt x;
id patient; strata K;
*where k in (&k1 &k2 &k3 &k4);
run;
ods output close;
ods listing;
%let k1=0; %let k2=0; %let k3=0; %let k4=0;
data x1a; set x1a x1a(in=a where=(variable='trt'));
if a then do;
z=estimate/stderr;
sample=&loop;
if z>&crit1 then call symput('count1',1);
else call symput('count1',0);
*--- run WLW cox model at interim analysis two ONLY IF fail to reject at interim analysis one;
/*proc freq data=temp noprint;
table k / out=k;
where event2=1;
run;
data _null_; set k;
if k=1 then call symput('k1','1');
if k=2 then call symput('k2','2');
if k=3 then call symput('k3','3');
if k=4 then call symput('k4','4');
run;*/
ods output Phreg.ParameterEstimates=x2;
ods listing close;
proc phreg data=temp(keep=time2 event2 trt x patient K )
covsandwich(aggregate) covs;
model time2*event2(0)=trt x;
  id patient; strata K;
  *where k in (&k1 &k2 &k3 &k4);
run;
ods output close;
ods listing;
%let k1=; %let k2=; %let k3=; %let k4=;
data x2a; set x2a x2(in=a where=(variable='trt'));
  if a then do;
    z=estimate/stderr;
    sample=&loop;
    if z>&crit2 then call symput('count2',1);
    else call symput('count2',0);
  end;
run;
%let int2=%sysevalf(&int2+&count2);
*--- run WLW cox model at final analysis ONLY IF fail to reject at interim analysis two;
ods output Phreg.ParameterEstimates=x3;
ods listing close;
proc phreg data=temp(keep=time event trt x patient K )
covsandwich(aggregate) covs;
model time*event(0)=trt x;
  id patient; strata K;
  *where k in (&k1 &k2 &k3 &k4);
run;
ods output close;
ods listing;
%let k1=; %let k2=; %let k3=; %let k4=;
data x3a; set x3a x3(in=a where=(variable='trt'));
  if a then do;
    z=estimate/stderr;
    sample=&loop;
    if z>&crit3 then call symput('count3',1);
    else call symput('count3',0);
end;
run;
%let int3=%sysevalf(&int3+&count3);
%end;
%end;
%let int1=%sysevalf(&int1+&count1);

%put i=&loop;
%end;
*--- end of the 5000 iterations. Now we count the overall rejections;

%put i=&loop int1=&int1 int2=&int2 int3=&int3;
%let time1=%sysfunc(time(),time5.);
%put time1=&time1 time2=&time2;
%let R1 = %sysevalf(&int1 / &b); %put R1 = &R1;
data counts2; set x2a nobs=n; call symput('denom',n); run;
%let R2 = %sysevalf(&int2 / &denom); %put R2 = &R2;
data counts3; set x3a nobs=n; call symput('denomb',n); run;
%let R3 = %sysevalf(&int3 / &denomb); %put R3 = &R3;
data sim.null_c&cens._rhoρ set x1a(in=a) x2a(in=b) x3a(in=c);
if a then interim=1;
else if b then interim=2;
else if c then interim=3;
run;
%mend;

*---- sample code to invoke the macro;

*-- simul under Ho;
ods noresults; options nonotes nomprint;
%simulate(b=5000, cens=0, rho=0, type=pocock);
%simulate(b=5000, cens=20, rho=0, type=pocock);
%simulate(b=5000, cens=40, rho=0, type=pocock);
%simulate(b=5000, cens=60, rho=0, type=pocock);
%simulate(b=5000, cens=80, rho=0, type=pocock);
%simulate(b=5000, cens=90, rho=0, type=pocock);
%simulate(b=5000, cens=95, rho=0, type=pocock);
%simulate(b=5000, cens=0, rho=25, type=pocock);
**** ... etc.;

