PROCEDURES FOR MULTIPLE OUTCOME MEASURES WITH APPLICATIONS TO MULTIPLE SCLEROSIS CLINICAL TRIALS

By

Payhsuan Daphne Guh

B.Sc. University of British Columbia, 1995

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in

THE FACULTY OF GRADUATE STUDIES

DEPARTMENT OF STATISTICS

We accept this thesis as conforming

to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA

November 1997

In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the head of my department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Department of Statistics
The University of British Columbia
2075 Wesbrook Place
Vancouver, Canada
V6T 1W5

Date: Dec. 5, 1997

Abstract

In planning clinical trials in many subject areas, researchers often find it difficult to designate one single outcome measure as the primary endpoint to describe treatment efficacy. When a disease affects a patient's functions in multiple dimensions, expecting one outcome measure to assess treatment efficacy in a comprehensive way may not be realistic. Multiple sclerosis (MS) is one such complex disease. The topic addressed in this thesis concerns approaches for the design and analysis of clinical trials where a multidimensional outcome measure is used to measure treatment efficacy. The most common approach is to select a single primary endpoint for formal statistical testing with all other outcome measures considered as secondary. This thesis is concerned with the situation where agreement on a single primary endpoint is not possible so that methods based on multiple endpoints are required.

Five methods, Bonferroni adjustment, Hotelling's T^2, O'Brien's OLS and GLS statistics and disjunctive outcome measures are examined and compared through power and sample size calculations. Our discussion of these methods is focused on two-armed (placebo and treatment) randomized clinical trials based on continuous outcome measures. We assume that the data to be analyzed are the changes in the responses from the baseline to the end of the trial and the underlying distribution of the multiple outcome measures can be approximated as multivariate normal. Our investigation is focused on the features of the configuration of the standardized differences in the underlying population means and the correlation structure among the multiple outcome measures. Specifically, several special cases are examined to highlight the main differences among the statistical properties of these methods. We also apply the methods considered to two

MS clinical trial data sets for a more focused comparison of these methods for actual MS patient populations.

Table of Contents

List of Tables

List of Figures

Acknowledgment

I am grateful to a number of people for helping me in the preparation of this thesis.

Foremost I would like to thank my supervisor, Dr. John Petkau, for his advice, ideas, support and encouragement throughout the last two years. I would also like to thank Dr. Harry Joe for his C programs and his very helpful suggestions and comments on improving the manuscript. As well, thanks to Dr. Donald E. Goodkin and to Dr. Gary Cutter and Ms. Monika Baier of the National Multiple Sclerosis Society Clinical Outcomes Assessment Task Force for providing the data sets used in Chapter 4 of the thesis.

I must also express my gratitude towards my family and friends for their support and encouragement. Thank you Dad, Mom, Michelle, Michael, Brandon, Karen, Tiffany, Gladys, Kathy, Jennifer, Jessie and Friendy.

Finally, I would like to take this opportunity to especially thank my close friend, Howard Chang, for his constant encouragement and invaluable help in C programming.

Chapter 1

Introduction

In planning clinical trials in many subject areas, researchers often find it difficult to designate one single outcome measure as the primary endpoint to describe treatment efficacy. When a disease affects a patient's functions in multiple dimensions, expecting one outcome measure to assess treatment efficacy in a comprehensive way may not be realistic. Multiple sclerosis (MS) is one such complex disease. The fact that the most widely used outcome measure for evaluating MS, Kurtzke's Expanded Disability Status Scale (EDSS) (Kurtzke 1983), is based on a neurological examination involving nine functional systems, such as ambulation, cognitive function, upper extremity function and so on, indicates the multidimensional nature of MS. The question of how to construct a multidimensional outcome measure is a fundamental and challenging problem but this is not our focus here. The topic addressed in this thesis is concerned with approaches for the design and analysis of clinical trials where a multidimensional outcome measure is used to measure treatment efficacy.

Suppose that the researchers have identified the most relevant dimensions for describing treatment efficacy. In addition, suppose they have selected what they believe to the most appropriate component measures for the individual dimensions. The focus throughout this thesis will be on the issue arising subsequently: what statistical methods can be applied to the design and analysis of clinical trials when treatment efficacy is described by multiple outcome measures?

The discussion of statistical and design issues for MS clinical trials with multiple

outcome measures in Petkau (1996) motivates our work in this thesis. In that chapter, three statistical methods for dealing with multiple outcome measures were examined for the case of equally correlated outcome measures. Our investigation is within the same general framework but includes several extension. We investigate two additional methods and study the case of unequally correlated outcome measures. We also apply these methods to two MS clinical trial data sets for a more focused comparison among these methods for actual MS patient populations.

In the thesis, our discussion of the statistical methods will be focused on two-armed (placebo and treatment) and randomized clinical trials with continuous outcome measures. The data to be analyzed are the changes in the responses from the baseline to the end of the trial. If $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ denote the vector of the mean changes of all outcome measures on the placebo arm and the treatment arm respectively, then the parameter of interest is $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$, the difference in the population mean changes. In MS clinical trials, these changes measure the patients' functional deterioration in the relevant dimensions, so a lowering of the mean change will correspond to a beneficial effect of the therapy.

A simple approach to the problem of assessing treatment efficacy described by multiple outcome measures is to carry out the comparisons on the individual outcome measures separately with adjusted Type I error levels. Simplicity is the main advantage of this approach as the assessment of treatment efficacy is made for the individual outcome measures. However, this approach may result in a lack of power as the relationships among the outcome measures are not taken into account; this is its main limitation. An alternate approach is to combine all the information from the individual outcome measures into a single prespecified composite outcome measure and use this composite outcome measure as the single primary endpoint to assess the relative efficacy of the two arms. The EDSS is an example of such a prespecified composite outcome measure. Although this approach provides an overall summary of treatment efficacy, the interpretation of

the treatment effect on a composite outcome measure can be difficult as the roles of the individual outcome measures are no longer clear. The main difficulty with this approach is that it is not obvious how to construct such an composite outcome measure. Developing a reliable, sensitive and widely accepted prespecified composite outcome measure requires a great deal of empirical assessment and validation of the outcome measures in current use.

We consider five statistical methods employing one of the above two approaches. Two of these have long been available while the remaining three are more recently developed statistical methodology. The methods based upon Bonferroni adjustment and Hotelling's T^2 which are commonly used for comparisons of multivariate samples are discussed first in Chapter 2. In the former method, separate tests for comparing the treatment arms are carried out on each of the outcome measures. For Hotelling's T^2, a single summary statistic based on the vector of all outcome measures is used. The method based on the Hotelling's T^2 can be thought of as being based on a combination of the individual Z-statistics for comparing the two arms. Due to limitations of these two standard methods (see Sections 2.1 and 2.2), several new methods have been proposed. Two composite outcome measures introduced by O'Brien (1984) consisting of linear combinations of the individual Z-statistics which we will refer to as OLS and GLS statistics, are also discussed in Chapter 2. In Chapter 3, we consider a different type of composite outcome measure, called a disjunctive composite outcome measure. With this method, the original individual outcome measures are first transformed to binary responses indicating changes of clinical significance on the individual outcome measures. The composite outcome measure employed as the single primary endpoint is then defined as an indication of treatment failure if a patient has a significant clinical change on any of the individual outcome measures. Thus, the disjunctive composite outcome measure is simply a binary response.

The methods are compared through power and sample size calculations. Specifically, we evaluate and compare the power achieved by each method with a fixed sample size per arm, at specified alternatives. In addition, we compare the sample size required for each method to achieve a specified power at specific alternatives. In our power and sample size calculations, we consider the number of outcome measures ranging from 1 to 20. In most MS studies, the number of clinical dimensions range from 3 to 5; therefore, these will be of most interest to us.

Our investigation is restricted to the case where the underlying distribution of the multiple outcome measures follows the multivariate normal distribution. We can therefore focus our investigation on the features of the configuration of the standardized differences in the underlying population means and the correlation structure among the multiple outcome measures. A thorough comparison of these methods requires consideration of many possibilities for these aspects of the probabilistic structure. We focus on several special cases intended to highlight the main differences among the statistical properties of these methods. In Chapters 2 and 3, three configurations of the standardized differences are considered. In the first configuration, only one of the multiple outcome measures is effective in comparing the two arms (Case A). This case is intended to illustrate the impact of the inclusion of ineffective outcome measures. The second configuration involves successive outcome measures of diminishing effectiveness in comparing the two arms (Case B). This case allows examination of whether it is beneficial to include such outcome measures. In the third configuration, all of the multiple outcome measures are equally effective in comparing the two arms (Case C). These configurations represent three special cases of multivariate problems: Case A and Case C represent worst and best case scenarios and Case B is intermediate.

With respect to the pattern of correlations, much of our investigation is focused on equally correlated outcome measures with common correlations of 0, 0.3 and 0.5.

Only moderate values of ρ are considered, because researchers in MS clinical trials are aware of the fact that the inclusion of highly correlated outcome measures adds little information. In fact, it often adds noise to the assessment of the relative efficacy of the treatment. Therefore, avoiding the inclusion of highly correlated outcomes is one criterion for designing MS studies; see Rudick et al. (1996). These configurations of the standardized differences and patterns of correlations among the outcome measures are used throughout our work in Chapters 2 and 3.

In Chapter 2, we show that O'Brien's OLS and GLS statistics are equivalent for equally correlated outcome measures. Therefore, our subsequent investigation is focused on other correlation structures which highlight the differences between these two procedures. This work is limited to three and five outcome measures as that covers a reasonable range of the number of clinical dimensions relevant to MS clinical trials. Throughout our investigation for unequally correlated outcomes, the correlations between any two outcomes are classified either low ($\rho = 0.2$), mild ($\rho = 0.5$) or high ($\rho = 0.7$). We also compare O'Brien's GLS to the methods based on Bonferroni adjustment and Hotelling's T^2 for unequally correlated outcomes.

Because the disjunctive composite outcome measure involves the use of dichotomized outcome measures, before investigating its performance in Chapter 3, we first examine dichotomized tests on a single continuous outcome variable (see Section 3.1). The issue of how much information is lost by dichotomized tests compared to the Z-test on the sample means of the continuous outcome variable is addressed. Percent power loss and asymptotic relative efficiency are used to compare these two tests.

Several other methods are available for comparing samples with multiple endpoints. As discussed in Pocock et al. (1987), the most common approach is to select a single primary endpoint for formal statistical testing with all other outcome measures considered as secondary. The design of the study and the assessment of the relative efficacy

of the two arms is based on the primary endpoint while the information provided by the other outcomes is viewed as exploratory. This thesis is concerned with the situation where agreement on a single primary endpoint is not possible so that methods based on multiple endpoints are required. Tang et al. (1993) discussed the dramatic decrease in power of GLS when the true directions of the treatment effects are not consistent. They noted that with the GLS procedure it is possible for endpoints to receive negative weights. This feature can result in that the directions of the components of GLS statistic are inconsistent and motivated them to consider a modification of the O'Brien's approach. They proposed an approximate likelihood ratio (ALR) statistic to account for that limitation. The statistic consists only nonnegative components. Wittes has provided a maximum score test based on the average of the maximum of the responses on the individual outcome measures; see Follmann (1995). For the special case of two outcome measures, Follmann (1995) discussed settings where O'Brien's GLS test and the ALR test can be clinically misleading. This motivated him to propose the risk score test whose rejection boundary corresponds to contours of constant risk and therefore clinically appealing. This test requires the multiple outcome measures to be surrogates; that is, some analysis of the endpoints on an ancillary data set is required to determine the risk score weights of these endpoints. This may not be applicable in some settings. In addition, because the risk score test was examined only for the case of two endpoints with paired data, its general definition and performance are not clear. Due to difficulties of computational implementation and practical issues, these methods are not considered further in this thesis.

In Chapter 4, we apply the methods discussed in Chapters 2 and 3 to two MS clinical trial data sets. Power and sample size calculations guided by patient characteristics in these data sets provide a more focused comparison among these methods for actual MS patient populations. The sample correlations among the outcome measures guide

our choices of the pattern of correlations and the configurations of the standardized differences considered in the underlying population means are suggested by the treatment effects observed in these data sets.

The thesis concludes with Chapter 5 where we make some concluding remarks based on the work reported in the earlier chapters.

Chapter 2

Several Approaches to Multiple Outcome Measures

Suppose we are in the following two-armed clinical trial setting: A total of $2n$ patients participate in the study with an equal number of patients assigned to the placebo arm and to the treatment arm. The experimenter will take measurements on m outcome measures for each patient and these m outcome measures are continuous response variables. We will assume that the variability of the responses and the correlation structure of the responses are the same on both arms in the population of interest. The experimenter's objective is to assess the treatment efficacy.

We will use the notation X_{ijk} to represent the jth outcome variable for the kth patient in the treatment group i where $j = 1, \ldots, m; k = 1, \ldots, n$, and $i = 1$ (placebo), 2 (treatment). Let $\boldsymbol{X}_{ik}$ denote the column vectors of length m containing the responses of the kth patient on all the outcome variables. We will assume that $\boldsymbol{X}_{ik}$ are independently distributed and each follows a multivariate normal distribution with mean vector $\boldsymbol{\mu}_i$ and known common variance-covariance matrix $\boldsymbol{\Sigma}$:

$$\boldsymbol{X}_{ik} \sim \boldsymbol{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}).$$

We can express $\boldsymbol{\Sigma}$ as a product of the matrix containing the information on the variances, $\boldsymbol{V}$, and the correlation matrix, $\boldsymbol{M}_\rho$:

$$\boldsymbol{\Sigma} = \boldsymbol{V}^{\frac{1}{2}} \boldsymbol{M}_\rho \boldsymbol{V}^{\frac{1}{2}},$$

where

$$\boldsymbol{V}^{\frac{1}{2}} = \begin{pmatrix} \sigma_1 & 0 & 0 & \cdots & 0 \\ 0 & \sigma_2 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \sigma_{m-1} & 0 \\ 0 & 0 & \cdots & 0 & \sigma_m \end{pmatrix}$$

and

$$\boldsymbol{M}_\rho = \begin{pmatrix} 1 & \rho_{12} & \rho_{13} & \cdots & \rho_{1m} \\ \rho_{12} & 1 & \rho_{23} & \cdots & \rho_{2m} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \rho_{1,m-1} & \rho_{2,m-1} & \cdots & 1 & \rho_{m,m-1} \\ \rho_{1m} & \rho_{2m} & \cdots & \rho_{m-1,m} & 1 \end{pmatrix}.$$

To simplify the notation, let

$$\overline{\boldsymbol{X}}_1 = \frac{1}{n}\sum_{k=1}^{n} \boldsymbol{X}_{1k},$$

$$\overline{\boldsymbol{X}}_2 = \frac{1}{n}\sum_{k=1}^{n} \boldsymbol{X}_{2k},$$

$$\mu_{1j} - \mu_{2j} = \delta_j,$$

$$\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \boldsymbol{\delta} = (\delta_1, \ldots, \delta_m)'.$$

Then,

$$\overline{\boldsymbol{X}}_1 - \overline{\boldsymbol{X}}_2 \sim \boldsymbol{N}(\boldsymbol{\delta}, \frac{2}{n}\boldsymbol{\Sigma}).$$

Marginally, $\overline{X}_{1j} - \overline{X}_{2j}$ follows a normal distribution with mean δ_j, and variance $\frac{2}{n}\sigma_j^2$. The Z-statistic for comparing the two arms on the jth outcome measure, Y_j can be obtained by standardizing $\overline{X}_{1j} - \overline{X}_{2j}$:

$$Y_j = \frac{\overline{X}_{1j} - \overline{X}_{2j}}{\sqrt{\frac{2}{n}}\sigma_j}.$$

Let $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_m)'$ and $\boldsymbol{\Delta} = (\Delta_1, \Delta_2, \ldots, \Delta_m)'$, where $\Delta_j = \frac{\delta_j}{\sigma_j}$, the standardized difference between the two arms on jth outcome measure. Then, we have

$$Y_j \sim N(\sqrt{\frac{n}{2}}\Delta_j, 1) \tag{2.1}$$

and

$$\boldsymbol{Y} \sim \boldsymbol{N}(\sqrt{\frac{n}{2}}\boldsymbol{\Delta}, \boldsymbol{M}_\rho). \tag{2.2}$$

The objective is to make inferences about the difference between the mean vectors, $\boldsymbol{\delta} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$. The first question of interest might be "Is $\boldsymbol{\delta} = \mathbf{0}$?". In this chapter, we will consider methods to address that question. We will explore the statistical properties of each method and compare the performance of the methods under a few specific circumstances such as uncorrelated outcome measures and equally correlated outcome measures. The methods are evaluated and compared through power and sample size calculations. To be more specific, the power achieved by each method with a fixed sample size per arm, at a fixed significance level, α, and a specific alternative, $\boldsymbol{\delta} = \boldsymbol{\delta}^*$ will be computed and compared. Similarly, the sample size required for each method with a significance level of α to achieve a specified power at a specific alternative will be compared as well.

A complete and comprehensive comparison of the methods requires consideration of many possibilities which arise from different configurations of the standardized differences in the underlying means, $\boldsymbol{\Delta}$, and different correlation structures, $\boldsymbol{M}_\rho$. Only a few

special cases will be investigated and hopefully the main differences among the statistical properties of these approaches will be apparent. We will consider the following three configurations of the standardized differences:

Case A : $\Delta_1 = \Delta^*, \Delta_2 = \cdots = \Delta_m = 0$. In this configuration, only the first outcome measure effectively compares the two arms. We are interested in seeing how these procedures penalize the inclusion of outcome measures which do not effectively compare the two arms.

Case B : $\Delta_1 = \Delta^*, \Delta_2 = \Delta^*/2, \cdots, \Delta_m = \Delta^*/m$. In this configuration, successive outcome measures are of diminishing effectiveness in comparing the two arms. We want to examine whether it is beneficial to include such outcome measures.

Case C : $\Delta_1 = \Delta_2 = \cdots = \Delta_m = \Delta^*$. In this configuration, the individual outcome measures are all equally effective in comparing the two arms.

The correlation structures to be considered will be specified together with the discussion in Sections 2.4 and 2.5.

2.1 Bonferroni Adjustment

The first method to be discussed in this chapter is perhaps the most common approach to multiple comparisons of two arms on different response variables. The idea of this method is to carry out individual comparisons separately but to adjust the Type I error level of the individual comparisons so that the overall probability of making any Type I error is no larger than the desired significance level α. Suppose we carry out each test at the significance level α^*. The statistical test for the jth outcome measure is

$$\text{Reject } H_{oj} : \delta_j = 0 \text{ in favour of } H_{aj} : \delta_j \neq 0 \text{ if } \mid \overline{X}_{1j} - \overline{X}_{2j} \mid > \ t_j,$$

where $t_j = \sqrt{\frac{2}{n}}\sigma_j z_{1-\frac{\alpha^*}{2}}$ is chosen such that $P(\mid \overline{X}_{1j} - \overline{X}_{2j} \mid > t_j \mid \delta_j = 0) \leq \alpha^*$. If we think of testing the equality of the standardized differences of the population means, the test can be re-expressed as follows:

$$\text{Reject } H_{oj} : \Delta_j = 0 \text{ in favour of } H_{aj} : \Delta_j \neq 0 \text{ if } \mid Y_j \mid > z_{1-\alpha^*/2}.$$

Based upon the Bonferroni inequality (Miller, 1981), we have the following result:

Result 2.1 *If* $P(\mid Y_j \mid > z_{1-\alpha^*/2} \mid \delta_j = 0) \leq \frac{\alpha}{m}$ *for* $j = 1, 2, \ldots, m$, *then*
$P(\mid Y_1 \mid > z_{1-\alpha^*/2}, or \mid Y_2 \mid > z_{1-\alpha^*/2}, \ldots, or \mid Y_m \mid > z_{1-\alpha^*/2} \mid \boldsymbol{\delta} = \mathbf{0}) \leq \alpha$.

In other words, if we carry out each of the individual comparisons at the significance level $\alpha^* = \alpha/m$, the overall probability of making any Type I error is ensured to be no larger than the desired significance level α. Note that Result 2.1 does not depend on the assumed normality.

For the special case of independent outcome measures, there is an exact adjustment based on the use of $\alpha^* = 1 - (1 - \alpha)^{1/m}$; that is, if the probability of making a Type I error for the individual comparisons is $1 - (1 - \alpha)^{1/m}$, the overall probability of making any Type I error is α.

2.1.1 Power and Sample Size Calculations

The overall power of this procedure is the probability of making one or more rejections of the individual null hypotheses; in other words, it is the probability that the two samples show a significant difference in one or more outcome variables. It is easier to evaluate this as the complement of the probability that the two samples show no difference in all of the outcome variables:

$$\begin{aligned} Power_{\boldsymbol{\Delta}=\boldsymbol{\Delta}^*} &= P(\text{one or more rejection of } H_{0j} \text{ for } j = 1, \ldots, m \mid \boldsymbol{\delta} = \boldsymbol{\delta}^*) \\ &= 1 - P(\text{no rejection of } H_{0j} \text{ for } j = 1, \ldots, m \mid \boldsymbol{\delta} = \boldsymbol{\delta}^*). \end{aligned}$$

Letting $Z_j = Y_j - \sqrt{\frac{n}{2}}\Delta_j^*$, then $\boldsymbol{Z} = (Z_1, \ldots, Z_m)'$ is distributed as $\boldsymbol{N}(\boldsymbol{0}, \boldsymbol{M}_\rho)$, and we obtain

$$\begin{aligned}
Power_{\boldsymbol{\Delta}=\boldsymbol{\Delta}^*} &= 1 - P(|\, Y_1 \,| \leq z_{1-\alpha^*/2}, |\, Y_2 \,| \leq z_{1-\alpha^*/2}, \ldots, |\, Y_m \,| \leq z_{1-\alpha^*/2} \mid \boldsymbol{\Delta} = \boldsymbol{\Delta}^*) \\
&= 1 - P(-z_{1-\alpha^*/2} \leq Y_j \leq z_{1-\alpha^*/2} \ \textit{for}\ j = 1, \ldots, m \mid \boldsymbol{\Delta} = \boldsymbol{\Delta}^*) \\
&= 1 - P(-z_{1-\alpha^*/2} - \sqrt{\frac{n}{2}}\Delta_j^* \leq Z_j \leq z_{1-\alpha^*/2} - \sqrt{\frac{n}{2}}\Delta_j^* \ \textit{for}\ j = 1, \ldots, m) \\
&= 1 - \int_{a_m}^{b_m} \cdots \int_{a_1}^{b_1} f_{\mathbf{Z}}(\mathbf{z})\, dz_1 \cdots dz_m \qquad (2.3)
\end{aligned}$$

where $a_j = -z_{1-\alpha^*/2} - \sqrt{\frac{n}{2}}\Delta_j^*$ and $b_j = z_{1-\alpha^*/2} - \sqrt{\frac{n}{2}}\Delta_j^*$.

For the special case of uncorrelated outcome measures, this expression can be simplified to the following:

$$\begin{aligned}
Power &= 1 - \prod_{j=1}^{m} P(a_j \leq Z_j \leq b_j) \\
&= 1 - \prod_{j=1}^{m} \int_{a_j}^{b_j} \phi(z_j)\, dz_j \\
&= 1 - \prod_{j=1}^{m} \left[\Phi(b_j) - \Phi(a_j)\right], \qquad (2.4)
\end{aligned}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are the univariate standard normal density function and cumulative distribution function.

In general, the expression for power for the correlated outcome measures would have to be evaluated by numerical integration. The C codes, written by Dr. H. Joe, for approximating multivariate normal rectangle probabilities based on conditional expectations (Joe, 1995) are used here. To be more specific, second order approximation is used. Note that for the special case of equally correlated outcome measures, this calculation can be reduced to one dimensional normal probability calculation (Johnson and Kotz, 1972).

As there is no closed form expression for the sample size required to achieve a specified power at an alternative $\boldsymbol{\Delta}^*$, we evaluate it numerically. Writing (2.3) and (2.4) in the

form of $g(n) = 0$, we can express the sample size required as the root of $g(n)$. The Newton-Raphson method is perhaps the most well-known numerical method for solving such a root-finding problems but it requires exact evaluation of derivatives of the non-linear equations. To avoid this inconvenience, the quasi-Newton method is used instead; a C routine is used to numerically obtain the derivatives at each iteration.

2.2 Hotelling's T^2 Statistic

One can imagine the possibility that the evidence of differences between the samples on each individual outcome measure is not strong, but all the evidence combined results in a significant overall difference. That is, when we perform statistical tests for each individual outcome separately, no significant difference between the two arms is shown; however, when we carry out a single global comparison of the two arms which combines the evidence from the individual outcomes, a significant difference is detected. The Bonferroni adjustment approach does not allow one to explore this possibility since it only carries out separate individual comparisons. Now, we will look at methods which allow us to compare the two samples on all m outcome measures simultaneously.

One simple and common approach is based on the use of the Hotelling's T^2 statistic, the multivariate version of the Student's t statistic. For testing $H_0 : \boldsymbol{\delta} = \mathbf{0}$ against $H_a : \boldsymbol{\delta} = \boldsymbol{\delta}^*$, Hotelling's T^2 statistic is given by:

$$\begin{aligned} T^2 &= (\overline{\boldsymbol{X}}_1 - \overline{\boldsymbol{X}}_2)'(\frac{2}{n}\boldsymbol{\Sigma})^{-1}(\overline{\boldsymbol{X}}_1 - \overline{\boldsymbol{X}}_2) \\ &= \frac{n}{2}(\overline{\boldsymbol{X}}_1 - \overline{\boldsymbol{X}}_2)'\boldsymbol{V}^{-1/2}\boldsymbol{M}_\rho{}^{-1}\boldsymbol{V}^{-1/2}(\overline{\boldsymbol{X}}_1 - \overline{\boldsymbol{X}}_2) \\ &= \boldsymbol{Y}'\boldsymbol{M}_\rho{}^{-1}\boldsymbol{Y}. \end{aligned}$$

For the special case of uncorrelated outcome measures,

$$T^2 = Y_1^2 + Y_2^2 + \cdots + Y_m^2,$$

which is simply the sum of the squared Z-statistics for comparing the two arms on the individual outcomes.

As the T^2 statistic sums up the squared Z-statistics, it does not take account of the direction of the differences between the two arms on the individual outcomes. This is the main limitation of this procedure. The question of whether one arm is better than the other is not being addressed; rather this procedure simply addresses the question of whether the two arms are different. We will therefore consider approaches which attempt to overcome this limitation in the next section.

The T^2 statistic has a χ^2_m distribution when $\overline{\boldsymbol{X}}_1 - \overline{\boldsymbol{X}}_2$ is multivariate normally distributed. Under the null hypothesis, $H_0 : \boldsymbol{\delta} = \mathbf{0}$, T^2 is distributed as χ^2_m and under the alternative hypothesis, $H_a : \boldsymbol{\delta} = \boldsymbol{\delta}^*$, T^2 is distributed as $\chi^2_m(\lambda^2)$ where λ^2 is the noncentrality parameter given by

$$\begin{aligned} \lambda^2 &= (\boldsymbol{\delta}^*)'(\frac{2}{n}\boldsymbol{\Sigma})^{-1}(\boldsymbol{\delta}^*) \\ &= \frac{n}{2}(\boldsymbol{\Delta}^*)'\boldsymbol{M}_\rho{}^{-1}(\boldsymbol{\Delta}^*) \end{aligned} \tag{2.5}$$

2.2.1 Power and Sample Size Calculations

The power of this procedure can be easily obtained once the distribution of the T^2 statistic under H_a is specified. The function **pchisq** in S-Plus calculates the cumulative probability for the χ^2 distribution and for the non-central χ^2 distribution as well. To determine the sample size required to achieve a specified power for a level α statistical test, the magnitude of the noncentrality parameter required to achieve this power needs to be calculated first. Once the magnitude of the noncentrality parameter, λ^2, is determined,

the required sample size, n, can be easily evaluated using (2.5). For the special case of equally correlated outcome measures, the expression for λ^2 can be simplified to:

$$\lambda^2 = \frac{n}{2(1-\rho)}(\boldsymbol{\Delta}^*)' \left(\boldsymbol{I} - \frac{\rho}{1+(m-1)\rho}\boldsymbol{J} \right) (\boldsymbol{\Delta}^*),$$

where $\boldsymbol{J}$ is a $m \times m$ matrix with all elements equal to 1. The derivation of this expression appears in Appendix A.

2.2.2 The Non-centrality Parameters for Cases A, B, and C

The non-centrality parameter plays an important role in the power achieved and sample size required for this procedure. Let m^{ij} denote $(\boldsymbol{M}_\rho{}^{-1})_{ij}$. For the three configurations of the standardized differences between the underlying means, Cases A, B, and C, we can simplify the expression for λ^2:

For Case A where $\boldsymbol{\Delta}^* = (\Delta^*, 0, 0, \cdots, 0)'$,

$$\lambda^2 = \frac{n}{2}(\boldsymbol{\Delta}^*)^2 m^{11}. \tag{2.6}$$

For Case B where $\boldsymbol{\Delta}^* = (\Delta^*, \Delta^*/2, \cdots, \Delta^*/m)'$,

$$\lambda^2 = \frac{n}{2}(\Delta^*)^2 \sum_{i=1}^{m}\sum_{j=1}^{m} \frac{m^{ij}}{ij}. \tag{2.7}$$

For Case C where $\boldsymbol{\Delta}^* = (\Delta^*, \Delta^*, \cdots, \Delta^*)'$,

$$\lambda^2 = \frac{n}{2}(\Delta^*)^2 \sum_{i=1}^{m}\sum_{j=1}^{m} m^{ij}. \tag{2.8}$$

For Case A, (2.6), (2.7), and (2.8) indicate that $\boldsymbol{M}_\rho{}^{-1}$ affects λ^2 only through m^{11}. For Case C, $\boldsymbol{M}_\rho{}^{-1}$ affects λ^2 only through the sum of all its elements. Case B is more complicated as λ^2 is affected through the weighted sum of the elements of $\boldsymbol{M}_\rho{}^{-1}$. The impact of the individual elements in $\boldsymbol{M}_\rho{}^{-1}$ on λ^2 is different; the further from m^{11} the element lies, the more its contribution to λ^2 is diluted through the weights. For later reference, we will denote m^{11}, $\sum_{i=1}^{m}\sum_{j=1}^{m}\frac{m^{ij}}{ij}$ and $\sum_{i=1}^{m}\sum_{j=1}^{m} m^{ij}$ as $\tilde{\lambda}_A^2$ $\tilde{\lambda}_B^2$, and $\tilde{\lambda}_C^2$ respectively.

2.3 Linear Combinations of Z-Statistics

In this section, two additional composite outcome measures based on linear combinations of the Z-statistics for comparing the two arms on individual outcomes will be discussed. A randomized clinical trial comparing two therapies for the treatment of diabetes with responses on 34 outcome measures on a total of 11 patients motivated O'Brien (1985) to examine procedures for comparing samples with multiple endpoints. O'Brien indicated that the approaches based on Bonferroni adjustment and Hotelling's T^2 statistic were perhaps most commonly used in the comparison of multivariate samples; however, there are some limitations of these two approaches. He suggested that the Bonferroni procedure may lack power when all the outcome measures are effective in comparing the two arms, particularly when the number of outcome measures is large relative to the sample size. He also argued that the Hotelling's T^2 procedure basically addresses the wrong question as we have already indicated. Therefore, he proposed three alternative composite outcome measures for the comparison of multivariate samples which are intended to overcome these limitations. Only two of these will be considered here as the other proposed method is a rank-based procedure which may be most suitable for use with ordinal outcome measures.

2.3.1 O'Brien's OLS Statistic

Suppose we consider each of the Z-statistics on the m outcome measures an unbiased estimator of the true standardized treatment efficacy, ξ. O'Brien's OLS statistic, denoted by $\widehat{\beta}_{OLS}$, is the linear combination of these Z-statistics which minimizes the sum of squares between the estimate of ξ and the individual Z-statistics:

$$\begin{aligned} \widehat{\beta}_{OLS} &= \arg_{\xi} \min \sum_{j=1} m(Y_j - \xi)^2 \\ &= \frac{Y_1 + Y_2 + \ldots + Y_m}{m}. \end{aligned}$$

The expectation and variance of $\widehat{\boldsymbol{\beta}}_{OLS}$ can be easily obtained from (2.1):

$$\begin{aligned} E(\widehat{\boldsymbol{\beta}}_{OLS}) &= \frac{1}{m} E(\sum_{j=1}^{m} Y_j) \\ &= \frac{1}{m} \sum_{j=1}^{m} \sqrt{\frac{n}{2}} \Delta_j \\ &= \sqrt{\frac{n}{2}} \overline{\Delta}; \end{aligned} \tag{2.9}$$

$$\begin{aligned} Var(\widehat{\boldsymbol{\beta}}_{OLS}) &= \frac{1}{m^2} Var\left(\sum_{j=1}^{m} Y_j\right) \\ &= \frac{1}{m^2}\left(\sum_{j=1}^{m} Var(Y_j) + \sum_{i \neq j}^{m} Cov(Y_i, Y_j)\right) \\ &= \frac{1}{m^2}\left(m + \sum_{i \neq j}^{m} \rho_{ij}\right) \\ &= \frac{1}{m^2}\left[m + m(m-1)\overline{\rho}\right] \\ &= \frac{1}{m}\left[1 + (m-1)\overline{\rho}\right]. \end{aligned} \tag{2.10}$$

As we assume that the underlying data follows a multivariate normal distribution, $\widehat{\boldsymbol{\beta}}_{OLS}$ is normally distributed with mean $\sqrt{\frac{n}{2}}\overline{\Delta}$ and variance $\frac{1}{m}\left[1 + (m-1)\overline{\rho}\right]$. To be specific, under the null hypothesis: $\boldsymbol{\Delta} = \mathbf{0}, \widehat{\boldsymbol{\beta}}_{OLS} \sim N\left(0, \frac{1}{m}\left[1 + (m-1)\overline{\rho}\right]\right)$, and under the alternative hypothesis: $\boldsymbol{\Delta} = \boldsymbol{\Delta}^*, \widehat{\boldsymbol{\beta}}_{OLS} \sim N\left(\sqrt{\frac{n}{2}}\overline{\Delta}^*, \frac{1}{m}\left[1 + (m-1)\overline{\rho}\right]\right)$.

The general formulae for the power achieved and the sample size required per arm by a level α test for comparing two population means are derived in Appendix B. Based on the results in Appendix B, the formulae for the power and the approximate sample size required per arm for O'Brien's OLS procedure are readily obtained:

$$Power_{\boldsymbol{\Delta}=\boldsymbol{\Delta}^*} = 1 - \Phi\left(z_{1-\alpha/2} - \frac{\sqrt{n/2}\overline{\Delta}^*}{\sqrt{\frac{1}{m}(1 + (m-1)\overline{\rho})}}\right)$$

$$+\Phi\left(-z_{1-\alpha/2}-\frac{\sqrt{n/2}\overline{\Delta}^{*}}{\sqrt{\frac{1}{m}(1+(m-1)\overline{\rho})}}\right) \tag{2.11}$$

and

$$n \approx \frac{2\left(\frac{1}{m}\left(1+(m-1)\overline{\rho}\right)\right)\left(z_{1-\frac{\alpha}{2}}+z_{1-\beta}\right)^{2}}{\overline{\Delta}^{*2}} \tag{2.12}$$

O'Brien's OLS statistic is simply the equally weighted average of the Z-statistics for comparing the two arms on the individual outcomes. Unlike the T^2 statistic, $\widehat{\boldsymbol{\beta}}_{OLS}$ takes the direction of the differences between the placebo arm and the treatment arm on each outcome measure into consideration. However, the limitation of this approach is that correlations among the outcome measures are not taken into account. This motivates the proposal of O'Brien's GLS statistic discussed next.

2.3.2 O'Brien's GLS Statistic

O'Brien's GLS statistic, denoted as $\widehat{\boldsymbol{\beta}}_{GLS}$, is the linear combination of the Z-statistics which minimizes the weighted sum of squares between ξ and the individual Z-statistics. The idea is to weight the individual Z-statistics according to the correlation among the outcome measures. It is sensible to down-weight any two highly correlated outcomes since they provide very similar information concerning the relative efficacy of the two arms. On the other hand, if two outcomes are almost uncorrelated, the weights on these two outcomes should be relatively larger. We will first define $\widehat{\boldsymbol{\beta}}_{GLS}$:

$$\widehat{\boldsymbol{\beta}}_{GLS}=\frac{\mathbf{1}'\boldsymbol{M}_{\rho}{}^{-1}\boldsymbol{Y}}{\mathbf{1}'\boldsymbol{M}_{\rho}{}^{-1}\mathbf{1}},$$

where $\mathbf{1}$ is a $m \times 1$ vector with all elements equal to 1.

The expectation and variance of $\widehat{\boldsymbol{\beta}}_{GLS}$ can be easily obtained from (2.2):

$$E(\widehat{\boldsymbol{\beta}}_{GLS}) = \frac{\mathbf{1}'\boldsymbol{M}_{\rho}{}^{-1}E(\boldsymbol{Y})}{\left(\mathbf{1}'\boldsymbol{M}_{\rho}{}^{-1}\mathbf{1}\right)}$$

$$= \sqrt{\frac{n}{2}}\frac{\mathbf{1}'\boldsymbol{M}_\rho^{-1}\boldsymbol{\Delta}}{(\mathbf{1}'\boldsymbol{M}_\rho^{-1}\mathbf{1})}$$

$$\begin{aligned} Var(\widehat{\beta}_{GLS}) &= \frac{Var(\mathbf{1}'\boldsymbol{M}_\rho^{-1}\boldsymbol{Y})}{(\mathbf{1}'\boldsymbol{M}_\rho^{-1}\mathbf{1})^2} \\ &= \frac{\mathbf{1}'\boldsymbol{M}_\rho^{-1}Var(\boldsymbol{Y})(\mathbf{1}'\boldsymbol{M}_\rho^{-1})'}{(\mathbf{1}'\boldsymbol{M}_\rho^{-1}\mathbf{1})^2} \\ &= \frac{\mathbf{1}'\boldsymbol{M}_\rho^{-1}\mathbf{1}}{(\mathbf{1}'\boldsymbol{M}_\rho^{-1}\mathbf{1})^2} \\ &= \frac{1}{\mathbf{1}'\boldsymbol{M}_\rho^{-1}\mathbf{1}}. \end{aligned}$$

As we assume that the underlying data follows a multivariate normal distribution, $\widehat{\beta}_{GLS}$ is normally distributed. Based on the results in Appendix B, the formulae for the power and sample size required per arm for the O'Brien's GLS procedure can be easily obtained as follows:

$$\begin{aligned} Power_{\boldsymbol{\Delta}=\boldsymbol{\Delta}^*} &= 1-\Phi\left(z_{1-\alpha/2}-\frac{\sqrt{n/2}(\mathbf{1}'\boldsymbol{M}_\rho^{-1}\boldsymbol{\Delta}^*)}{\sqrt{\mathbf{1}'\boldsymbol{M}_\rho^{-1}\mathbf{1}}}\right) \\ &\quad + \Phi\left(-z_{1-\alpha/2}-\frac{\sqrt{n/2}(\mathbf{1}'\boldsymbol{M}_\rho^{-1}\boldsymbol{\Delta}^*)}{\sqrt{\mathbf{1}'\boldsymbol{M}_\rho^{-1}\mathbf{1}}}\right) \end{aligned} \tag{2.13}$$

and

$$n \approx \frac{2(\mathbf{1}'\boldsymbol{M}_\rho^{-1}\mathbf{1})(z_{1-\frac{\alpha}{2}}+z_{1-\beta})^2}{(\mathbf{1}'\boldsymbol{M}_\rho^{-1}\boldsymbol{\Delta}^*)^2} \tag{2.14}$$

For the special cases of either uncorrelated or equally correlated outcome measures, $\widehat{\boldsymbol{\beta}}_{GLS}=\widehat{\boldsymbol{\beta}}_{OLS}$ as is shown in Appendix C. We use the S-plus functions **qnorm** and **pnorm** to evaluate the quantiles and the cumulative probabilities for the power and sample size calculations for both the OLS and GLS tests.

We have limited our computations to the case of normally distributed data. However, as long as the joint distribution of $\boldsymbol{Y}$, the vector of Z-statistics, can be reasonably approximated by the multivariate normal distribution, the procedures we have discussed can be applied and the numerical results which follow will be relevant.