ods noresults; options nonotes nomprint;
*-- simul under HA;
%simulate(b=5000, cens=0, rho=0, type=pocock, HA=1, MUa=2.8);
%simulate(b=5000, cens=20, rho=0, type=pocock, HA=1, MUa=2.8);
%simulate(b=5000, cens=40, rho=0, type=pocock, HA=1, MUa=2.8);
%simulate(b=5000, cens=60, rho=0, type=pocock, HA=1, MUa=2.8);
%simulate(b=5000, cens=80, rho=0, type=pocock, HA=1, MUa=2.8);
%simulate(b=5000, cens=90, rho=0, type=pocock, HA=1, MUa=2.8);
%simulate(b=5000, cens=95, rho=0, type=pocock, HA=1, MUa=2.8);
%simulate(b=5000, cens=0, rho=25, type=pocock, HA=1, MUa=2.8);
**** ... etc.;
D.3 SAS Code to Generate Simulations with Non-Constant Correlation Structure

*--- include SAS/IML multivariate normal macro;
%inc 'C:\Documents and Settings\vhasfcscherr\Desktop\misc\pgms\mvn.sas';

*--- location to store simulation results;
libname sim 'C:\...\simulation data';

*--- set up starting values for MVN data;

%macro generate(cens=, rhoa=, rhob=, rhoc=);
data varcov1 means1 means2 test1 test2 samples;
run;

%let censdec=%sysevalf(&cens / 100);
%let rhodeca=%sysevalf(&rhoa / 100);
%let rhodecb=%sysevalf(&rhob / 100);
%let rhodecc=%sysevalf(&rhoc / 100);
%put censdec = &censdec;

* Store the variance-covariance matrix in a data set;
data varcov1;
m1=1.000; m2=&rhodeca; m3=&rhodecb; output;
m1=&rhodeca; m2=1.000; m3=&rhodecc; output;
m1=&rhodecb; m2=&rhodecc; m3=1.000; output;
run;

* Store the mean vector in a data set (times on normal scale);
data means1;
m1=3.1; output; m1=3.1; output; m1=3.1; output;
run;

* censoring times;
data means2(rename=(m2=m1)); set means1;
x = 1-%censdec; *---92% of the time event is censored, if censdec=0.92;
if x=0 then x=0.00000001;
if x=1 then x=0.999999;
var = 2*var;
m2 = -1* (probit(1-x) * sqrt(var) - m1);
keep m2;
run;

*--- generate first set of MVN survival times;
*--- we need N=170 pts, 5000 replications => 850,000 obs;
%mvn(varcov=varcov1o, means=means1, n=850000, seed=5000, sample=test1)
data test1; set test1;
rename col1=time1 col2=time2 col3=time3 ;
run;
*--- generate second set of MVN survival times, use to determine censoring;*
%mvn(varcov=varcov1o, means=means2, n=850000, seed=1000, sample=test2)
data test2: set test2;
rename col1=c1 col2=c2 col3=c3;
run;

data samples: merge test1 test2;
array times time1-time3;
array cs c1-c3;
array events event1-event3;
*--- transform to lognormal;*
array lntimes lntime1-lntime3;
do over times:
  if times<cs then events=1;
  else events=0;
  if &cens=0 then events=1;
  lntimes=exp(times);
  if lntimes>48 then lntimes=ranuni(5000)*48;
end;
patient=_n_; keep patient time1-time3 event1-event3 lntime1-lntime3 c1-c3;
run;

*-- sampling under Ho;*
data s_rhoa&rhoa._rhob&rhob._rhoc&rhoc._c&cens;
set samples(keep=event1-event3 lntime1-lntime3 patient);
  sample=ceil(patient/170);
  if mod(patient,2)=0 then trt=1;
  else trt=2;
x=ranuni(50); *-- random covariate;*
run;
%mend generate;

*--- Note: remaining code follows macro simulate given in Appendix D.1;
*--- Macro is invoked for non-constant correlation as;
%simulate(b=5000, cens=0, rhoa=25, rhob=50, rhoc=75, type=pocock);
D.4 SAS Code to Summarize BHM and Hochberg Simulations