2.4 Comparisons for Equally Correlated Outcome Measures

Because GlS and OLS are equivalent for the case of equally correlated outcome measures, the comparisons in this section are made among Bonferroni adjustment, Hotelling's T^2 and OLS. To investigate these procedures, the correlation structure among the m outcome measures needs to be specified. First, we will consider the exchangeable form for the correlation structure where all outcome measures are equally correlated:

$$\boldsymbol{M}_\rho = \begin{pmatrix} 1 & \rho & \rho & \cdots & \rho \\ \rho & 1 & \rho & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \rho & \rho & \cdots & 1 & \rho \\ \rho & \rho & \cdots & \rho & 1 \end{pmatrix} = (1-\rho)\boldsymbol{I} + \rho\boldsymbol{J}.$$

Among all the possible values of ρ, we specifically examine:

1. $\rho = 0$ which corresponds to uncorrelated outcome measures.
2. $\rho = 0.3$ which corresponds to mildly correlated outcome measures.
3. $\rho = 0.5$ which corresponds to modestly correlated outcome measures.

The powers of two-sided tests of 5% significance level in comparing two arms with 100 patients per arm for Cases A, B, and C are presented in Table 2.1. Note that for a single

outcome measure, the three methods are equivalent. The value of Δ^* which identified the specific alternative is chosen so that for a single outcome measure with 100 patients per arm, a level $\alpha = 5\%$ two-sided comparison of the two arms will achieve a power of 0.80; this value is $\Delta^* = 0.396232$.

We will first discuss the special case of uncorrelated outcome measures. For Case A, the power of all procedures decreases monotonically with the inclusion of additional outcomes. All three procedures penalize the inclusion of outcome measures which are not effective in comparing the two arms although the penalty is substantially heavier for O'Brien's OLS procedure as the decrease in power is dramatic even with the inclusion of only one ineffective outcome measure. The performance of Bonferroni adjustment and T^2 are roughly comparable although the former has a slight advantage over the latter and this advantage become noticeable as the number of outcome measures increases.

On the contrary, for Case C, the power of all three procedures increases monotonically with the inclusion of additional equally effective outcomes. Moreover, this increase in power is dramatic for all procedures. All procedures perform very well. However, O'Brien's OLS has a clear advantage over the other two and Bonferroni adjustment is least competitive.

Case B is more complicated. The inclusion of additional outcome measures with diminishing effectiveness has a deleterious effect on all three procedures, except for the inclusion of the first additional one. OLS has a clear advantage over the other two as this deleterious effect is only mild on OLS. The impact on the method based on Bonferroni adjustment is substantially larger; for example, the inclusion of a single ineffective outcome measure already has a detrimental effect on Bonferroni adjustment. The impact on T^2 is mild when m is small but becomes modest as m gets larger. Similarly to Case C, Bonferroni adjustment is not competitive with OLS and T^2.

Table 2.1 indicates that the inclusion of ineffective outcomes can result in drastic

deterioration in the performance of these procedures. Furthermore, the inclusion of too many weakly effective outcomes leads to such deterioration as well. On the other hand, the performance of these procedures is impressive for Case C. In particular, OLS and T^2, which share the characteristic that the evidence provided by the individual outcomes is summarized into a global statistic to give an overall assessment of the treatment efficacy, are very powerful.

We now translate this comparison of power into comparison of the required sample sizes. Table 2.2 provides the sample sizes required to achieve a power of 0.80 when the common correlation $\rho = 0$, 0.3 or 0.5. The results for $\rho = 0$ indicate that the required sample sizes differ substantially even when the power differs only slightly. For example, in Cases B and C, slight differences in power between OLS and T^2 lead to moderate differences in the required sample sizes. The results for Case A indicate that the substantial differences in power between OLS and the other two procedures lead to huge differences in the required sample sizes.

We next turn to the examination of the impact of positive correlation on the performance of these procedures. The discussion for the case of $\rho = 0$ has highlighted the issue relevant to trials with multiple outcome measures. Our discussion for $\rho = 0.3$ and 0.5 will focus on the sample sizes required rather than power because the former provides equivalent comparisons which are of greater relevance for designing clinical trials.

We first examine the effect of positive correlation on the Bonferroni adjustment procedure. Positive correlation among the multiple outcomes has a negative impact on the Bonferroni adjustment procedure. In Case A, for a fixed number of ineffective outcomes, the magnitude of the common correlation has very small negative impact on the required sample size. In addition, regardless of the magnitude of the common correlation, the impact of the inclusion of additional ineffective outcomes on the performance of this procedure is roughly the same. In Case C, any positive correlation dilutes the evidence

Table 2.1: Power of procedures with n =100 for equally correlated outcome measures

			m=total number of outcome measures						
Case	ρ	Procedure	1	2	3	4	5	10	20
A	0.0	Bonferroni	0.80	0.72	0.67	0.63	0.61	0.52	0.44
		Hotelling's T^2	0.80	0.71	0.64	0.60	0.56	0.43	0.31
		O'Brien's OLS	0.80	0.51	0.37	0.29	0.24	0.14	0.10
	0.3	Bonferroni	0.80	0.71	0.66	0.62	0.59	0.50	0.42
		Hotelling's T^2	0.80	0.75	0.72	0.69	0.66	0.59	0.43
		O'Brien's OLS	0.80	0.41	0.25	0.17	0.13	0.07	0.06
	0.5	Bonferroni	0.80	0.71	0.66	0.62	0.59	0.50	0.41
		Hotelling's T^2	0.80	0.83	0.83	0.82	0.81	0.73	0.61
		O'Brien's OLS	0.80	0.37	0.21	0.14	0.11	0.07	0.05
B	0.0	Bonferroni	0.80	0.77	0.73	0.70	0.68	0.58	0.49
		Hotelling's T^2	0.80	0.81	0.79	0.77	0.75	0.65	0.51
		O'Brien's OLS	0.80	0.84	0.84	0.83	0.82	0.74	0.62
	0.3	Bonferroni	0.80	0.74	0.70	0.66	0.63	0.54	0.45
		Hotelling's T^2	0.80	0.73	0.67	0.62	0.59	0.52	0.46
		O'Brien's OLS	0.80	0.74	0.65	0.56	0.49	0.27	0.14
	0.5	Bonferroni	0.80	0.73	0.68	0.64	0.61	0.52	0.43
		Hotelling's T^2	0.80	0.71	0.66	0.65	0.64	0.64	0.63
		O'Brien's OLS	0.80	0.68	0.55	0.45	0.38	0.20	0.11
C	0.0	Bonferroni	0.80	0.92	0.96	0.98	0.988	0.999	1.000
		Hotelling's T^2	0.80	0.95	0.990	0.998	1.000	1.000	1.000
		O'Brien's OLS	0.80	0.98	0.998	1.000	1.000	1.000	1.000
	0.3	Bonferroni	0.80	0.88	0.91	0.93	0.94	0.96	0.98
		Hotelling's T^2	0.80	0.89	0.91	0.92	0.92	0.91	0.85
		O'Brien's OLS	0.80	0.94	0.97	0.98	0.988	0.996	0.998
	0.5	Bonferroni	0.80	0.85	0.87	0.88	0.89	0.91	0.92
		Hotelling's T^2	0.80	0.83	0.83	0.82	0.81	0.73	0.61
		O'Brien's OLS	0.80	0.90	0.93	0.94	0.95	0.97	0.97

provided by the different outcome measures resulting in a larger required sample size. Moreover, the benefit gained from including more equally effective outcomes diminishes as the common correlation increases. As the total number of outcomes increases, the negative impact of positive correlation increases. In Case B, the effect of positive correlation for a fixed number of outcome is intermediate. It can be seen that for each fixed number of outcomes in Table 2.2, an increase of the common correlation results in an increase in the required sample size.

We now turn to the impact of positive correlation on the procedure based on Hotelling's T^2 statistic. The effects of a common positive correlation among the multiple outcomes on T^2 are more complicated. In Case A, for a fixed number of outcome measures, the required sample size decreases as the correlation increases. In Case C, similar dilution of evidence occurs as for the Bonferroni adjustment procedure, but the effect on the required sample size is greater for Hotelling's T^2. In Case B, it can be shown that for each fixed number of outcomes, there exists a particular value of ρ such that smaller positive correlation has a negative impact on T^2 and larger positive correlation has a positive impact. For instance, when the total number of outcomes is equal to 5, for a common correlation above about 0.30, the required sample size decreases as the correlation increases.

Positive correlation among the multiple outcome measures has a deleterious effect on O'Brien's OLS. As shown in (2.12), the sample size for OLS is directly proportional to $1 + (m-1)\overline{\rho}$ and inversely proportional to $(\overline{\Delta}^*)^2$. As we can see from Tables 2.1 and 2.2, any positive correlation has a negative impact on OLS and this impact becomes dramatic as the total number of outcomes gets larger. Since the correlations among the outcomes affect the properties of OLS only through $\overline{\rho}$ and the standardized differences in the underlying means on the two arms affect its properties only through their average, $\overline{\Delta}^*$, any correlation structures with the same value of $\overline{\rho}$ or similarly any configurations of the standardized differences with the same value of $\overline{\Delta}^*$ will yield the same properties.

Table 2.2: Sample size required to achieve power of 0.80 with equally correlated outcome measures

			m=total number of outcome measures						
Case	ρ	Procedure	1	2	3	4	5	10	20
A	0.0	Bonferroni	100	120	132	140	147	167	187
		Hotelling's T^2	100	123	139	152	163	207	267
		O'Brien's OLS	100	200	300	400	500	1000	2000
	0.3	Bonferroni	100	121	133	142	149	169	190
		Hotelling's T^2	100	112	120	126	132	158	196
		O'Brien's OLS	100	260	480	760	1100	3700	13400
	0.5	Bonferroni	100	121	133	142	149	170	190
		Hotelling's T^2	100	92	93	95	98	114	140
		O'Brien's OLS	100	300	600	1000	1500	5500	21000
B	0.0	Bonferroni	100	108	116	123	129	150	172
		Hotelling's T^2	100	95	102	107	112	134	167
		O'Brien's OLS	100	89	89	92	96	117	155
	0.3	Bonferroni	100	114	125	134	141	162	184
		Hotelling's T^2	100	118	133	144	152	170	184
		O'Brien's OLS	100	116	143	175	211	431	1035
	0.5	Bonferroni	100	118	130	139	146	167	188
		Hotelling's T^2	100	123	133	137	137	134	136
		O'Brien's OLS	100	133	179	230	288	641	2900
C	0.0	Bonferroni	100	72	61	55	50	40	33
		Hotelling's T^2	100	61	46	38	33	21	13
		O'Brien's OLS	100	50	33	25	20	10	5
	0.3	Bonferroni	100	80	73	68	65	58	54
		Hotelling's T^2	100	80	74	72	72	77	89
		O'Brien's OLS	100	65	53	48	44	37	34
	0.5	Bonferroni	100	87	82	79	78	74	72
		Hotelling's T^2	100	92	93	95	98	114	140
		O'Brien's OLS	100	75	67	63	60	55	53

Taking the three procedures together for an overall comparison, in Case A, with the exception that Hotelling's T^2 benefits from the inclusion of a few ineffective outcomes when the correlation among them is large enough, positive correlation has a deleterious effect on all procedures. This effect is substantially more dramatic for OLS than the other two procedures. Generally speaking, when the correlation is very mild, Bonferroni adjustment has the advantage especially when m is large. On the other hand, when the correlation becomes modest, Hotelling's T^2 performs better. In Case C, OLS has a clear advantage. Hotelling's T^2 has a clear advantage over Bonferroni adjustment when the common correlation is small, but the latter performs better when ρ becomes larger and, in particular, when m is large as well. In Case B, except for the special case of $\rho = 0$, the inclusion of rather weakly effective outcome measures results in a deterioration of the performance for all three procedures. For the special case of uncorrelated outcome measures, Bonferroni adjustment is the least powerful procedure and OLS performs best. However, as the deleterious effect of the magnitude of ρ on the performance of Bonferroni adjustment is slight and that on OLS is substantial, Bonferroni adjustment has a clear advantage over OLS even when ρ is modest. When $\rho = 0.3$, Bonferroni adjustment has a slight advantage over T^2. On the other hand, when $\rho = 0.5$, T^2 has the advantage over Bonferroni adjustment, although Bonferroni adjustment is competitive with T^2 when only a few weakly effective outcomes are included.

2.5 Comparisons for Unequally Correlated Outcome Measures

In this section, the main focus will be the comparison between O'Brien's OLS and GLS procedures as we want to explore the potential of GLS to be a more powerful procedure than OLS. Subsequently, we will bring the methods based on Bonferroni adjustment and T^2 into the comparison with GLS.

As O'Brien's OLS and GLS are equivalent when the outcomes are equally correlated, we now want to consider a few other correlation structures to get a better understanding of the properties of GLS. In the following, we will classify the correlations between any two outcomes as either weakly correlated (L), mildly correlated (M) or highly correlated (H) and restrict ourselves to consideration of corresponding ρ to be 0.2, 0.5 and 0.7 respectively. In addition, our examination will be limited to $m = 3$ and 5 as in most MS clinical trials, the number of clinical dimensions ranges from 3 to 5.

2.5.1 Three Outcome Measures

If the total number of outcomes is 3, we are limited to a total of 27 possible patterns of the correlations among the three outcome variables. (Note that 3 of these correspond to equally correlated cases for which OLS and GLS are equivalent.) The comparisons between OLS and GLS are summarized in Tables 2.3 to 2.9, where the results for the Bonferroni adjustment and Hotelling's T^2 are also provided. We will make the comparison between OLS and GLS through their effect sizes as the procedure with a large effect size has the larger power. The effect sizes (taken here for convenience, as mean/sd, ignoring the common factor of $\sqrt{\frac{n}{2}}$) of the OLS and GLS statistics, denoted ζ_{OLS} and ζ_{GLS} are:

$$\zeta_{OLS} = \frac{\overline{\Delta}^*}{\sqrt{\frac{1}{m}\left[1 + (m-1)\overline{\rho}\right]}},$$

$$\zeta_{GLS} = \frac{\mathbf{1}'\boldsymbol{M}_\rho{}^{-1}\boldsymbol{\Delta}^*}{\sqrt{\mathbf{1}'\boldsymbol{M}_\rho{}^{-1}\mathbf{1}}}.$$

Table 2.3 provides the effect sizes of OLS and GLS statistics, as well as $\overline{\rho}$ and the weights GLS assigns to the individual Z-statistics, denoted as w_j. Table 2.4 presents the power achieved with 100 patients per arm for the OLS and GLS statistics for each of the 27 possible patterns of correlations for Case A. Table 2.5 translates the comparison of

power for Case A into the comparison of the sample size requirements. Tables 2.6 and 2.7 present the corresponding results for Case B and those for Case C appear in Tables 2.8 and 2.9.

GLS versus OLS

As indicated by (2.9) and (2.10), the correlations among the outcome measures affect the properties of OLS only through $\overline{\rho}$. Therefore, the results presented in Tables 2.1 and 2.2 for equally correlated outcomes can be relevant, depending upon the values of $\overline{\rho}$ for the correlation structures under consideration here. For example, $\overline{\rho} = 0.30$ the patterns $(\rho_{12}, \rho_{23}, \rho_{23})$ = (L, L, M), (L, M, L), and (M, L, L). Hence, the performance of OLS will be identical for these patterns and for three equally correlated outcomes with common correlation of 0.30, where the configuration of standardized differences has the same value of $\overline{\Delta^*}$. The patterns (L, M, M), (M, L, M), and (M, M, L) have an average correlation of $\overline{\rho} = 0.40$. In this situation, OLS will be less powerful than with the previous three patterns but more powerful than for the case of three equally correlated outcomes with $\rho = 0.50$ illustrated in Tables 2.1 and 2.2.

We now turn to the comparison between GLS and OLS. In Case A where only the first outcome measure is effective in comparing the two arms, if the weight GLS assigns to the effective outcome is larger than that assigned by OLS, GLS will have a larger effect size and therefore an advantage. The magnitude of this advantage depends upon how much larger ζ_{GLS} is than ζ_{OLS} . For example, the patterns (L, H, H) and (H, L, H) represent situations where one of the two ineffective outcomes is highly correlated and the effective outcome is almost independent of the highly correlated outcome. In this situation, the advantage of GLS is most substantial. For GLS, the weight on the effective outcome is very large, $w_1 = 0.75$, while the weight OLS assigns is only 0.33. As a result, the effect size of GLS is substantially larger than that for OLS (0.40 compared

Table 2.3: For $m = 3$: weights, effect sizes, and noncentrality parameters

Correlation				Weights			Case A			Case B			Case C		
ρ_{12}	ρ_{13}	ρ_{23}	$\bar{\rho}$	w_1	w_2	w_3	$\tilde{\lambda}_A^2$	ζ_{OLS}	ζ_{GLS}	$\tilde{\lambda}_B^2$	ζ_{OLS}	ζ_{GLS}	$\tilde{\lambda}_C^2$	ζ_{OLS}	ζ_{GLS}
L	L	L	.20	.33	.33	.33	1.07	.19	.19	1.10	.35	.35	2.14	.58	.58
L	L	M	.30	.40	.30	.30	1.06	.18	.22	1.09	.33	.36	1.90	.54	.55
L	L	H	.37	.44	.28	.28	1.05	.17	.24	1.10	.32	.36	1.79	.52	.53
L	M	L	.30	.30	.40	.30	1.35	.18	.16	1.15	.33	.33	1.90	.54	.55
L	M	M	.40	.42	.42	.16	1.34	.17	.22	1.24	.31	.35	1.71	.51	.52
L	M	H	.47	.50	.50	.00	1.42	.16	.26	1.43	.30	.38	1.67	.49	.51
L	H	L	.37	.28	.44	.28	1.98	.17	.15	1.39	.32	.31	1.79	.52	.53
L	H	M	.47	.50	.50	.00	2.08	.16	.26	1.71	.30	.38	1.67	.49	.51
L	H	H	.53	.75	.75	−.50	2.90	.16	.40	2.69	.29	.51	1.82	.48	.53
M	L	L	.30	.30	.30	.40	1.35	.18	.16	1.02	.33	.32	1.90	.54	.55
M	L	M	.40	.42	.16	.42	1.34	.17	.22	1.02	.31	.33	1.71	.51	.52
M	L	H	.47	.50	.00	.50	1.42	.16	.26	1.04	.30	.34	1.67	.49	.51
M	M	L	.40	.16	.42	.42	1.71	.17	.09	1.04	.31	.27	1.71	.51	.52
M	M	M	.50	.33	.33	.33	1.50	.16	.16	1.04	.30	.30	1.50	.49	.49
M	M	H	.57	.42	.29	.29	1.42	.16	.19	1.06	.29	.31	1.42	.47	.47
M	H	L	.47	.00	.50	.50	2.67	.16	.00	1.28	.30	.21	1.67	.49	.51
M	H	M	.57	.29	.42	.29	2.08	.16	.14	1.28	.29	.28	1.42	.47	.47
M	H	H	.63	.43	.43	.14	1.96	.15	.20	1.39	.28	.32	1.35	.46	.46
H	L	L	.37	.28	.28	.44	1.98	.17	.15	1.10	.32	.30	1.79	.52	.53
H	L	M	.47	.50	.00	.50	2.08	.17	.26	1.19	.30	.34	1.67	.49	.51
H	L	H	.53	.75	−.50	.75	2.90	.16	.40	1.44	.29	.40	1.82	.48	.53
H	M	L	.47	.00	.50	.50	2.67	.16	.00	1.15	.30	.21	1.67	.49	.51
H	M	M	.57	.29	.29	.42	2.08	.16	.14	1.09	.29	.27	1.42	.47	.47
H	M	H	.63	.43	.14	.43	1.96	.15	.20	1.08	.28	.30	1.35	.46	.46
H	H	L	.53	−.50	.75	.75	5.45	.16	−.27	1.75	.29	.07	1.82	.48	.53
H	H	M	.63	.14	.43	.43	2.88	.15	.07	1.34	.28	.23	1.35	.46	.46
H	H	H	.70	.33	.33	.33	2.36	.15	.15	1.27	.27	.27	1.25	.44	.44

to 0.16). The simultaneous down-weighting of one of the ineffective outcomes enables GLS to make better use of the evidence provided by the effective outcome. On the contrary, if GLS assigns a smaller weight to the effective outcome than OLS does, OLS will have a larger effect size and it will have the advantage. Consider the patterns, (M, H, L) and (H, M, L) as examples. In these two situations, the effective outcome is highly correlated and therefore considered redundant, and GLS assigns it zero weight which results in zero effect size. Consequently, the power achieved is identical to the level α and the required sample size is ∞. One might wonder about the even more extreme pattern, (H, H, L) which represents two weakly correlated but ineffective outcomes plus one highly correlated but effective outcome. With this structure, GLS actually assigns negative weight to the effective outcome. As a result, the effect size is negative. If we perform a one-sided test, GLS will have no power.