Note: code below is provided for Hochberg, but BHM summary is structured analogously.

```sas
%macro summarize(GS=P, cens=0, rho=0, NULL=Y, ha);
%clean;
%if &GS=P %then %do;
*--- critical values for Pocock;
%let crit1=2.1; %let crit2=1.902; %let crit3=1.999;
%let alpha1=0.01787; %let alpha2=0.02353; %let alpha3=0.00860;
%end;

%else %if &GS=OF %then %do;
*--- critical values for O'Brien-Fleming;
%let crit1=3.75; %let crit2=1.985; %let crit3=1.715;
%let alpha1=0.0001; %let alpha2=0.0235; %let alpha3=0.0264;
%end;

%if &NULL=Y %THEN %let indat=misc.null_c&cens._rhoρ
%else %let indat=misc.ha_c&cens._rho&rho._ha&ha;

proc transpose data=&indat out=samples prefix=p;
by interim sample; var probchisq; id k;
where k ne .;
run;

proc transpose data=&indat out=samples2 prefix=z;
by interim sample; var z; id k;
where k ne .;
run;

data samples; merge samples samples2 &indat(keep=interim sample z k where=(k=.));
by interim sample;
keep z interim p1-p4 z1-z4 sample;
if interim=. or sample=. then delete;
rename interim=analysis;
run;

data hoch; set samples;
if analysis=1 then do; alpha=&alpha1; crit=&crit1; end;
else if analysis=2 then do; alpha=&alpha2; crit=&crit2; end;
else if analysis=3 then do; alpha=&alpha3; crit=&crit3; end;

*--- rescale one-sided p-values to not reject;
array rawp p1-p4;
array rawz z1-z4;
do over rawp; if .<rawz<0 then rawp=0.9999; end;

format p1o p2o p3o p4o 6.4;
*--- calculate adjusted p-values;
p1o=min(p1,p2,p3,p4);
p4o=max(p1,p2,p3,p4);
med=median(p1,p2,p3,p4);
```

165
if \( z < \text{crit} \) then reject = 0; else if \( z > \text{crit} \) then reject = 1;

if \( p_1 > p_{10} \) and \( p_1 < \text{med} \) then \( p_2 = p_1 \);
else if \( p_1 = \text{med} \) and \( p_1 < p_{40} \) then \( p_3 = p_1 \);
if \( p_2 = \text{med} \) then do;
if \( p_2 > p_{10} \) and \( p_2 < \text{med} \) then \( p_2 = p_2 \);
if \( p_3 > p_{10} \) and \( p_3 < \text{med} \) then \( p_2 = p_3 \);
if \( p_4 > p_{10} \) and \( p_4 < \text{med} \) then \( p_2 = p_4 \);
end;
if \( p_3 = \text{med} \) then do;
if \( p_2 > p_{40} \) and \( p_3 = p_3 \) then \( p_3 = p_2 \);
if \( p_3 = \text{med} \) and \( p_3 < p_{40} \) and \( p_4 = p_4 \) then \( p_3 = p_3 \);
if \( p_4 > p_{40} \) and \( p_4 = p_4 \) and \( p_3 = p_3 \) then \( p_3 = p_4 \);
end;
p_{4\_adj} = p_{40} ;
p_{3\_adj} = \min(p_{4\_adj}, 2*p_{30});
p_{2\_adj} = \min(p_{3\_adj}, 3*p_{40});
p_{1\_adj} = \min(p_{2\_adj}, 4*p_{10});

if \( p_{1\_adj} < \alpha \) then reject 1 = 1; else if \( p_{1\_adj} > \) then reject 1 = 0;
if \( p_{2\_adj} < \alpha \) then reject 2 = 1; else if \( p_{2\_adj} > \) then reject 2 = 0;
if \( p_{3\_adj} < \alpha \) then reject 3 = 1; else if \( p_{3\_adj} > \) then reject 3 = 0;
if \( p_{4\_adj} < \alpha \) then reject 4 = 1; else if \( p_{4\_adj} > \) then reject 4 = 0;

if reject 1 = 1 or reject 2 = 1 or reject 3 = 1 or reject 4 = 1 then at least 1 = 1;
else at least 1 = 0;

if reject 1 = 1 and reject 2 = 1 and reject 3 = 1 and reject 4 = 1 then all four = 1;
else all four = 0;

*-- if inferior on all or superior on any, stop the "study" for sample i;
if \( \text{all four} = 1 \) and \( z < 0 \) or at least 1 = 1 then stop = 1;
run;
data stop1(rename=(stop=stop1 at least 1 = at least 11 all four = all four 1)); set hoch;
where analysis = 1;
if stop = 1;
keep sample stop at least 1 all four;
run;

data int2_hoch; merge hoch(where=(analysis=2)) stop1(in=a keep=sample);
by sample;
if not a;
* c stop = max(stop, stop1);
* cat least1 = max(at least 1, at least 11);
* call four = max(all four, all four 1);
run;

data stop2; set int2_hoch;
where analysis = 2;
if stop = 1;
keep sample;

data final_hoch; merge hoch(where=(analysis=3)) stop1(in=a) stop2(in=b);
by sample;
if not (a or b);
title "Interim 1: hoch, GS Method &GS, Censoring at &Cens, Rho at &rho, NULL = &NULL";
proc freq data=hoch;
  where analysis=1;
  table stop / missing out=stop; table atleast1 / out=atleast1; table allfour / out=allfour;
run;
title "Interim 2: hoch, GS Method &GS, Censoring at &Cens, Rho at &rho, NULL = &NULL";
proc freq data=int2_hoch noprint;
  where analysis=2;
  table atleast1 / out=a; table allfour / out=b; table stop / out=c;
run;
data c; set c stop;
  proc freq data=c;
  weight count;
  table stop; run;
data a; set a atleast1;
  proc freq data=a;
  weight count;
  table atleast1; run;
data b; set b allfour;
  proc freq data=b;
  weight count;
  table allfour; run;
title "Final: hoch, GS Method &GS, Censoring at &Cens, Rho at &rho, NULL = &NULL";
proc freq data=final_hoch noprint;
  where analysis=3;
  table atleast1 / out=aa; table allfour / out=bb; table stop / out=cc;
run;
data c; set c cc;
  proc freq data=c;
  weight count;
  table stop; run;
data a; set a aa;
  proc freq data=a;
  weight count;
  table atleast1; run;
data b; set b bb;
  proc freq data=b;
  weight count;
  table allfour; run;
%mend;
*--- sample code to invoke summary macro;
%summarize(GS=OF, cens=0, rho=0, NULL=N, ha=28);
%summarize(GS=OF, cens=40, rho=0, NULL=N, ha=28);
%summarize(GS=OF, cens=80, rho=0, NULL=N, ha=28);
%summarize(GS=OF, cens=95, rho=0, NULL=N, ha=28);
D.5 SAS Code to Summarize Fisher Simulations

%macro summarize(GS=P, cens=0, rho=0, NULL=Y, ha=);
%clean;
*--- critical values for Pocock:
%let crit1=2.1; %let crit2=1.902; %let crit3=1.999;
%let alpha1=0.01787; %let alpha2=0.02353; %let alpha3=0.00860;
%end;

*--- critical values for O'Brien-Fleming:
%let crit1=3.75; %let crit2=1.985; %let crit3=1.715;
%let alpha1=0.0001; %let alpha2=0.0235; %let alpha3=0.0264;
%end;

%if &GS=P %then %do;
* if &GS=P %then %do;
%let indat=misc.null_c&cens._rhoρ;
%end;