Both OLS and GLS are expected to be powerful in Case C because in this configuration, all outcomes are equally effective in comparing the two arms. The means for the OLS and GLS statistics are the same; therefore, the differences between ζ_{GLS} and ζ_{OLS} arise only through differences in their standard deviations; whichever has the smaller standard deviation will have the advantage. For the cases under consideration, this advantage will only be modest as the differences between the effect sizes of GLS and OLS are quite small. The greatest difference in effect sizes, 0.53 for GLS versus 0.48 for OLS, is for the patterns (L, H, H), (H, L, H), and (H, H, L); this results in a modest advantage for GLS.

Similar to Case A, in Case B, the weight GLS assigns to the most effective outcome dominates its effect size and consequently its performance. If this outcome is weighted more heavily by GLS than OLS, ζ_{GLS} is larger than ζ_{OLS} and vise versa. We first look at the patterns for which GLS is substantially more powerful than OLS. The pattern (L, H, H) represents a situation where the two outcomes of more effectiveness are almost

Table 2.4: Case A with $m = 3$: Power achieved with $n = 100$

Correlation			Procedures			
ρ_{12}	ρ_{13}	ρ_{23}	Bon.	T^2	OLS	GLS
L	L	L	0.66	0.68	0.28	0.28
L	L	M	0.66	0.67	0.25	0.35
L	L	H	0.66	0.67	0.23	0.39
L	M	L	0.66	0.79	0.25	0.21
L	M	M	0.66	0.78	0.23	0.33
L	M	H	0.66	0.81	0.21	0.44
L	H	L	0.66	0.93	0.23	0.18
L	H	M	0.66	0.94	0.21	0.44
L	H	H	0.66	0.99	0.20	0.81
M	L	L	0.66	0.79	0.25	0.21
M	L	M	0.66	0.78	0.23	0.33
M	L	H	0.66	0.81	0.21	0.44
M	M	L	0.66	0.88	0.23	0.09
M	M	M	0.66	0.83	0.21	0.21
M	M	H	0.66	0.81	0.20	0.28
M	H	L	0.66	0.98	0.21	0.05
M	H	M	0.66	0.94	0.20	0.17
M	H	H	0.66	0.92	0.19	0.29
H	L	L	0.66	0.93	0.23	0.18
H	L	M	0.66	0.94	0.21	0.44
H	L	H	0.66	0.99	0.20	0.81
H	M	L	0.66	0.98	0.21	0.05
H	M	M	0.66	0.94	0.20	0.17
H	M	H	0.66	0.92	0.19	0.29
H	H	L	0.66	1.00	0.20	0.47
H	H	M	0.66	0.99	0.19	0.08
H	H	H	0.66	0.96	0.18	0.18

Table 2.5: Case A with $m = 3$: Sample size required to achieve power of 0.80

Correlation			Procedures			
ρ_{12}	ρ_{13}	ρ_{23}	Bon.	T^2	OLS	GLS
L	L	L	133	130	420	420
L	L	M	133	131	480	317
L	L	H	133	132	520	278
L	M	L	133	103	480	599
L	M	M	133	104	540	336
L	M	H	133	98	580	240
L	H	L	133	70	520	734
L	H	M	133	67	580	240
L	H	H	133	48	620	98
M	L	L	133	103	480	599
M	L	M	133	104	540	336
M	L	H	133	98	580	240
M	M	L	133	81	540	2100
M	M	M	133	93	600	600
M	M	H	133	98	640	416
M	H	L	133	52	580	∞
M	H	M	133	67	640	816
M	H	H	133	71	680	404
H	L	L	133	70	520	734
H	L	M	133	67	580	240
H	L	H	133	48	620	98
H	M	L	133	52	580	∞
H	M	M	133	67	640	816
H	M	H	133	71	680	404
H	H	L	133	25	620	220
H	H	M	133	48	680	3640
H	H	H	133	59	720	720

Table 2.6: Case B with $m = 3$: Power achieved with $n = 100$

Correlation			Procedures			
ρ_{12}	ρ_{13}	ρ_{23}	Bon.	T^2	OLS	GLS
L	L	L	0.71	0.69	0.71	0.71
L	L	M	0.70	0.69	0.65	0.71
L	L	H	0.70	0.69	0.62	0.72
L	M	L	0.70	0.71	0.65	0.64
L	M	M	0.70	0.75	0.60	0.70
L	M	H	0.70	0.81	0.57	0.77
L	H	L	0.70	0.80	0.62	0.60
L	H	M	0.70	0.88	0.57	0.77
L	H	H	0.70	0.98	0.54	0.95
M	L	L	0.69	0.65	0.65	0.61
M	L	M	0.69	0.66	0.60	0.65
M	L	H	0.68	0.66	0.57	0.67
M	M	L	0.68	0.66	0.60	0.47
M	M	M	0.68	0.66	0.55	0.55
M	M	H	0.68	0.67	0.53	0.59
M	H	L	0.67	0.76	0.57	0.33
M	H	M	0.67	0.76	0.53	0.51
M	H	H	0.67	0.80	0.50	0.61
H	L	L	0.68	0.69	0.62	0.56
H	L	M	0.68	0.73	0.57	0.67
H	L	H	0.67	0.82	0.54	0.81
H	M	L	0.67	0.71	0.57	0.33
H	M	M	0.67	0.69	0.53	0.49
H	M	H	0.67	0.68	0.50	0.55
H	H	L	0.66	0.89	0.54	0.08
H	H	M	0.66	0.78	0.50	0.37
H	H	H	0.66	0.76	0.48	0.48

Table 2.7: Case B with $m = 3$: Sample size required to achieve power of 0.80

Correlation			Procedures			
ρ_{12}	ρ_{13}	ρ_{23}	Bon.	T^2	OLS	GLS
L	L	L	122	126	125	125
L	L	M	123	127	143	123
L	L	H	124	126	155	121
L	M	L	124	121	143	147
L	M	M	124	112	161	126
L	M	H	125	97	173	107
L	H	L	125	100	155	159
L	H	M	125	81	173	107
L	H	H	125	52	184	60
M	L	L	127	136	143	156
M	L	M	128	136	161	143
M	L	H	128	134	173	135
M	M	L	130	134	161	221
M	M	M	130	133	179	179
M	M	H	130	131	190	164
M	H	L	130	109	173	346
M	H	M	130	109	190	197
M	H	H	130	100	202	156
H	L	L	130	126	155	176
H	L	M	130	117	173	135
H	L	H	130	98	184	98
H	M	L	132	121	173	346
H	M	M	132	127	190	211
H	M	H	132	129	202	180
H	H	L	133	79	184	3520
H	H	M	133	104	202	297
H	H	H	133	109	214	214

Table 2.8: Case C with $m = 3$: Power achieved with $n = 100$

Correlation			Procedures			
ρ_{12}	ρ_{13}	ρ_{23}	Bon.	T^2	OLS	GLS
L	L	L	0.93	0.95	0.98	0.98
L	L	M	0.91	0.92	0.97	0.97
L	L	H	0.90	0.90	0.96	0.96
L	M	L	0.91	0.92	0.97	0.97
L	M	M	0.89	0.88	0.95	0.96
L	M	H	0.88	0.87	0.94	0.95
L	H	L	0.90	0.90	0.96	0.96
L	H	M	0.88	0.87	0.94	0.95
L	H	H	0.87	0.90	0.92	0.97
M	L	L	0.91	0.92	0.97	0.97
M	L	M	0.89	0.88	0.95	0.96
M	L	H	0.88	0.87	0.94	0.95
M	M	L	0.89	0.88	0.95	0.96
M	M	M	0.87	0.83	0.93	0.93
M	M	H	0.86	0.81	0.91	0.92
M	H	L	0.88	0.87	0.94	0.95
M	H	M	0.86	0.81	0.91	0.92
M	H	H	0.84	0.79	0.90	0.90
H	L	L	0.90	0.90	0.96	0.96
H	L	M	0.88	0.87	0.94	0.95
H	L	H	0.87	0.90	0.92	0.97
H	M	L	0.88	0.87	0.94	0.95
H	M	M	0.86	0.81	0.91	0.92
H	M	H	0.84	0.79	0.90	0.90
H	H	L	0.87	0.90	0.92	0.97
H	H	M	0.84	0.79	0.90	0.90
H	H	H	0.83	0.75	0.88	0.88

Table 2.9: Case C with $m = 3$: Sample size required to achieve power of 0.80

correlation			procedures			
ρ_{12}	ρ_{13}	ρ_{23}	Bon.	T^2	OLS	GLS
L	L	L	68	65	47	47
L	L	M	73	73	53	53
L	L	H	76	78	58	56
L	M	L	73	73	53	53
L	M	M	77	81	60	58
L	M	H	81	83	64	60
L	H	L	76	78	58	56
L	H	M	81	83	64	60
L	H	H	85	76	69	55
M	L	L	73	73	53	53
M	L	M	77	81	60	58
M	L	H	81	83	64	60
M	M	L	77	81	60	58
M	M	M	82	93	67	67
M	M	H	86	98	71	71
M	H	L	81	83	64	60
M	H	M	86	98	71	71
M	H	H	90	103	76	74
H	L	L	76	78	58	56
H	L	M	81	83	64	60
H	L	H	85	76	69	55
H	M	L	81	83	64	60
H	M	M	86	98	71	71
H	M	H	90	103	76	74
H	H	L	85	76	69	55
H	H	M	90	103	76	74
H	H	H	94	111	80	80

independent of each other and the least effective outcome is highly correlated. The weights on the first two rather effective outcomes are relatively larger ($w_1 = w_2 = 0.75$) than that on the least effective one ($w_3 = -0.50$) which results in a substantial difference between the effect sizes of GLS and OLS (0.51 versus 0.29). Similar but less extreme examples are the patterns (L, M, H) and (L, H, M) which correspond to two relatively weakly correlated and effective outcomes plus one highly correlated and less effective outcome. For these patterns, GLS assigns no weight to the least effective outcome and equal weights to the other two. For the patterns (M, L, H) and (H, L, M), the second outcome is relatively weakly correlated, as was the third outcome in (L, M, H) or (L, H, M), and it is assigned zero weight; however, GLS is again more powerful than OLS. Another similar example is the pattern (H, L, H) which corresponds to the most and least effective outcome being only weakly correlated with each other and the moderately effective outcome being highly correlated. In this case, equal weights are assigned to the first and third outcomes ($w_1 = w_3 = 0.75$) and a large negative weight to the second outcome ($w_2 = -0.50$). Even when the moderately effective outcome is so negatively weighted, as long as the most effective outcome is heavily weighted, GLS still has a clear advantage over OLS.

Nevertheless, for a few patterns of correlations, GLS is less powerful than OLS. Contrary to (L, H, H), the pattern (H, H, L) describes a situation where the most effective outcome is highly correlated with the other weakly correlated outcomes. A large negative weight therefore is assigned to the first outcome and GLS performs poorly in this situation. Less extreme examples are provided by the patterns (M, H, L) and (H, M, L), as contrasted with (L, M, H) and (L, H, M). With these patterns, the first outcome which is relatively highly correlated with the other moderately correlated outcome variables is assigned no weight by GLS. OLS is much more powerful than GLS in these situations.

Bonferroni Adjustment

The results in Tables 2.4 and 2.5 show that in Case A, the correlation structures seem to have essentially no impact on the performance of Bonferroni adjustment. This agrees with the results in Table 2.2 where in Case A, positive correlation has only a small impact.

The results in Tables 2.6 and 2.7 indicate that in Case B, the correlation structure has only a little more impact on the performance of Bonferroni adjustment than in Case A. Generally speaking, when the more effective outcomes are less correlated, this procedure performs slightly better. For example, the pattern (L, M, H) describes a situation where the most effective outcome is least correlated and the least effective outcome is most correlated. On the contrary, the pattern (H, M, L) represents a situation where the most effective outcome is heavily correlated and the least effective is modestly correlated. Bonferroni adjustment performs better in the former situation. (Compare power of 0.67 to 0.70.)

In Case C, the impact of the correlation structure on the performance of Bonferroni adjustment is more apparent. The results in Tables 2.8 and 2.9 reveal that a smaller degree of correlation among the outcome measures results in better performance of this procedure. Among the 27 patterns of correlation, it performs the best for the pattern (L, L, L) (power of 0.93) and the worst for the pattern (H, H, H) (power of 0.83). Generally speaking, in this case, the required sample size is roughly proportional to $\overline{\rho}$. For example, the average of the correlation for each of the patterns (L, L, M), (L, M, L) and (M, L, L) is 0.30 and the performance of Bonferroni adjustment for these patterns is identical to that for the pattern of three equally correlated outcomes with the common correlation of 0.30.

Hotelling's T^2

The effects of correlation structures on the procedure based on Hotelling's T^2 are more complicated. As indicated by (2.6), (2.7), and (2.8), the magnitude of the non-centrality parameter, λ^2 is proportional to m^{11}, the weighted sum of the elements in $\boldsymbol{M}_\rho{}^{-1}$ and the sum of all the elements in $\boldsymbol{M}_\rho{}^{-1}$ for Cases A, B, and C respectively. We will discuss the effect of the correlation structure on T^2 through these quantities; Table 2.3 provides the relevant information.

In Case A, among the 27 patterns, T^2 performs most powerfully for the pattern (H, H, L). The inverse of this correlation matrix is:

$$\boldsymbol{M}_\rho{}^{-1} = \begin{pmatrix} 5.45 & -3.18 & -3.18 \\ -3.18 & 2.90 & 1.65 \\ -3.18 & 1.65 & 2.90 \end{pmatrix}.$$

The relatively large $m^{11} = 5.45$ leads a large λ^2 and hence, large power for T^2. The results in Tables 2.4 and 2.5 suggest that this procedure performs better when the degree of correlation relevant to the effective outcome is higher. For example, for the patterns (L, M, L), (M, M, L), and (H, M, L), the power achieved with 100 patients per arm by T^2 is 0.79, 0.88 and 0.98. This agrees with the results for Case A in Table 2.1 where the common correlation increases, the power of the procedure based on Hotelling's T^2 increases. Not only ρ_{12} and ρ_{13} but also ρ_{23} impact on the performance of T^2. For example, when ρ_{12} and ρ_{13} have the patterns (ρ_{12}, ρ_{13}) = (L, L), (M, M), (M, H), (H, M), and (H, H), the power achieved by T^2 increases as ρ_{23} decreases. On the other hand, for (ρ_{12}, ρ_{13}) = (L, H) and (H, L), Hotelling's T^2 performs better when ρ_{23} is larger.

In Case C, the impact of the correlation structures on Hotelling's T^2 is similar to that on Bonferroni adjustment: when the degree of correlation among the outcomes is smaller, T^2 performs better. This impact could be quite substantial. Taking the two

most extreme patterns, (L, L, L) and (H, H, H), for comparison, the required sample size for the former is about $\frac{3}{5}$ that for the latter (65 versus 111).

In Case B, the effect of the correlation structures on T^2 is complicated. The largest power occurs for the pattern (L, H, H). The inverse of this correlation matrix is:

$$\boldsymbol{M}_\rho{}^{-1} = \begin{pmatrix} 2.90 & 1.65 & -3.18 \\ 1.65 & 2.90 & -3.18 \\ -3.18 & -3.18 & 5.45 \end{pmatrix}.$$

As the negative elements are removed from m^{11}, the resulting weighted sum relevant to Case B is quite large (2.69). Compare this to the pattern (M, L, L) whose inverse matrix is:

$$\boldsymbol{M}_\rho{}^{-1} = \begin{pmatrix} 1.34 & -0.71 & 0.09 \\ -0.71 & 1.71 & -0.71 \\ 0.09 & -0.71 & 1.34 \end{pmatrix}.$$

The small positive elements and negative elements lying close to m^{11} lead to a small weighted sum (1.02). As a result, T^2 is not very powerful in this situation.

Overall Comparison

Bringing all the procedures together for an overall comparison, we first discuss Case A where only one outcome measure is effective in comparing the two arms. In Case A, the results in Tables 2.4 and 2.5 reveal the potential of GLS to perform substantially better than OLS. When the correlation relevant to the effective outcome is weak and the correlation between the ineffective outcomes is strong, GLS performs more powerfully. When the correlation pattern is (L, H, H) or (H, L, H), GLS has a clear advantage over Bonferroni adjustment; otherwise, Bonferroni adjustment is substantially more powerful. In Case A, T^2 has a clear advantage over GLS in all 27 correlation structures. Bonferroni

adjustment is competitive with T^2 only when the correlations relevant to the effective outcomes are weak (patterns: (L, L, L), (L, L, M) and (L, L, H)).

In Case C where all outcomes are effective, all procedures perform well, but particularly GLS. In all 27 correlation structures considered, GLS always performs at least as well as OLS. However, because OLS also performs well in this case, the advantage of GLS is quite small. GLS also has a modest advantage over T^2 for the structures considered; the magnitude of this advantage in the required sample size is roughly the same for all patterns considered. For all the patterns of correlation considered, GLS also has a modest advantage over Bonferroni adjustment. In Case C, Bonferroni adjustment and T^2 are rather comparable. For patterns (L, H, H), (H, L, H) and (H, H, L), T^2 has a modest advantage over Bonferroni adjustment. On the other hand, when the degree of correlation among the outcome measures is relatively large, Bonferroni adjustment is more powerful. The patterns (M, M, H), (M, H, M), (M, H, H), (H, M, M), (H, M, H), (H, H, M), and (H, H, H) are examples where Bonferroni adjustment performs better.

In Case B where outcome measures are of diminishing effectiveness, the performance of GLS, OLS and T^2 depends strongly on the correlation structure; this is especially so for GLS. Only when the most effective outcome measure is weakly correlated with the other two outcome measures (patterns (L, L, L), (L, L, M), and (L, L, H)), does GLS have a slight advantage over T^2. For all other patterns considered, T^2 is more powerful. When the degree of correlation relevant to the first outcome measure is large, the advantage of T^2 over GLS can be substantial. The pattern (H, H, L) represents such a situation and T^2 is much more powerful (power of 0.89 versus 0.08). In Case B, GLS has a clear advantage over Bonferroni adjustment when the patterns of correlation are (L, L, H), (L, M, H), (L, H, M), (L, H, H) and (H, L, H). Generally speaking, when the correlation between the first and second outcomes is moderate or high, Bonferroni adjustment performs substantially better; the only exception is the pattern (H, L, H).