%else %if &GS=OF %then %do;
* if &GS=OF %then %do;
%let indat=misc.ha_c&cens._rho&rho._ha&ha;
%end;

proc transpose data=&indat out=samples prefix=p;
by interim sample; var probchisq; id k;
where k ne .;
run;

proc transpose data=&indat out=samples2 prefix=z;
by interim sample; var z; id k;
where k ne .;
run;

data samples; merge samples samples2 &indat(keep=interim sample z k where=(k=.));
by interim sample;
keep z interim p1-p4 z1-z4 sample;
if interim=. or sample=. then delete;
rename interim=analysis;
run;

data fisher; set samples;
if analysis=1 then do; alpha=&alpha1; crit=&crit1; end;
else if analysis=2 then do; alpha=&alpha2; crit=&crit2; end;
else if analysis=3 then do; alpha=&alpha3; crit=&crit3; end;
if .<z<crit then reject=0; else if z>crit then reject=1;
if reject then do;
if .<p1<alpha and z1>0 then reject1=1; else if p1 ne . then reject1=0;
if .<p2<alpha and z2>0 then reject2=1; else if p2 ne . then reject2=0;
if .<p3<alpha and z3>0 then reject3=1; else if p3 ne . then reject3=0;
if .<p4<alpha and z4>0 then reject4=1; else if p4 ne . then reject4=0;
end;
if reject1=1 or reject2=1 or reject3=1 or reject4=1 then atleast1=1;
else atleast1=0;
if reject1=1 and reject2=1 and reject3=1 and reject4=1 then allfour=1;
else allfour=0;
*-- if inferior on all or superior on any, stop the "study" for sample i;
stop=reject;

168
run;

data stop1(rename=(stop=stop1 atleast1=atleast11 allfour=allfour1)); set fisher;
where analysis=1;
if stop=1;
keep sample stop atleast1 allfour;
run;

data int2_fisher; merge fisher(where=(analysis=2)) stop1(in=a keep=sample);
by sample;
if not a;
 cstop=max(stop,stop1);
catleast1=max(atleast1,atleast11);
callfour=max(allfour,allfour1);
run;

data stop2; set int2_fisher;
where analysis=2;
if stop=1;
keep sample;

data final_fisher; merge fisher(where=(analysis=3)) stop1(in=a) stop2(in=b);
by sample;
if not (a or b);

title"Interim 1: fisher, GS Method &GS, Censoring at &Cens, Rho at &rho, NULL = &NULL";
proc freq data=fisher;
where analysis=1;
table stop / out=stop; table atleast1 / out=atleast1; table allfour / out=allfour;
run;

title"Interim 2: fisher, GS Method &GS, Censoring at &Cens, Rho at &rho, NULL = &NULL";
proc freq data=int2_fisher noprint;
where analysis=2;
table atleast1 / out=a; table allfour / out=b; table stop / out=c;
run;

data c; set c stop;
proc freq data=c;
weight count;
table stop; run;

data a; set a atleast1;
proc freq data=a;
weight count;
table atleast1; run;

data b; set b allfour;
proc freq data=b;
weight count;
table allfour; run;

title"Final: fisher, GS Method &GS, Censoring at &Cens, Rho at &rho, NULL = &NULL";
proc freq data=final_fisher noprint;
where analysis=3;
table atleast1 / out=aa; table allfour / out=bb; table stop / out=cc;
run;
data c; set c cc;
proc freq data=c;
weight count;
  table stop; run;

data a; set a aa;
proc freq data=a;
weight count;
  table atleast1; run;

data b; set b bb;
proc freq data=b;
weight count;
  table allfour; run;

%mend;

*--- sample code to invoke summary macro;

%summarize(GS=OF, cens=0, rho=25, NULL=Y);
%summarize(GS=OF, cens=40, rho=25, NULL=Y);
%summarize(GS=OF, cens=80, rho=25, NULL=Y);
%summarize(GS=OF, cens=95, rho=25, NULL=Y);