2.5.2 Five Outcome Measures

We now turn to the case of $m = 5$. A few examples where the differences between GLS and OLS are pronounced will be considered. We first present the patterns of correlations to be considered among the 5 outcome measures; the resulting weights GLS assigns to each individual outcome are provided in the last row adjoined to the correlation matrices:

$$\boldsymbol{M}_{\rho_1} = \left(\begin{array}{ccccc} 1 & L & L & M & H \\ L & 1 & M & H & H \\ L & M & 1 & H & H \\ M & H & H & 1 & H \\ H & H & H & H & 1 \\ \hline 1.62 & 1.38 & 1.38 & -1.15 & -2.23 \end{array}\right) \quad \boldsymbol{M}_{\rho_2} = \left(\begin{array}{ccccc} 1 & L & M & H & H \\ L & 1 & M & H & H \\ M & M & 1 & M & M \\ H & L & M & 1 & H \\ H & L & M & H & 1 \\ \hline 1.87 & 1.87 & -0.25 & -1.25 & -1.25 \end{array}\right)$$

$$\boldsymbol{M}_{\rho_3} = \left(\begin{array}{ccccc} 1 & L & M & H & H \\ L & 1 & H & H & H \\ M & H & 1 & H & H \\ H & H & H & 1 & H \\ H & H & H & H & 1 \\ \hline 1.50 & 1.50 & 0.00 & -1.00 & -1.00 \end{array}\right) \quad \boldsymbol{M}_{\rho_4} = \left(\begin{array}{ccccc} 1 & L & M & M & H \\ L & 1 & H & H & H \\ M & H & 1 & H & H \\ M & H & H & 1 & H \\ H & H & H & H & 1 \\ \hline 0.75 & 0.75 & 0.00 & 0.00 & -0.50 \end{array}\right)$$

$$\boldsymbol{M}_{\rho_5} = \left(\begin{array}{ccccc} 1 & L & L & L & M \\ L & 1 & L & L & M \\ L & L & 1 & L & M \\ L & L & L & 1 & M \\ M & M & M & M & 1 \\ \hline 0.31 & 0.31 & 0.31 & 0.31 & -0.25 \end{array}\right) \quad \boldsymbol{M}_{\rho_6} = \left(\begin{array}{ccccc} 1 & L & L & L & M \\ L & 1 & L & L & M \\ L & L & 1 & M & H \\ L & L & M & 1 & H \\ M & M & H & H & 1 \\ \hline 0.42 & 0.42 & 0.43 & 0.43 & -0.69 \end{array}\right)$$

$$\boldsymbol{M}_{\rho_7} = \left(\begin{array}{ccccc} 1 & L & L & L & H \\ L & 1 & L & L & H \\ L & L & 1 & L & M \\ L & L & L & 1 & M \\ H & H & M & M & 1 \\ \hline 0.75 & 0.75 & 0.42 & 0.42 & -1.33 \end{array}\right) \quad \boldsymbol{M}_{\rho_8} = \left(\begin{array}{ccccc} 1 & L & L & L & M \\ L & 1 & L & H & H \\ L & L & 1 & H & H \\ L & H & H & 1 & H \\ M & H & H & H & 1 \\ \hline 0.50 & 1.50 & 1.50 & -1.00 & -1.50 \end{array}\right)$$

$$\boldsymbol{M}_{\rho_9} = \left(\begin{array}{ccccc} 1 & L & L & L & M \\ L & 1 & L & L & H \\ L & L & 1 & M & H \\ L & L & M & 1 & H \\ M & H & H & H & 1 \\ \hline 0.50 & 0.97 & 0.71 & 0.71 & -1.88 \end{array}\right) \quad \boldsymbol{M}_{\rho_{10}} = \left(\begin{array}{ccccc} 1 & L & L & M & M \\ L & 1 & H & H & H \\ L & H & 1 & H & H \\ M & H & H & 1 & H \\ M & H & H & H & 1 \\ \hline 0.55 & 0.39 & 0.39 & -0.16 & -0.16 \end{array}\right)$$

$$\boldsymbol{M}_{\rho_{11}} = \left(\begin{array}{ccccc} 1 & L & L & L & M \\ L & 1 & H & H & H \\ L & H & 1 & H & H \\ L & H & H & 1 & H \\ M & H & H & H & 1 \\ \hline 0.50 & 0.25 & 0.25 & 0.25 & -0.25 \end{array}\right) \quad \boldsymbol{M}_{\rho_{12}} = \left(\begin{array}{ccccc} 1 & L & L & L & L \\ L & 1 & H & H & H \\ L & H & 1 & H & H \\ L & H & H & 1 & H \\ L & H & H & H & 1 \\ \hline 0.42 & 0.15 & 0.15 & 0.15 & 0.15 \end{array}\right)$$

$$\boldsymbol{M}_{\rho_{13}} = \left(\begin{array}{ccccc} 1 & H & L & L & L \\ H & 1 & L & L & L \\ L & L & 1 & L & L \\ L & L & L & 1 & L \\ L & L & L & L & 1 \\ \hline 0.15 & 0.15 & 0.24 & 0.24 & 0.24 \end{array}\right) \quad \boldsymbol{M}_{\rho_{14}} = \left(\begin{array}{ccccc} 1 & H & H & L & L \\ H & 1 & L & L & L \\ H & L & 1 & L & L \\ L & L & L & 1 & L \\ L & L & L & L & 1 \\ \hline -0.27 & 0.40 & 0.40 & 0.23 & 0.23 \end{array}\right)$$

$$\boldsymbol{M}_{\rho_{15}} = \left(\begin{array}{ccccc} 1 & H & H & H & M \\ H & 1 & H & M & M \\ H & H & 1 & M & L \\ H & M & M & 1 & L \\ M & M & L & L & 1 \\ \hline -0.35 & -0.06 & 0.46 & 0.43 & 0.52 \end{array}\right) \quad \boldsymbol{M}_{\rho_{16}} = \left(\begin{array}{ccccc} 1 & H & H & H & M \\ H & 1 & H & H & M \\ H & H & 1 & H & L \\ H & H & H & 1 & L \\ M & M & L & L & 1 \\ \hline -0.16 & -0.16 & 0.39 & 0.39 & 0.55 \end{array}\right)$$

GLS versus OLS

To compare GLS with OLS, we first provide their effect sizes and $\overline{\rho}$, the average correlation, in Table 2.10. The results of the power and sample size calculations for GLS and

OLS for these 16 correlation structures for Cases A, B, and C are presented in Tables 2.11 to 2.13.

The structure $\boldsymbol{M}_{\rho_1}$ represents one relatively weakly correlated outcome, the first, two moderately correlated outcomes, the second and third, plus two highly correlated outcomes. For this structure, GLS assigns large negative weights to the two highly correlated outcomes and large positive weights to the remaining three, especially the least correlated one. In Case A, since the outcome with the largest weight is the effective outcome, GLS makes very good use of the information provided by this outcome. OLS is not competitive with GLS because of the big difference between the effect sizes (3.65 versus 0.10). Similarly in Case B, as the most effective outcome is weighted the most, as expected, GLS has a clear advantage. In Case C, the much larger effect size for GLS results in its superiority in power. Taking Cases A, B, and C together, for this particular correlation structure, GLS is much more powerful than OLS. $\boldsymbol{M}_{\rho_2}$ is a similar example. In this structure, two outcomes are relatively less dependent, one moderately dependent and the remaining two are highly dependent. GLS considers the two highly dependent outcomes as redundant outcomes and hence weights them heavily and negatively and assigns large weights to the relatively weakly correlated outcomes. The moderately dependent outcome is assigned a small weight. With the same reasoning as in the previous example, GLS performs substantially better than OLS in all of Cases A, B, and C. $\boldsymbol{M}_{\rho_3}$ and $\boldsymbol{M}_{\rho_4}$ also describe similar situations.

$\boldsymbol{M}_{\rho_5}$ represents a structure where the first four outcomes are weakly correlated and the remaining outcome is equally and moderately dependent with each of the first four. In this situation, GLS weights the dependent outcome negatively while assigning equal and positive weights to the first four outcomes. In Case A, GLS is modestly more powerful than OLS as the weight GLS assigns to the effective outcome is not large; nevertheless, the required sample sizes differ substantially. Similar in Case B, GLS has a clear advantage.

In Case C, because OLS also performs very well, the advantage of GLS is rather limited. $\boldsymbol{M}_{\rho 6}$ is a similar example to $\boldsymbol{M}_{\rho 5}$. The main difference between these two structures is that the fifth outcome in $\boldsymbol{M}_{\rho 6}$ is even more dependent than in $\boldsymbol{M}_{\rho 5}$. As a result, this highly correlated outcome is more negatively weighted and the first four are more positively weighted by GLS. In each of Cases A, B, and C, the advantage of GLS becomes more clear.

$\boldsymbol{M}_{\rho 7}$ corresponds to four weakly correlated outcomes plus one very dependent outcome. This dependent outcome is equally and highly correlated with the first two outcomes; in addition, it is equally and moderately correlated with the remaining two outcomes. Not surprisingly, GLS down-weights this dependent outcome. However, it worth noting that the first two outcomes which seem to be more dependent are actually weighted more heavily than the remaining two. This result seems to suggest that GLS tends to transfer the weight from a redundant outcome to the outcomes which are relatively highly correlated with the redundant outcome. In Cases A and B, both the power and required sample size clearly demonstrate the superiority of GLS. In Case C, the gain by GLS in power is modest (the potential gain is limited as the power of OLS is 0.98); however, the difference in the required sample sizes for GLS and OLS is substantial. $\boldsymbol{M}_{\rho 8}$ and $\boldsymbol{M}_{\rho 9}$ represent similar structures as $\boldsymbol{M}_{\rho 7}$.

The correlation structure $\boldsymbol{M}_{\rho 10}$ describes one less correlated outcome, the first, and two equally and moderately correlated outcomes, the second and third, plus two equally and more heavily correlated outcomes. Under this structure, GLS assigns a larger weight to the first outcome, and least weight to the fourth and fifth outcomes. GLS again performs substantially better than OLS for Cases A and B and modestly better for Case C. The improvement of GLS over OLS is smaller in this example than in $\boldsymbol{M}_{\rho 10}$ since the differences in the weights GLS assigns to the outcome measures are less dramatic. $\boldsymbol{M}_{\rho 11}$ represents a situation where the first outcome is not strongly correlated with any of the

remaining outcomes while these four outcomes are highly correlated among themselves. GLS assigns most weight to the first outcome, least weight to the fifth, and equal weights to the remaining three outcomes. $\boldsymbol{M}_{\rho 12}$ is a similar example except that the degree of correlation relevant to the first outcome is smaller. The pattern of improvement of GLS over OLS for $\boldsymbol{M}_{\rho 11}$ and $\boldsymbol{M}_{\rho 12}$ is similar to that for the structure $\boldsymbol{M}_{\rho 10}$.

We now turn to examples where GLS may perform poorly and therefore OLS could be more powerful. $\boldsymbol{M}_{\rho 13}$ corresponds to a structure in which two outcomes are highly correlated, the remaining three are weakly correlated, and these two sets of outcomes are weakly correlated as well. With this structure, GLS weights the three weakly correlated outcomes more heavily. In Case A, as the weight GLS assigns to the effective outcome is small, GLS performs even more poorly than OLS. In Case B, since the weights on the more effective outcomes are smaller than those on the less effective outcomes, OLS again is more powerful. In Case C, both procedures perform very well and GLS has a slight advantage over OLS. $\boldsymbol{M}_{\rho 14}$ is similar to $\boldsymbol{M}_{\rho 13}$ except that the first outcome is even more dependent as it is also highly correlated with the third outcome. This time, GLS assigns negative weight to the first outcome. Consequently, for Case A, the mean of the GLS statistic is negative. As we have discussed for a similar situation earlier, this fact is not reflected in our power or sample size calculations as the tests we perform are two-sided. $\boldsymbol{M}_{\rho 15}$ represents outcomes of diminishing dependence. Roughly speaking, GLS assigns the weight in an increasing fashion. The first outcome is least and actually negatively weighted and the fifth outcome is most weighted. For Case A, the mean of GLS statistic is again negative. Furthermore, for Case B, the mean is not only negative but also very close to zero. As a consequence, the required sample size is very large. $\boldsymbol{M}_{\rho 16}$ represents a similar situation except that the first and second outcomes have the same correlations relative to the remaining outcomes and hence the first outcome is not so negatively weighted.

Table 2.10: For $m = 5$: average correlation, effect sizes, and noncentrality parameters

Correlation Structure	$\overline{\rho}$	Case A			Case B			Case C		
		$\tilde{\lambda}^2$	ζ_{OLS}	ζ_{GLS}	$\tilde{\lambda}^2$	ζ_{OLS}	ζ_{GLS}	$\tilde{\lambda}^2$	ζ_{OLS}	ζ_{GLS}
$\boldsymbol{M}_{\rho_1}$	0.56	88.50	0.10	3.65	138.68	0.22	4.60	32.50	0.49	2.26
$\boldsymbol{M}_{\rho_2}$	0.57	20.63	0.10	1.21	23.57	0.22	1.40	2.67	0.49	0.65
$\boldsymbol{M}_{\rho_3}$	0.63	11.63	0.09	0.94	14.11	0.22	1.13	2.50	0.47	0.63
$\boldsymbol{M}_{\rho_4}$	0.61	3.52	0.10	0.40	4.29	0.22	0.55	1.82	0.48	0.53
$\boldsymbol{M}_{\rho_5}$	0.32	1.35	0.12	0.20	1.64	0.27	0.39	2.67	0.59	0.65
$\boldsymbol{M}_{\rho_6}$	0.39	1.57	0.11	0.29	2.36	0.25	0.51	3.07	0.55	0.69
$\boldsymbol{M}_{\rho_7}$	0.36	6.09	0.11	0.81	10.79	0.26	1.20	7.50	0.57	1.09
$\boldsymbol{M}_{\rho_8}$	0.48	2.67	0.10	0.51	11.00	0.24	1.23	6.67	0.52	1.02
$\boldsymbol{M}_{\rho_9}$	0.41	8.08	0.11	1.05	30.44	0.25	2.15	28.33	0.55	2.11
$\boldsymbol{M}_{\rho_{10}}$	0.56	1.85	0.10	0.30	2.33	0.22	0.43	1.85	0.49	0.54
$\boldsymbol{M}_{\rho_{11}}$	0.53	1.48	0.10	0.27	1.80	0.23	0.39	1.90	0.50	0.55
$\boldsymbol{M}_{\rho_{12}}$	0.50	1.05	0.10	0.23	1.19	0.23	0.33	1.87	0.51	0.54
$\boldsymbol{M}_{\rho_{13}}$	0.25	1.99	0.13	0.09	1.11	0.29	0.26	2.57	0.63	0.64
$\boldsymbol{M}_{\rho_{14}}$	0.30	5.52	0.12	−0.17	1.81	0.27	0.11	2.59	0.60	0.64
$\boldsymbol{M}_{\rho_{15}}$	0.52	4.15	0.10	−0.20	2.13	0.23	−0.01	2.03	0.50	0.57
$\boldsymbol{M}_{\rho_{16}}$	0.56	3.27	0.10	−0.09	1.93	0.22	0.05	1.85	0.49	0.54

Table 2.11: Case A with $m = 5$: power achieved with $n = 100$ and sample size required to achieve power of 0.80

Correlation Structure	Power				n			
	Bon.	T^2	OLS	GLS	Bon.	T^2	OLS	GLS
$\boldsymbol{M}_{\rho_1}$	0.59	1.000	0.11	1.000	149	2	1620	1
$\boldsymbol{M}_{\rho_2}$	0.59	1.000	0.11	1.000	149	8	1640	11
$\boldsymbol{M}_{\rho_3}$	0.59	1.000	0.10	1.000	149	14	1760	18
$\boldsymbol{M}_{\rho_4}$	0.59	0.99	0.10	0.81	149	46	1720	98
$\boldsymbol{M}_{\rho_5}$	0.59	0.71	0.13	0.30	148	121	1140	384
$\boldsymbol{M}_{\rho_6}$	0.59	0.78	0.12	0.53	148	104	1280	188
$\boldsymbol{M}_{\rho_7}$	0.59	1.000	0.13	1.000	148	27	1220	24
$\boldsymbol{M}_{\rho_8}$	0.59	0.96	0.11	0.95	148	61	1460	60
$\boldsymbol{M}_{\rho_9}$	0.59	1.000	0.12	1.000	148	20	1320	14
$\boldsymbol{M}_{\rho_{10}}$	0.59	0.85	0.11	0.55	149	89	1620	180
$\boldsymbol{M}_{\rho_{11}}$	0.59	0.75	0.11	0.49	148	111	1560	210
$\boldsymbol{M}_{\rho_{12}}$	0.59	0.58	0.11	0.36	148	155	1500	306
$\boldsymbol{M}_{\rho_{13}}$	0.59	0.88	0.14	0.10	148	82	1000	1840
$\boldsymbol{M}_{\rho_{14}}$	0.59	1.000	0.13	0.22	149	30	1100	544
$\boldsymbol{M}_{\rho_{15}}$	0.59	0.998	0.11	0.29	149	39	1540	403
$\boldsymbol{M}_{\rho_{16}}$	0.59	0.99	0.11	0.09	149	50	1620	2080

Table 2.12: Case B with $m = 5$: power achieved with $n = 100$ and sample size required to achieve power of 0.80

Correlation	Power				n			
Structure	Bon.	T^2	OLS	GLS	Bon.	T^2	OLS	GLS
$\boldsymbol{M}_{\rho_1}$	0.63	1.000	0.36	1.000	139	1	311	1
$\boldsymbol{M}_{\rho_2}$	0.63	1.000	0.35	1.000	141	7	315	8
$\boldsymbol{M}_{\rho_3}$	0.62	1.000	0.33	1.000	141	12	338	12
$\boldsymbol{M}_{\rho_4}$	0.62	0.998	0.34	0.97	141	38	330	52
$\boldsymbol{M}_{\rho_5}$	0.64	0.80	0.47	0.79	138	99	219	104
$\boldsymbol{M}_{\rho_6}$	0.64	0.93	0.43	0.95	138	69	245	60
$\boldsymbol{M}_{\rho_7}$	0.64	1.000	0.45	1.000	138	15	234	11
$\boldsymbol{M}_{\rho_8}$	0.64	1.000	0.39	1.000	139	15	280	10
$\boldsymbol{M}_{\rho_9}$	0.64	1.000	0.42	1.000	138	5	253	3
$\boldsymbol{M}_{\rho_{10}}$	0.63	0.93	0.36	0.86	140	70	311	85
$\boldsymbol{M}_{\rho_{11}}$	0.63	0.84	0.37	0.80	140	91	299	101
$\boldsymbol{M}_{\rho_{12}}$	0.63	0.64	0.38	0.64	140	137	288	146
$\boldsymbol{M}_{\rho_{13}}$	0.62	0.61	0.53	0.44	144	148	192	239
$\boldsymbol{M}_{\rho_{14}}$	0.61	0.84	0.49	0.12	146	90	211	1310
$\boldsymbol{M}_{\rho_{15}}$	0.60	0.90	0.37	0.05	148	77	295	22300
$\boldsymbol{M}_{\rho_{16}}$	0.60	0.87	0.36	0.06	148	85	311	6200

Table 2.13: Case C with $m = 5$: power achieved with $n = 100$ and sample size required to achieve power of 0.80

Correlation	Power				n			
Structure	Bon.	T^2	OLS	GLS	Bon.	T^2	OLS	GLS
$\boldsymbol{M}_{\rho_1}$	0.88	1.000	0.94	1.000	82	5	65	3
$\boldsymbol{M}_{\rho_2}$	0.87	0.96	0.93	0.996	84	61	66	37
$\boldsymbol{M}_{\rho_3}$	0.85	0.95	0.92	0.993	89	65	70	40
$\boldsymbol{M}_{\rho_4}$	0.86	0.85	0.92	0.97	86	90	69	55
$\boldsymbol{M}_{\rho_5}$	0.94	0.96	0.99	0.996	66	61	46	37
$\boldsymbol{M}_{\rho_6}$	0.93	0.98	0.97	0.998	70	53	51	33
$\boldsymbol{M}_{\rho_7}$	0.94	1.000	0.98	1.000	67	22	49	13
$\boldsymbol{M}_{\rho_8}$	0.90	1.000	0.96	1.000	77	25	58	15
$\boldsymbol{M}_{\rho_9}$	0.93	1.000	0.97	1.000	70	6	53	4
$\boldsymbol{M}_{\rho_{10}}$	0.88	0.85	0.94	0.97	82	89	65	54
$\boldsymbol{M}_{\rho_{11}}$	0.89	0.87	0.94	0.97	80	86	62	52
$\boldsymbol{M}_{\rho_{12}}$	0.90	0.86	0.95	0.97	77	87	60	53
$\boldsymbol{M}_{\rho_{13}}$	0.95	0.95	0.993	0.994	63	64	40	39
$\boldsymbol{M}_{\rho_{14}}$	0.94	0.96	0.988	0.995	66	63	44	39
$\boldsymbol{M}_{\rho_{15}}$	0.89	0.89	0.95	0.98	79	80	62	49
$\boldsymbol{M}_{\rho_{16}}$	0.88	0.85	0.94	0.97	82	89	65	54

Despite the limited scope of these examples, the results presented in this section have highlighted some differences between the properties of GLS and OLS. In Cases A and B, if the weight GLS assigns to the most effective outcome is more than that OLS assigns, the former has a larger mean and is more likely to have a larger effect size. (Of course, the effect size also depend on the standard deviation but the examples considered reveal that the standard deviations of GLS and OLS do not differ much.) This can result either when the most effective outcome is nearly uncorrelated with the others or when one or more of the less effective outcomes are nearly redundant as GLS weights less correlated outcomes relatively more heavily. In Case C, because the means of GLS and OLS are identical, the procedure having a smaller standard deviation has a larger effect size and hence larger power. In this case, both GLS and OLS perform well and they are competitive in power in most situations; however, a mild advantage in power of GLS can result in substantial differences in the required sample sizes.

Bonferroni Adjustment

The effect of the correlation structures on the performance of Bonferroni adjustment for five outcome measures is similar to that for three outcome measures. Table 2.11 show that in Case A, these different correlation structures seem to have essentially no impact on its performance. This agrees with the results for three outcome measures in Tables 2.4 and 2.5.

Table 2.12 indicates that in Case B, the different structures have a small impact on the performance of Bonferroni adjustment. When the more effective outcomes are relatively weakly correlated, this procedure performs slightly better; $\boldsymbol{M}_{\rho_5}$, $\boldsymbol{M}_{\rho_6}$, $\boldsymbol{M}_{\rho_7}$, $\boldsymbol{M}_{\rho_9}$, $\boldsymbol{M}_{\rho_{13}}$, and $\boldsymbol{M}_{\rho_{14}}$, are examples of this kind.

In Case C, the procedure based on Bonferroni adjustment performs better when the

outcomes are generally more weakly correlated. For instance, with the correlation structures $\boldsymbol{M}_{\rho_5}$, $\boldsymbol{M}_{\rho_6}$, $\boldsymbol{M}_{\rho_7}$, $\boldsymbol{M}_{\rho_8}$, or $\boldsymbol{M}_{\rho_9}$, the outcomes are mostly only mildly correlated, whereas with structures like $\boldsymbol{M}_{\rho_1}$, $\boldsymbol{M}_{\rho_2}$, $\boldsymbol{M}_{\rho_3}$, $\boldsymbol{M}_{\rho_4}$, $\boldsymbol{M}_{\rho_{10}}$, $\boldsymbol{M}_{\rho_{11}}$, $\boldsymbol{M}_{\rho_{12}}$, or $\boldsymbol{M}_{\rho_{16}}$, there is a greater degree of correlation among the outcomes. Bonferroni adjustment performs more powerfully in the former situation. The results in Table 2.13 suggests that in Case C, the required sample size is roughly proportional to $\overline{\rho}$, but this effect is moderate for the case of $m = 5$.

Hotelling's T^2

The effects of the correlation structures on the procedure based on T^2 for five outcome measures are complicated. The magnitude of $\tilde{\lambda}^2$ for each case is displayed in Table 2.10. Taking $\boldsymbol{M}_{\rho_1}$ as an example, we first display $\boldsymbol{M}_{\rho_1}^{-1}$:

$$\boldsymbol{M}_{\rho_1}^{-1} = \begin{pmatrix} 88.5 & 75.0 & 75.0 & -63.5 & -122.5 \\ 75.0 & 66.0 & 64.0 & -55.0 & -105.0 \\ 75.0 & 64.0 & 66.0 & -55.0 & -105.0 \\ -63.5 & -55.0 & -55.0 & 48.5 & 87.5 \\ -122.5 & -105.0 & -105.0 & 87.5 & 172.5 \end{pmatrix}.$$

With this correlation structure, T^2 is more powerful in Case A than Case C as 88.5 (m^{11}) is larger than 32.5 (sum). The elements close to m^{11} are quite large and all the large negative elements are relatively far away from m^{11} leading to a large weighted sum (138.68). As a consequence, T^2 performs even better in Case B than in Case A.

As another example,

$$\boldsymbol{M}_{\rho_{13}}^{-1} = \begin{pmatrix} 1.99 & -1.34 & -0.09 & -0.09 & -0.09 \\ -1.34 & 1.99 & -0.09 & -0.09 & -0.09 \\ -0.09 & -0.09 & 1.10 & -0.15 & -0.15 \\ -0.09 & -0.09 & -0.15 & 1.10 & -0.15 \\ -0.09 & -0.09 & -0.15 & -0.15 & 1.10 \end{pmatrix}.$$

The matrix has a sum of 2.57 and a weighted sum of 1.11. The element m^{11}, the weighted sum and the sum are considerably smaller than those of $\boldsymbol{M}_{\rho_1}^{-1}$. For this structure, T^2 is less powerful than it was for $\boldsymbol{M}_{\rho_1}$ for each of Cases A, B, and C. The above examples and Table 2.7 show that the correlation structure can have a great impact on the performance of T^2.

Overall Comparison

Now, we want to bring all the procedures together for an overall comparison. In Case A where only the first outcome is effective in comparing the two arms, GLS has the potential to have a very clear advantage over OLS. The structures $\boldsymbol{M}_{\rho_1}, \boldsymbol{M}_{\rho_2}, \boldsymbol{M}_{\rho_3}, \boldsymbol{M}_{\rho_7}, \boldsymbol{M}_{\rho_8}$, and $\boldsymbol{M}_{\rho_9}$ are examples of patterns of correlations for which GLS is very powerful. In these examples, Hotelling's T^2 is comparable to GLS and sometimes Hotelling's T^2 performs slightly better than GLS. On the other hand, Bonferroni adjustment is not comparable to GLS and T^2 although it has a clear advantage over OLS. Still in Case A, for structures like $\boldsymbol{M}_{\rho_5}$ $\boldsymbol{M}_{\rho_6}$, $\boldsymbol{M}_{\rho_{10}}$, $\boldsymbol{M}_{\rho_{11}}$ and $\boldsymbol{M}_{\rho_{12}}$, where GLS does not perform well although it still has a clear advantage over OLS, T^2 is more powerful. $\boldsymbol{M}_{\rho_{13}}$ and $\boldsymbol{M}_{\rho_{16}}$ are patterns of correlations where in Case A, OLS performs poorly but better than GLS. In these situations, T^2 is most powerful and Bonferroni adjustment also performs considerably better than OLS although it is not comparable to T^2.

In Case C where all outcomes are effective, GLS has a clear advantage over the other procedures in all the 16 patterns of correlations considered. The performance of GLS is substantially better than Bonferroni adjustment. The advantage of GLS over T^2 in required sample size is quite clear except for the structures $\boldsymbol{M}_{\rho_1}$ and $\boldsymbol{M}_{\rho_9}$ in which the latter also requires only a small sample.

In Case B where outcomes are of diminishing effectiveness, the results in Tables 2.12 indicate that the correlation structures have a great impact on the performance of both GLS and T^2 . Tables 2.11 and 2.12 show that for structures where T^2 is comparable to GLS in Case A, T^2 is also comparable to GLS in Case B. Furthermore, under the situations where T^2 is more powerful than GLS in Case A, T^2 is also more powerful than GLS in Case B.

2.6 Discussion

Although the comparisons presented in this chapter are limited, we can still draw a few general conclusions. First, for the special case of equally correlated outcomes, the inclusion of ineffective outcome measures leads to detrimental effects on all the procedures (Case A). In this situation, if only one outcome measure is effective, identifying this single effective outcome becomes essential. However, if it is not clear which outcome measure is effective, the results suggest the use of the procedure based on Bonferroni adjustment as the impact of the inclusion of ineffective outcomes on this procedure is smallest.

When several equally correlated outcome variables with roughly equal effectiveness are included, procedures which combine the evidence provided by individual outcomes can be quite powerful in assessing the relative efficacy of the two arms. The results in Case C suggest that O'Brien's OLS and GLS procedures are the best way to proceed in this situation.

For outcome measures with diminishing effectiveness (Case B), the situation is more complicated. For equally correlated outcomes, O'Brien's OLS procedure is more powerful than the other procedures only when the common correlation is very small. With a modest common correlation, say 0.3, Bonferroni adjustment seems to be the best way to proceed. When the common correlation is moderate, say 0.5, Bonferroni adjustment performs better than Hotelling's T^2 only when a small number of outcomes are included.

For unequally correlated outcome measures, O'Brien's GLS and Hotelling's T^2 statistic demonstrate their potential to overcome the possible detrimental effects resulting from the inclusion of ineffective outcome measures. When the effective outcome is weakly correlated with the ineffective outcomes and the ineffective outcomes are intercorrelated, the resulting down-weighting of these ineffective outcomes in the GLS statistic (relative to the OLS statistic) results in enhanced sensitivity of the assessment of the relative efficacy of the two arms. As the approach based on Hotelling's T^2 does not take account of the direction of the differences between the two arms on the individual outcomes, when T^2 and O'Brien's GLS are comparable, the latter should be preferred. Generally speaking, the examples we have considered suggest that when the outcome measures with greater effectiveness are not highly correlated and the outcomes with less effectiveness are not weakly correlated, T^2 and O'Brien's GLS are competitive with each other.

However, the danger of using O'Brien's GLS procedure is that depending upon the correlation structure among the outcome measures, it is possible for GLS to perform very well or very poorly. Throughout our work, we have assumed that the correlation structure is known. This would typically not be the case when planning clinical trials. The ideal situation would be that the effectiveness of the individual outcome measures selected for inclusion is clear and high quality information on the relationship among the outcome measures to be used is available. In this case, the appropriateness of using the O'Brien's GLS or OLS procedure can be assessed. This will not be possible if the information on the

pattern of the correlation among the outcome measures is of low quality. There might be a situation when it is not clear which outcome measures are effective and therefore several outcomes need to be included, and the information about the underlying correlation structure among the outcomes is very limited. In such a situation, the use of the GLS procedure could be risky. To avoid that, Bonferroni adjustment is a reasonable way to proceed as our results indicate that the impact of correlation structures on this procedure is small.

Chapter 3

Disjunctive Composite Outcome Measures

We now consider another type of composite outcome measure called a "disjunctive" outcome measure. The low dose oral methotrexate clinical trial in chronic progressive MS, the results of which are presented in Goodkin et al. (1992), is one example of a MS clinical trial which used this type of composite outcome measure in its design and analysis. The idea of this method of combining multiple outcome variables is as follows: The researcher first dichotomizes each outcome measure; a measurement on the jth outcome variable exceeding a pre-assigned cutoff value is taken to indicate a significant clinical change on this particular outcome. An indication of significant clinical change on **any** of the m outcome measures is taken to indicate a treatment failure. In other words, the responses on the original individual outcome measures are first converted to binary responses, and the information on the binary responses is then combined into an overall binary response. To assess the effect of the treatment relative to the placebo, the proportions of treatment failure on the two arms are compared.

As all the evidence from the individual outcome measures is summarized into a single response, the simplicity of this method makes it attractive to some researchers. However, there are some potential difficulties with this method. To construct meaningful pre-assigned cutoff values for individual outcome measures requires substantial knowledge on these outcomes. Additionally, the best rule for combining the binary responses from the individual outcome measures into a composite outcome measure is not obvious. Here these binary responses are combined disjunctively, but they could be combined in other

ways. For instance, the most strict way would be that an indication of significant clinical change on **all** of the m outcome measures is required to indicate a treatment failure.

In this chapter, we first examine the properties of dichotomized tests in the univariate setting and compare such tests to those based upon a continuous variable. Second, we investigate the statistical properties of disjunctive outcome measures. Finally, we compare this method to the procedures discussed in Chapter 2.

3.1 Dichotomized Tests for One Outcome Variable

This section is devoted to the examination of dichotomized tests on a single continuous outcome variable. For consistency of notation with Chapter 2, let X_{i1k} represent this particular outcome variable for the kth patient in treatment group i, where $i = 1$ for the placebo arm and $i = 2$ for the treatment arm. We will assume that X_{11k} are independently and identically distributed with the distribution function F_1 which has mean μ_{11} and known variance σ_1^2; similarly, X_{21k} are i.i.d as F_2 with mean μ_{21} and the same variance σ_1^2. Further, we will assume that F_1 and F_2 belong to the same location shift family; that is,

$$F_1(x) = F_2(x - \delta_1),$$

where $\delta_1 = \mu_{11} - \mu_{21}$. In other words, the difference between the two population distributions can be expressed by a shift in location. Let F denote the *standard* cdf for this family which has mean 0 and variance 1. Then $F_1(x)$ and $F_2(x)$ can be expressed in terms of F:

$$F_1 = F(\frac{x - \mu_{11}}{\sigma_1}),$$

and

$$F_2 = F(\frac{x - \mu_{12}}{\sigma_1}).$$

Let η_1 represent the pre-assigned cutoff point for this outcome measure; a patient has a significant clinical change on this outcome if his or her response is greater than η_1. We will express η_1 as the sum of the underlying mean of this outcome on the placebo arm, μ_{11}, and the standardized distance between μ_{11} and η_1; that is,

$$\eta_1 = \mu_{11} + c_1\sigma_1,$$

where $c_1 \geq 0$ which means η_1 is greater than the placebo mean.

If π_i denote the probability that a patient has a significant clinical change on the ith treatment arm, then π_i can be expressed as:

$$\begin{aligned}
\pi_i &= P(\text{a patient on the } i\text{th treatment arm has a significant clinical change}) \\
&= 1 - P(X_{i1k} \leq \eta_1) \\
&= 1 - P(X_{i1k} \leq \mu_{11} + c_1\sigma_1) \\
&= 1 - P\left(\tfrac{X_{i1k}-\mu_{i1}}{\sigma_1} \leq c_1 + \tfrac{\mu_{11}-\mu_{i1}}{\sigma_1}\right).
\end{aligned}$$

In terms of F, we can express π_1 and π_2 as:

$$\pi_1 = 1 - F(c_1) \tag{3.1}$$

and

$$\pi_2 = 1 - F(c_1 + \Delta_1), \tag{3.2}$$

where $\Delta_1 = \frac{\mu_{11}-\mu_{21}}{\sigma_1}$ is the standardized difference of the underlying population means of this continuous outcome variable.

The difference between π_1 and π_2 depends upon the cutoff value, c_1, and the standardized difference between the population means of the two arms. If we compare the two arms by a dichotomized test, we will test $H_{0d} : \pi_1 = \pi_2$ against $H_{ad} : \pi_1 \neq \pi_2$. On the other hand, if we compare the two arms by the Z-test on the sample means of the continuous outcome variable, we will test $H_{0c} : \mu_{11} = \mu_{12}$ against $H_{ac} : \mu_{11} \neq \mu_{12}$. As π_1 and

π_2 are equal if and only if μ_{11} and μ_{12} are equal, the null hypotheses, $H_{0d} : \pi_1 = \pi_2$ and $H_{0c} : \mu_{11} = \mu_{12}$ are equivalent and a meaningful comparison between the dichotomized test and the Z-test can be made. Moreover, for every specified c_1, the difference to be detected between π_1 and π_2 is determined by Δ_1, the standardized difference to be detected between the underlying means. Hence, once H_{ac} is specified, the corresponding alternative hypothesis, H_{ad}, is specified as well. In particular, the alternative corresponding to $\Delta_1 = \Delta$ is $\pi_1 - \pi_2 = \theta$, where $\theta = F(c_1 + \Delta) - F(c_1)$. Under such situations where the hypotheses to be tested by the two statistics are equivalent, a meaningful comparison between the dichotomized test and the Z-test can be made.

3.1.1 How Much Is Lost by Dichotomizing?

It is clear that use of the dichotomized outcome variable involves a certain degree of loss of information due to the transformation of the continuous variable to the binary response variable. The issue now becomes how much information is lost. We will try to address this issue in this section through a comparison between the dichotomized test and the Z-test using two criteria: percent power loss and asymptotic relative efficiency.

Criterion 1: Percent Power Loss

Percent power loss is defined as the difference between the powers achieved by the two tests for equivalent alternative hypothesis, expressed as a percentage of the power achieved by the Z-test; that is,

$$\text{Percent power loss} = \frac{P_c - P_d}{P_c} 100\%,$$

where P_c and P_d denote the power achieved by the Z-test and the dichotomized test respectively.

To evaluate the percent power loss, we need formulae for P_c and P_d. From Appendix B, the formula for the power of the Z-test on the continuous outcome variable evaluated at the specified alternative, $\Delta_1 = \Delta$, is:

$$P_c(\Delta) \approx 1 - \Phi\left(z_{1-\alpha/2} - \sqrt{\frac{n}{2}}\Delta\right) + \Phi\left(-z_{1-\alpha/2} - \sqrt{\frac{n}{2}}\Delta\right).$$

This approximate formula is derived based upon the Central Limit Theorem. Provided that n is reasonably large, this approximation should be adequate. From Appendix D which presents the general formulae for the power and the sample size required per arm for a level α test for comparing two population proportions, we have:

$$\begin{aligned} P_d(\theta) \approx\ & 1 - \Phi\left(z_{1-\alpha/2}\sqrt{\frac{2\overline{\pi}(1-\overline{\pi})}{\pi_1(1-\pi_1)+\pi_2(1-\pi_2)}} - \frac{\sqrt{n}\theta}{\sqrt{\pi_1(1-\pi_1)+\pi_2(1-\pi_2)}}\right) \\ & + \Phi\left(-z_{1-\alpha/2}\sqrt{\frac{2\overline{\pi}(1-\overline{\pi})}{\pi_1(1-\pi_1)+\pi_2(1-\pi_2)}} - \frac{\sqrt{n}\theta}{\sqrt{\pi_1(1-\pi_1)+\pi_2(1-\pi_2)}}\right) \end{aligned}$$

where $\overline{\pi} = \frac{\pi_1+\pi_2}{2}$ and $\theta = \pi_1 - \pi_2$. As this approximate formula for the power of the dichotomized test relies on the normal approximation for binomial probabilities, it will not provide an accurate approximation when π_1 and π_2 are close to 0 or 1.

As in Chapter 2, we assume X_{i1k} follows the normal distribution and examine the property of percent power loss under this assumption. (Note that under this normality assumption, the formula for P_c is exact.) From (3.1) and (3.2), π_1 and π_2 can then be calculated as:

$$\pi_1 = 1 - \Phi(c_1), \tag{3.3}$$

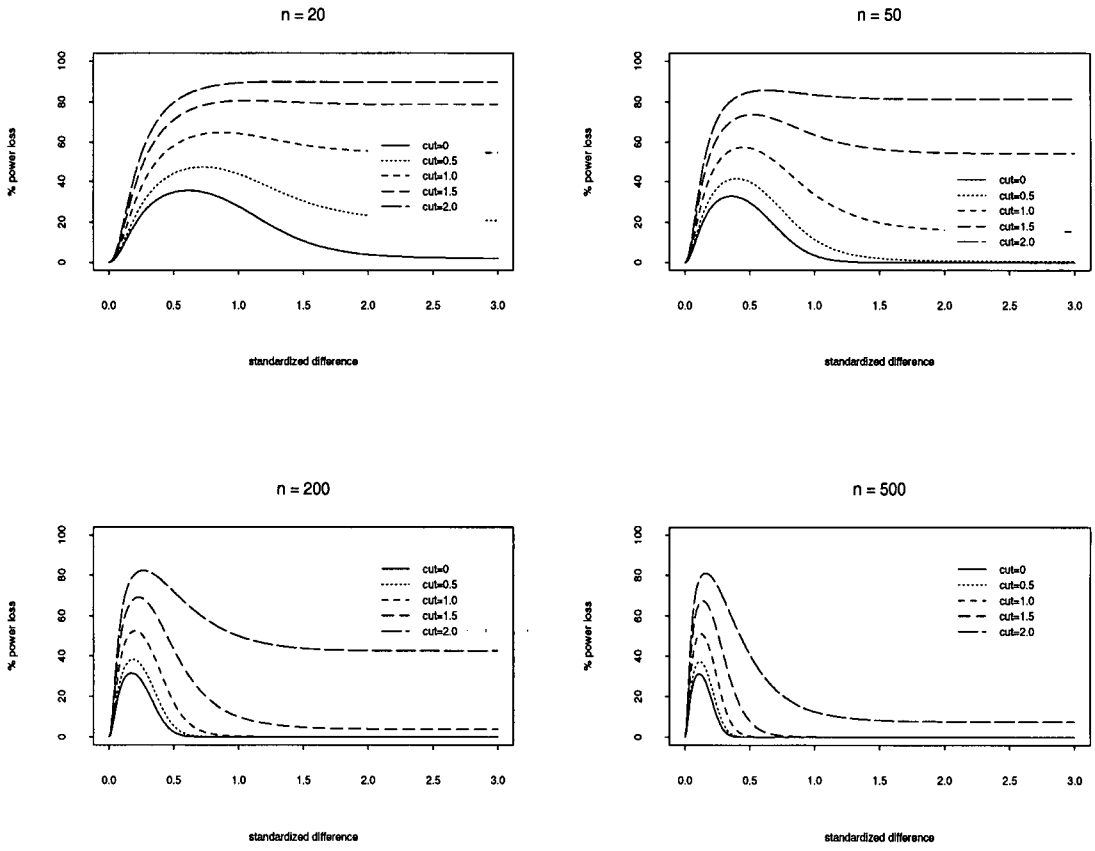
and

$$\pi_2 = 1 - \Phi(c_1 + \Delta_1). \tag{3.4}$$

Figure 3.1 presents the percent power loss as a function of Δ_1, the standardized difference between the population means, for a few specified cutoff points for each of four different sample sizes when F is taken to be normal distribution. Comparing the five specified values of the cutoff point, regardless of the sample size, using a cutpoint of $\eta_1 = \mu_{11}$ or $c_1 = 0$ (i.e. dichotomizing at the placebo mean) provides the minimal percent loss of power for every fixed value of Δ, the standardized difference. Moreover, for every fixed value of Δ, as the pre-assigned cutoff point gets larger, the percent power loss increases. If the continuous variable is dichotomized at the placebo mean (i.e. $\eta_1 = \mu_{11}$), the percent power loss never exceeds 30%, no matter what the value of the standardized difference.

Figure 3.1 also shows that for each specific cutoff point, the value of the standardized difference at which the percent power loss achieves its maxima changes with the sample size. With the sample sizes of 50, 200 and 500, the value ranges from 0.2 to 0.5, increasing only slightly as the pre-assigned cutoff point gets larger. With $n = 20$, this value lies beyond 0.5. The figure illustrates the dramatic impact of sample size on the relationships among the percent power loss, the cutoff point and the standardized difference. When we have very large samples, say $n = 500$, with the cutoff point of 1.5 or smaller, the percent power loss decreases very quickly from its maxima to 0 as the standardized difference gets moderately large. In this case, the power for both the dichotomized test and the Z-test approaches 1 very quickly. However, when the cutoff point is too large, the power of the dichotomized test can never approach 1 even with a large standardized difference. For example, with a sample size of 500, when $c_1 = 2$, the percent power loss does not approach 0 as Δ increases. On the contrary, for small samples such as $n = 20$, Figure 3.1 shows that a small difference in the cutoff point can result in a substantial difference in the percent power loss. Also, when the cutoff point is 0.5 or larger and Δ_1 is reasonably large, the percent power loss for small samples is substantially larger than

Figure 3.1: Percent power loss for different values of c_1 and sample sizes



for large samples.

From (3.3) and (3.4), it is clear that for any positive values of Δ_1, once c_1 is beyond about 2, π_2 is close to 0. Similarly, as the cutoff point gets large no matter where the standardized difference lies, both π_1 and π_2 approach 0. In either case, the use of normal approximation is no longer valid; it is used here only for illustration purposes.

Criterion 2: Asymptotic Relative Efficiency

As Percent Power Loss, for a fixed sample size, depends upon both where the continuous outcome variable is dichotomized and where the alternative lies, it does not provide a general comparison between the dichotomized test and the Z-test when the alternative hypothesis is composite as in our case. One criterion often used for comparing two test statistics which overcomes this disadvantage is the *asymptotic relative efficiency* (ARE), also often called *the Pitman efficiency.* Suppose we want to compare two tests, A and B, having the same level. An obvious comparison would be of the sample sizes required to achieve the same power at a specified alternative. The idea of the ARE of test A relative to test B is to examine the limiting behaviour of the ratio n_A/n_B of these required sample sizes, as the specified alternative approaches the null hypothesis.

We first provide the theoretical basis for calculating the ARE. Suppose that we have two test statistics, T_n and T_n^*, for samples of size n and the parameter of interest is ν. Both tests are used to test $H_0 : \nu \in \Omega_0$ versus $H_a : \nu \in \Omega - \Omega_0$. Further, suppose that a subset of the space Ω can be indexed in terms of a sequence of parameters $\{\nu_0, \nu_1, \cdots, \nu_n, \cdots\}$ such that ν_0 specifies a value in Ω_0 and the remaining $\nu_1, \nu_2, \cdots$ are in $\Omega - \Omega_0$ and that $\lim_{n\to\infty} \nu_n = \nu_0$. Under these conditions, we can give a formal definition of the ARE of T relative to T^* (Gibbons, 1971):

Definition 3.1 *Let T_n and T_n^* be two sequences of test statistics, all with the same significance level α. Let $\{n_i\}$ and $\{n_i^*\}$ be two monotonic increasing sequences of positive integers such that*

$$\lim_{i\to\infty} Power(T_{n_i} \mid \nu = \nu_i) = \lim_{i\to\infty} Power(T^*_{n_i^*} \mid \nu = \nu_i) = \gamma$$

where γ is not equal to 0 or 1. Then the asymptotic relative efficiency of test T relative

to test T^* *is*

$$ARE(T, T^*) = \lim_{i \to \infty} \frac{n_i^*}{n_i}$$

provided this limit exists and is constant for all sequences of $\{n_i\}$ *and* $\{n_i^*\}$.

To calculate the ARE directly from its definition is complicated. The calculation of the ARE can be simplified if the following regularity assumptions are satisfied by the sequences of test statistics T_n, and analogously for T_n^* ($E(T_n)$ and $\sigma(T_n)$ denote the expectation and standard deviation of the test statistic T_n):

1. $dE(T_n)/d\nu$ exists and is nonzero for $\nu = \nu_0$, and is continuous at ν_0.

2. There exists a positive constant c such that

$$\lim_{n \to \infty} \frac{dE(T_n)/d\nu|_{\nu=\nu_0}}{\sqrt{n}\sigma(T_n)|_{\nu=\nu_0}} = c.$$

3. There exists a sequence of alternatives $\{\nu_n\}$ such that for some constant $d > 0$, we have

$$\nu_n = \nu_0 + \frac{d}{\sqrt{n}}$$

$$\lim_{n \to \infty} \frac{dE(T_n)/d\nu|_{\nu=\nu_n}}{dE(T_n)/d\nu|_{\nu=\nu_0}} = 1$$

$$\lim_{n \to \infty} \frac{\sigma(T_n)\,|_{\nu=\nu_n}}{\sigma(T_n)\,|_{\nu=\nu_0}} = 1.$$

4.

$$\lim_{n \to \infty} P\left[\frac{T_n - E(T_n)\,|_{\nu=\nu_n}}{\sigma(T_n)\,|_{\nu=\nu_n}} \leq z \,\middle|\, \nu = \nu_n\right] = \Phi(z)$$

Theorem 3.1 *Under these four regularity conditions, the limiting power of the test* T_n *is*

$$\lim_{n \to \infty} Power(T_n \mid \nu = \nu_n) = 1 - \Phi(z_{1-\alpha} - dc)$$

Theorem 3.2 *If T_n and T_n^* are two tests satisfying these four regularity conditions, the ARE of T relative to T^* is*

$$ARE(T, T^*) = \lim_{n\to\infty} \frac{e(T_n)}{e(T_n^*)},$$

where $e(T_n)$ is called the efficacy of the test statistic T_n when used to test the hypothesis $\nu = \nu_0$ and

$$e(T_n) = \frac{[dE(T_n)/d\nu]^2}{\sigma^2(T_n)} \Big|_{\nu=\nu_0} .$$

Theorem 3.3 *The statement in Theorem 3.2 remains valid as stated if both tests are two-sided, with rejection region*

$$T_n \in R \qquad for\ t_n \geq t_{n,1-\alpha_1} \qquad or \qquad t_n \leq t_{n,1-\alpha_2}$$

where the size is still α, and a corresponding rejection region is defined for T_n^* with the same α_1 and α_2. Then the alternative is also two-sided, as $H_a : \nu \neq \nu_0$.

We now use the above theorems to calculate the ARE of the dichotomized test for comparing two population proportions relative to the Z-test for comparing two population means with known variance. Suppose we have the distribution model

$$F_1(x) = F_2(x - \nu)$$

and the null hypothesis is $H_0 : \nu = 0$. With samples of size n, the corresponding Z-statistic for populations with a common known variance σ_1^2 is

$$\begin{aligned} T_n^* &= \frac{\overline{X}_{11} - \overline{X}_{21}}{\sigma_1\sqrt{2/n}} \\ &= \sqrt{\frac{n}{2}}\left(\frac{\overline{X}_{11} - \overline{X}_{21} - \nu}{\sigma_1} + \frac{\nu}{\sigma_1}\right). \end{aligned}$$

Since

$$E(T_n^*) = \sqrt{\frac{n}{2}}\frac{\nu}{\sigma_1},$$

$$\frac{d}{d\nu}E(T_n^*) \mid_{\nu=0} = \sqrt{\frac{n}{2}}\frac{1}{\sigma_1},$$

and

$$Var(T_n^*) \mid_{\nu=0} = 1,$$

the efficacy of this Z-test for any population within the location shift family is

$$e(T_n^*) = \frac{n}{2\sigma_1^2}. \tag{3.5}$$

For the dichotomized test, the test statistic is

$$T_n = \frac{p_1 - p_2}{\sqrt{2\overline{p}(1-\overline{p})/n}}$$

where $\overline{p} = \frac{p_1+p_2}{2}$. It can also be written as

$$T_n = \sqrt{\frac{n}{2}}\left[\frac{(p_1 - p_2)}{\sqrt{\overline{\pi}(1-\overline{\pi})}}\right]\sqrt{\frac{\overline{\pi}(1-\overline{\pi})}{\overline{p}(1-\overline{p})}}$$

To evaluate the efficacy of this test statistic, first note that as $n \to \infty, \sqrt{\frac{\overline{\pi}(1-\overline{\pi})}{\overline{p}(1-\overline{p})}} \to 1$ in probability. Therefore, T_n is asymptotically equivalent to

$$T_n' = \sqrt{\frac{n}{2}}\left[\frac{(p_1 - p_2)}{\sqrt{\overline{\pi}(1-\overline{\pi})}}\right]$$

and it suffices to calculate the expectation and variance of T_n'. But

$$E(T_n') = \sqrt{\frac{n}{2}}\frac{\pi_1 - \pi_2}{\sqrt{\overline{\pi}(1-\overline{\pi})}},$$

and the null variance ($\pi_1 - \pi_2 = 0$ under H_0) is

$$Var(T_n') \mid_{H_0} = \frac{n}{2} \frac{2\pi_2(1-\pi_2)/n}{\pi_2(1-\pi_2)} = 1.$$

Thus, to evaluate the efficacy of T_n, it remains to evaluate

$$\frac{d}{d\nu} E(T_n') \Big|_{H_0} = \sqrt{\frac{n}{2}} \frac{d}{d\nu} \left(\frac{\pi_1 - \pi_2}{\sqrt{\overline{\pi}(1-\overline{\pi})}} \right) \Bigg|_{H_0}.$$

We will first evaluate (keep in mind that under the null hypothesis, $\nu = 0$, and $\pi_1 = \pi_2 = \overline{\pi}$):

$$\begin{aligned}
\frac{d}{d\nu} \left(\frac{\pi_1 - \pi_2}{\sqrt{\overline{\pi}(1-\overline{\pi})}} \right) \mid_{H_0} &= \frac{\sqrt{\overline{\pi}(1-\overline{\pi})} \frac{d}{d\nu}(\pi_1 - \pi_2) - (\pi_1 - \pi_2) \frac{d}{d\nu} \left[\sqrt{\overline{\pi}(1-\overline{\pi})} \right]}{\overline{\pi}(1-\overline{\pi})} \mid_{H_0} \\
&= \frac{\sqrt{\overline{\pi}(1-\overline{\pi})} \frac{d}{d\nu}(\pi_1 - \pi_2)}{\overline{\pi}(1-\overline{\pi})} \mid_{H_0} \\
&= \frac{\frac{d}{d\nu}(\pi_1 - \pi_2)|_{H_0}}{\sqrt{\pi_1(1-\pi_1)}} \\
&= \frac{\frac{d}{d\nu}(F_1(\eta_1 + \nu) - F_1(\eta_1))|_{H_0}}{\sqrt{\pi_1(1-\pi_1)}} \\
&= \frac{f_1(\eta_1 + \nu) \mid_{H_0}}{\sqrt{\pi_1(1-\pi_1)}} \\
&= \frac{f_1(\eta_1)}{\sqrt{\pi_1(1-\pi_1)}} \\
&= \frac{f_1(\eta_1)}{\sqrt{F_1(\eta_1)[1-F_1(\eta_1)]}}.
\end{aligned}$$

Thus,

$$\frac{d}{d\nu} E(T_n') \mid_{H_0} = \sqrt{\frac{n}{2}} \frac{f_1(\eta_1)}{\sqrt{F_1(\eta_1)[1-F_1(\eta_1)]}},$$

and the efficacy of T_n' therefore is

$$e(T_n') = \frac{n}{2} \frac{[f_1(\eta_1)]^2}{F_1(\eta_1)[1-F_1(\eta_1)]}. \tag{3.6}$$

Result 3.1 *From (3.5) and (3.6), the ARE of the dichotomized test relative to the Z-test for location shift family F_1 with known variance σ_1^2 is*

$$\begin{aligned} ARE(T,T^*) &= \lim_{n\to\infty} \frac{e(T_n)}{e(T_n^*)} \\ &= \lim_{n\to\infty} \frac{e(T_n')}{e(T_n^*)} \\ &= \lim_{n\to\infty} \frac{n}{2} \frac{[f_1(\eta_1)]^2}{\pi_1(1-\pi_1)} \frac{2\sigma_1^2}{n} \\ &= \left(\frac{\sigma_1 f_1(\eta_1)}{F_1(\eta_1)[1-F_1(\eta_1)]} \right)^2. \end{aligned}$$

Applying this result to normal and logistic populations leads to:

Result 3.2

1. *If F_1 is taken to be normal,*

$$ARE(T,T^*) = \frac{[e^{-c_1^2/2}]^2}{2\pi[1-\Phi(c_1)]\Phi(c_1)}.$$

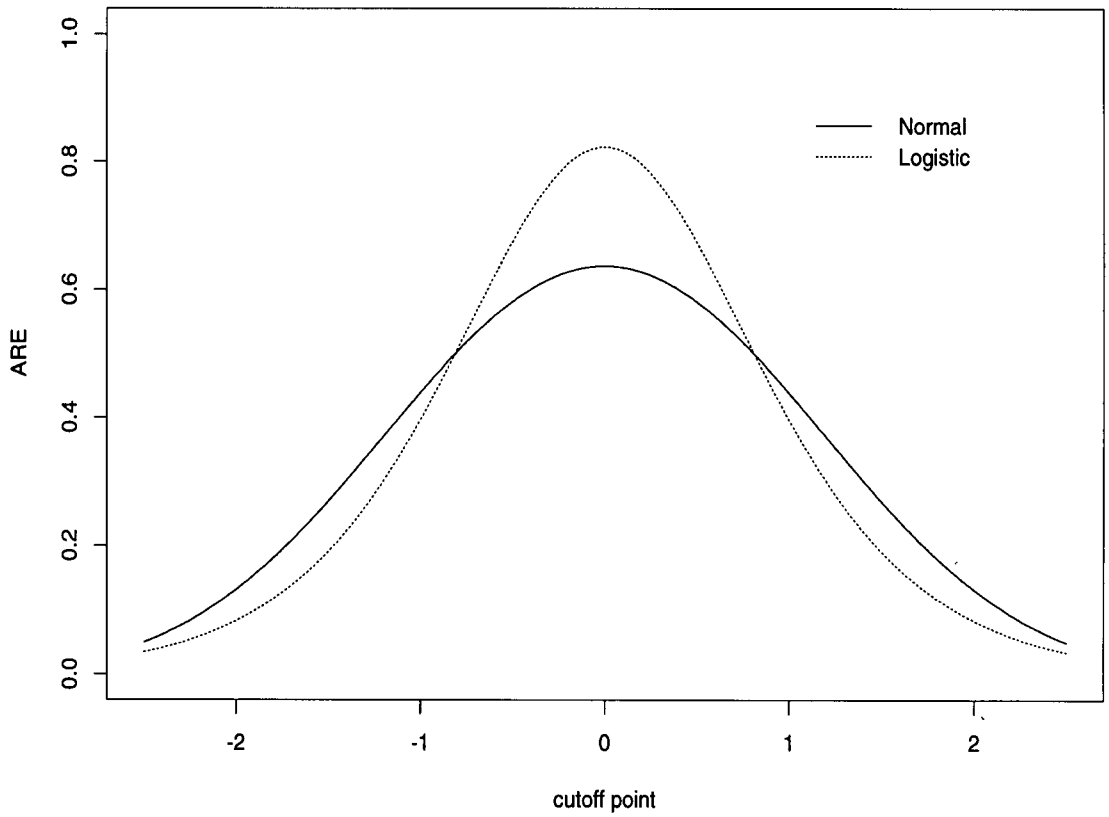
2. *If F_1 is taken to be logistic, where*

$$F_1(x \mid \mu_{11}, \sigma_1) = \frac{1}{1+e^{-\pi(x-\mu_{11})/(\sigma_1\sqrt{3})}},$$

$$ARE(T,T^*) = \frac{\pi^2}{3} \frac{e^{-\pi c_1/\sqrt{3}}}{[1+e^{-\pi c_1/\sqrt{3}}]^2}.$$

Note that the ARE does not depend upon the standardized difference; it is a function only of the cutoff point, c_1. Figure 3.2 plots the ARE of the dichotomized test relative to the Z-test for normal and logistic populations as a function of the cutoff point c_1. The ARE is symmetric about the cutoff point of 0 and decreases as the cutoff point moves away from 0; in particular, for both the normal and logistic populations, the ARE is maximized at the cutoff point of 0. As the cutoff point moves away from 0, the ARE

Figure 3.2: ARE of the dichotomous test relative to the Z-test



decreases more rapidly for logistic populations than for normal populations. The ARE has a maximum of $2/\pi \cong 0.64$ for normal populations while it has a maximum of $\pi^2/12 \cong 0.82$ for logistic populations. For both distributions, the ARE is about 0.5 when the cutoff point is close to 1 or -1; the ARE is only about 0.3 when c_1 is at 1.5 and the ARE is very small once c_1 moves above 2 or below -2.

Compared to the percent power loss, the ARE generalizes the comparison of the dichotomous test and the Z-test in the sense that it does not depend on the significance level and the standardized difference. However, the disadvantage of the ARE is that

because it is an asymptotic concept, it may not accurately reflect the relative sample sizes required to achieve the same power when the samples are finite and/or H_a is not approaching H_0. Nevertheless, the message from both criteria we have examined is clear: dichotomizing at the placebo mean is the best choice among the various values of c_1 we have examined for the dichotomous test on one outcome variable. With this choice, if the underlying distributions are normal, the percent power loss is ensured to be no more than about 30% and the ARE is about 60%.

3.2 Properties of Disjunctive Composite Outcome Measures

After some basic understanding of the statistical properties of dichotomized tests based on one outcome variable, we now turn to the examination of the disjunctive composite outcome measure. In this section, we will work under the assumption, as in the previous chapter, that the underlying data follows a multivariate normal distribution; that is to say, $\boldsymbol{X}_{ik}$ are independently distributed and each follows a multivariate normal distribution with mean vector $\boldsymbol{\mu}_i$ and known common variance-covariance matrix $\boldsymbol{\Sigma}$. It is difficult to provide a thorough investigation of the properties of this composite measure because there are many possibilities that could be considered. The objective here, as in Chapter 2, is to examine a few cases to highlight the main aspects of its statistical properties.

3.2.1 Power and Sample Size Calculations

We first give a formal definition of this composite measure and provide the formulae for its power and the required sample size. Let η_j represent the pre-assigned cutoff point for the jth outcome measure. A patient has a significant clinical change on the jth outcome if his or her response on the jth outcome is greater than η_j. Again, we will express η_j as the sum of the underlying mean on the jth outcome of the placebo arm, μ_{1j}, and the

standardized distance between μ_{1j} and η_j, $\eta_j = \mu_{1j} + c_j\sigma_j$. An indication of significant clinical change in **any** of these m outcome measures is taken to indicate a treatment failure. In other words, the information obtained from the individual outcome variables is summarized by a binary response, treatment failure or treatment success.

If π_i denotes the probability that a patient on the ith treatment arm has a treatment failure, then π_i can be expressed as:

$$\begin{aligned}
\pi_i &= P(\text{a patient on the } i\text{th treatment arm has a treatment failure}) \\
&= P(\text{a patient on the } i\text{th treatment arm has a significant change} \\
&\quad \text{on any of the } m \text{ outcomes}) \\
&= 1 - P(X_{ijk} \leq \eta_j, for \;\; j = 1, 2, \ldots, m) \\
&= 1 - P(X_{ijk} \leq \mu_{1j} + c_j\sigma_j, for \;\; j = 1, 2, \ldots, m) \\
&= 1 - \int_{-\infty}^{\mu_{1m}+c_m\sigma_m} \cdots \int_{-\infty}^{\mu_{11}+c_1\sigma_1} f_{\boldsymbol{X}_i}(\mathbf{x})\, dx_1 \cdots dx_m
\end{aligned}$$

where $f_{\boldsymbol{X}_i}$ denotes the pdf of $\boldsymbol{X}_{ik}$, the pdf of the multivariate normal distribution with mean vector $\boldsymbol{\mu}_i$ and variance-covariance matrix $\boldsymbol{\Sigma}$.

We can simplify the expression for π_i with the standardization of X_{ijk} by:

$$Z_j = \frac{X_{ijk} - \mu_{ij}}{\sigma_j}.$$

Marginally, Z_j follows the standard normal distribution. With $\boldsymbol{Z} = (Z_1, Z_2, \cdots, Z_m)'$, the joint distribution of $\boldsymbol{Z}$ is the multivariate normal distribution with $\mathbf{0}$ mean vector and correlation matrix $\boldsymbol{M}_\rho$. Now, π_i can be repressed as:

$$\pi_i \;\; = \;\; 1 - \int_{-\infty}^{c_m + \frac{\mu_{1m} - \mu_{im}}{\sigma_m}} \cdots \int_{-\infty}^{c_1 + \frac{\mu_{11} - \mu_{i1}}{\sigma_1}} f_{\boldsymbol{Z}}(\mathbf{z})\, dz_1 \cdots dz_m,$$

where $f_{\boldsymbol{Z}}(\mathbf{z})$ denotes the pdf of $\boldsymbol{Z}$. Thus,

$$\pi_1 = 1 - \int_{-\infty}^{c_m} \cdots \int_{-\infty}^{c_1} f_{\boldsymbol{Z}}(\mathbf{z})\, dz_1 \cdots dz_m, \tag{3.7}$$

$$\pi_2 = 1 - \int_{-\infty}^{c_m+\Delta_m} \cdots \int_{-\infty}^{c_1+\Delta_1} f_{\boldsymbol{Z}}(\mathbf{z})\, dz_1 \cdots dz_m. \tag{3.8}$$

For the special case of uncorrelated outcome measures, the expressions for π_1 and π_2 simplify to:

$$\pi_1 = 1 - \prod_{j=1}^{m} P(Z_j \leq c_j) = 1 - \prod_{j=1}^{m} \Phi(c_j) \tag{3.9}$$

$$\pi_2 = 1 - \prod_{j=1}^{m} P(Z_j \leq c_j + \Delta_j) = 1 - \prod_{j=1}^{m} \Phi(c_j + \Delta_j) \tag{3.10}$$

Now with all the information from the individual outcome measures being combined and summarized by this disjunctive outcome measure, comparison between the two arms reduces to a comparison of the two population proportions, π_1 and π_2. From Appendix D, the approximate formulae for the power and the required sample size for the test comparing two population proportions are:

$$\begin{aligned} Power_{\pi_1-\pi_2=\theta} \approx\ & 1 - \Phi\left(z_{1-\alpha/2}\sqrt{\frac{2\overline{\pi}(1-\overline{\pi})}{2\overline{\pi}(1-\overline{\pi}) - \theta^2/2}} - \frac{\sqrt{n}\theta}{\sqrt{2\overline{\pi}(1-\overline{\pi}) - \theta^2/2}} \right) \\ & + \Phi\left(-z_{1-\alpha/2}\sqrt{\frac{2\overline{\pi}(1-\overline{\pi})}{2\overline{\pi}(1-\overline{\pi}) - \theta^2/2}} - \frac{\sqrt{n}\theta}{\sqrt{2\overline{\pi}(1-\overline{\pi}) - \theta^2/2}} \right) \end{aligned}$$

$$n \approx \frac{\left(z_{1-\alpha/2}\sqrt{2\overline{\pi}(1-\overline{\pi})} - z_\beta \sqrt{2\overline{\pi}(1-\overline{\pi}) - \frac{1}{2}\theta^2} \right)^2}{\theta^2}$$

where $\overline{\pi} = \frac{\pi_1+\pi_2}{2}$.

In order to connect the power and sample size calculations for the disjunctive outcome measure to those for the procedures discussed in Chapter 2, the calculations must be made for equivalent hypotheses. Testing the equivalence of the mean vectors, $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$, is

the same as testing the equivalence of the population proportions, π_1 and π_2. As shown by (3.7) and (3.8), the difference between the population proportions depends upon the cutoff values, c_j, and the standardized differences, Δ_j, between the population means of the two arms. Therefore, for every specified set of c_j's, the difference to be detected between π_1 and π_2 is determined by the standardized differences between the underlying means. In other words, for every alternative hypothesis considered in Chapter 2, the corresponding alternative hypothesis for the population proportions, π_1 and π_2, can be determined. Therefore, we will consider the same configurations of the standardized differences in the underlying means, namely, Case A where only one outcome measure is effective, Case B where outcome measures are of diminishing effectiveness, and Case C where the individual outcome measures are all equally effective. Due to the complicated nature of this disjunctive outcome measure, our power and sample size calculations are mainly for the special case of equal cutoff points for all outcome measures; that is, for the special case where $c_1 = c_2 = \cdots = c_m = c^*$ say.

3.2.2 Optimal Common Cutoff Point for Equally Correlated Outcomes

The suggestion from the previous section that for a single outcome measure, dichotomizing at the placebo mean, $c_1 = 0$, is most powerful among the various values of c_1 we have investigated, motivates us to examine the value of the common cutoff point, c^*, which maximizes the power of the disjunctive outcome measure. An analytic examination is difficult but using C and S-Plus, it is straightforward to numerically evaluate the optimal cutoff point under the constraint of equal cutoff points for each of Cases A, B, and C.

We examine the special case of equally correlated outcomes with common correlation of $\rho = 0.0$, 0.3 and 0.5. For each scenario, the power is evaluated at values of the common cutoff point lying within a reasonable range and the value at which the power is

Table 3.14: Optimal common cutoff point (expressed as a multiple of Δ^*) for the disjunctive composite outcome measure with $n = 100$ for equally correlated outcome measures

		m = total number of outcome measures						
Case	ρ	1	2	3	4	5	10	20
A	0.0	0	0.669	1.316	1.752	2.077	3.016	3.658
	0.3	0	0.567	1.193	1.564	1.863	2.721	3.481
	0.5	0	0.443	0.968	1.319	1.582	2.326	2.978
B	0.0	0	0.718	1.372	1.812	2.140	3.086	3.936
	0.3	0	0.630	1.239	1.648	1.951	2.820	3.588
	0.5	0	0.514	1.054	1.414	1.749	2.434	3.049
C	0.0	0	0.631	1.238	1.647	1.953	2.844	3.862
	0.3	0	0.540	1.096	1.467	1.743	2.536	3.241
	0.5	0	0.422	0.907	1.245	1.640	2.139	2.735

maximized is identified as the optimal. Table 3.14 presents the results for the case of 100 patients per arm with the same choice of Δ^* as used in Chapter 2. The optimal common cutoff point, c^*_{opt}, is presented as a multiple of Δ^*, identified to a precision of 0.001. The optimal common cutoff point increases with a decreasing rate as the number of outcome measures increases. For example, when $m = 2$, c^*_{opt} lies in the range of 0.4 to about 0.7; on the other hand, when $m = 10$, c^*_{opt} is in the range of 2 to 3. This observation can be explain by the following reasoning: The power of this procedure is maximized when π_1 and π_2 are widely separated. When the total number of outcomes is large, π_1 and π_2 will be well separated only if the probabilities that a patient has a significant clinical change on the individual outcome measures are already widely separated across the two arms. Regardless of the standardized differences of the underlying means, a larger value of the cutoff point widens the separation on the individual outcome measures.

Table 3.14 shows that for any configuration of the m standardized differences considered, the optimal common cutoff point decreases as the positive correlation among the

multiple outcome measures increases. The optimal common cutoff point for Case A lies between that for Case B (largest) and for Case C (smallest). However, the differences are small, ranging from 0.1 to 0.3 multiples of Δ^* (i.e. ranging from 0.04 to 0.12). Thus, the configurations of the standardized differences of the underlying means considered do not have a great impact on the value of the optimal common cutoff point.

Additional numerical results (not presented) suggest that the common optimal cutoff point for equally correlated outcomes does not depend on n. However, without analytic verification, we will consider the results in Table 3.14 to be applicable only to the case of $n = 100$.

3.2.3 Properties for Equally Correlated Outcome Measures

The statistical properties of this approach will be explored for equally correlated outcomes. We will again consider Cases A, B, and C where different configurations of the standardized differences between the underlying means are examined. In addition, there is one more feature to be specified: the cutoff points, c_j. Dichotomizing at the placebo means on all m outcomes ($c_1 = c_2 = \cdots = c_m = c^* = 0$) seems a natural choice for several reasons. First, the mean of the responses of the patients on the placebo arm is then used as a guideline of a significant clinical change. Moreover, dichotomizing at the placebo mean helps to ensure a reasonable proportion (away from 0 and 1) of patients with significant clinical change on the placebo arm; if the histograms of responses on the placebo arm are roughly symmetric, then about 50% of the placebo patients will exhibit a significant clinical change on each of the individual outcome measures.

Table 3.15 provides the sample size required to achieve a power of 0.80 as well as the corresponding π_1 and π_2 when the common correlation $\rho = 0$, 0.3, 0.5. Before our discussion, it is worth emphasizing again that our calculations of the required sample size use the normal approximation and hence their accuracy relies on the appropriateness of

Table 3.15: Sample size required to achieve power of 0.80 for the disjunctive composite outcome with equally correlated outcome variables when all the cutoff points are at the placebo mean (i.e. $c_j = 0$ for $j = 1, 2, \ldots, m$). Second lines contain (π_1, π_2).

		m = total number of outcome measures						
Case	ρ	1	2	3	4	5	10	20
A	0.0	160	542	1300	2830	5890	195,000	200,000,000
		(.50, .35)	(.75, .67)	(.88, .84)	(.94, .92)	(.968, .959)	(.9990, .9987)	(1.0000, 1.0000)
	0.3	160	651	1630	3300	5930	44,400	434,000
		(.50, .35)	(.70, .63)	(.80, .76)	(.86, .83)	(.90, .88)	(.964, .960)	(.9893, .9887)
	0.5	160	742	1910	3840	6700	40,600	267,000
		(.50, .35)	(.67, .60)	(.75, .71)	(.80, .77)	(.83, .81)	(.909, .903)	(.9524, .9507)
B	0.0	160	204	308	495	828	14,900	8,980,000
		(.50, .35)	(.75, .62)	(.86, .79)	(.94, .89)	(.97, .94)	(.9990, .9977)	(1.0000, 1.0000)
	0.3	160	233	358	534	770	3190	19,000
		(.50, .35)	(.70, .58)	(.80, .71)	(.86, .80)	(.90, .85)	(.964, .949)	(.989, .986)
	0.5	160	252	391	571	795	2600	10,300
		(.50, .35)	(.67, .55)	(.75, .66)	(.80, .73)	(.83, .78)	(.91, .89)	(.952, .944)
C	0.0	160	110	105	115	139	668	38800
		(.50, .35)	(.75, .57)	(.86, .72)	(.94, .82)	(.97, .88)	(.999, .986)	(1.0000, .9998)
	0.3	160	125	121	125	133	196	365
		(.50, .35)	(.70, .53)	(.80, .64)	(.86, .72)	(.90, .77)	(.96, .89)	(.99, .95)
	0.5	160	134	131	133	137	164	216
		(.50, .35)	(.67, .50)	(.75, .59)	(.80, .65)	(.83, .69)	(.91, .80)	(.95, .88)

the normal approximation in each case. There are a few examples in the table where π_1 and π_2 approach 1 and the approximation may not be accurate; the inclusion of these examples is for illustration purposes only.

We first examine the special case of uncorrelated outcome measures. In Case A where only the first outcome is effective in comparing the two arms, the inclusion of even one ineffective outcome has a dramatic deleterious effect on this method. The separation between π_1 and π_2 decreases very quickly as additional ineffective outcomes are included because both approach 1 very quickly; for example, with $m = 4$, the difference between π_1 and π_2 is only about 0.02. Hence, for large m, the sample size required is huge as the difference between π_1 and π_2 is vanishingly small. In Case B where the outcomes have diminishing effectiveness in comparing the two arms, the effect of including one additional outcome is detrimental as well but much less dramatic than in Case A. The required sample size increases rather gradually as the number of outcomes increases. In Case C where all outcomes are equally effective, Table 3.15 shows that there is a value of the number of outcome measures below which the inclusion of an additional outcome is beneficial but above which such inclusion is detrimental. The results in Table 3.15 indicate that when the number of outcomes included is not larger than 3, the inclusion of an additional outcome is beneficial. On the contrary, such inclusion is detrimental when the number of outcomes included is larger than 3. Once m is 10 or more, the probability that a patient has a significant clinical change on any of the outcomes is close to 1 for both treatment arms. Thus, the difference to be detected between π_1 and π_2 is very small. The detrimental effect is then substantial but still much less dramatic than in Cases A and B.

We now turn to the examination of positively correlated outcomes. The effect of positive correlation among the multiple outcomes on the disjunctive measure is quite interesting. The results in Table 3.15 show that in Cases B and C, the effect of positive

correlation on this procedure depends upon the number of outcomes included. For example, with $m = 5$, for both Cases B and C, there is a value of ρ below which positive correlation has a positive impact but above which the effect is detrimental. In addition, in both Cases B and C with less than 5 outcomes, within the range of values of ρ considered, the effect of positive correlation is deleterious as the required sample size increases. On the other hand, with more than 5 outcomes, the required sample size decreases substantially as ρ increases. In Case A, there is a value of m below which the effect of positive correlation is negative but above which the effect is beneficial. For all of Cases A, B, and C, when the number of outcomes is large, say greater than 10, the effect of positive correlation is beneficial. For example, with $m = 10$, when ρ changes from 0 to 0.3, the impact is substantial: for all three cases, the required sample size decreases about 70% to 80%. However, as ρ changes from 0.3 to 0.5, the beneficial impact is only mild: there is only about 10% to 20% additional decrease in the required sample size.

So far, we have considered two choices for the common cutoff points, $c^* = c^*_{opt}$ and $c^* = 0$. We now examine the improvement of the performance of this procedure made by dichotomizing each outcome measure at the optimal common cutoff point instead of at the placebo means. We will abbreviate the disjunctive outcome measure with $c_j = c^*_{opt}$ as DCM^* and with $c_j = 0$ as DCM^0.

Table 3.16 presents the power achieved by DCM^* and DCM^0 with equally correlated outcomes for $n = 100$. In Case A where only one outcome measure is effective in comparing the two arms, the improvement in power made by c^*_{opt} is very limited for all three values of ρ. In Case C, the difference in power is only mild for $m = 2$, but as the number of outcome measures increases, the difference in power becomes more and more apparent. In addition, when m is fixed, the improvement in power made by DCM^* diminishes as the common correlation increases. We also notice that for Case C, the impact of including an additional outcome measure on DCM^* and DCM^0 differs.

Table 3.16: Power achieved by the disjunctive composite outcome measure with $n = 100$ for equally correlated outcome measures

			m = total number of outcome measures						
Case	ρ	c_j	1	2	3	4	5	10	20
A	0.0	0	0.60	0.224	0.121	0.082	0.065	0.051	0.050
		c^*_{opt}	0.60	0.232	0.140	0.104	0.087	0.061	0.053
	0.3	0	0.60	0.194	0.106	0.077	0.065	0.052	0.050
		c^*_{opt}	0.60	0.198	0.114	0.085	0.072	0.055	0.051
	0.5	0	0.60	0.176	0.098	0.074	0.063	0.052	0.053
		c^*_{opt}	0.60	0.178	0.102	0.077	0.067	0.054	0.051
B	0.0	0	0.60	0.498	0.356	0.241	0.163	0.056	0.050
		c^*_{opt}	0.60	0.518	0.434	0.370	0.320	0.193	0.116
	0.3	0	0.60	0.448	0.314	0.226	0.172	0.078	0.055
		c^*_{opt}	0.60	0.459	0.353	0.282	0.234	0.130	0.081
	0.5	0	0.60	0.420	0.292	0.215	0.168	0.085	0.059
		c^*_{opt}	0.60	0.426	0.314	0.245	0.201	0.111	0.073
C	0.0	0	0.60	0.761	0.782	0.741	0.660	0.191	0.052
		c^*_{opt}	0.60	0.776	0.857	0.902	0.929	0.978	0.995
	0.3	0	0.60	0.706	0.720	0.705	0.678	0.517	0.309
		c^*_{opt}	0.60	0.716	0.767	0.797	0.816	0.863	0.893
	0.5	0	0.60	0.675	0.686	0.679	0.666	0.589	0.477
		c^*_{opt}	0.60	0.681	0.714	0.734	0.750	0.777	0.798

For DCM^0, there is a value of m below which the inclusion of an additional outcome is beneficial but above which such inclusion is detrimental. On the contrary, for DCM^*, the power increases as the number of outcome measures included increases. Case B is more complicated. When m is 5 or less, for a fixed value of the common correlation, the differences in power increase as m increases. On the other hand, when m increases from 10 to 20, the difference in power decreases. Also for Case B, for a fixed number of outcomes, positive correlation has a negative impact on the improvement made by DCM^*. For both Cases B and C, DCM^* and DCM^0 are comparable only when the number of outcomes is 2 or less.

The results in Table 3.16 indicate the substantial improvement DCM^* can achieve over DCM^0 for Cases B and C. Nevertheless, use of DCM^* does not seem practical for at least two reasons. First, the numerically optimal common cutoff points might not be clinically meaningful. When the cutoff points used are not clinically meaningful, the interpretation of the results can be difficult. Second, the determination of these optimal cutoff points depends heavily on knowledge of the configuration of the standardized differences between the underlying means and the pattern of correlations. As high quality information on the properties of some of the outcome measures in current use in MS is scarce, the determination of the optimal cutoff points seems very difficult as well.

3.3 Comparisons to O'Brien's GLS Statistic

Now we want to bring the methods based on the disjunctive composite measure and O'Brien's GLS statistic together for comparison. We select O'Brien's GLS for comparison for two reasons. First, it is also a composite measure although the information from the individual outcome measures is combined in a very different way. Second, among all the procedures discussed in Chapter 2, it appears generally to be the most sensitive

Table 3.17: Power of DCM^* relative to GLS with 100 patients per arm

		m = total number of outcome measures						
Case	ρ	1	2	3	4	5	10	20
A	0.0	0.75	0.45	0.38	0.36	0.36	0.43	0.53
	0.3	0.75	0.48	0.46	0.50	0.56	0.79	0.85
	0.5	0.75	0.48	0.48	0.55	0.61	0.77	1.02
B	0.0	0.75	0.62	0.52	0.45	0.39	0.26	0.19
	0.3	0.75	0.62	0.54	0.50	0.48	0.48	0.58
	0.5	0.75	0.63	0.57	0.54	0.53	0.56	0.66
C	0.0	0.75	0.79	0.86	0.90	0.93	0.98	1.00
	0.3	0.75	0.76	0.79	0.81	0.83	0.87	0.89
	0.5	0.75	0.76	0.77	0.78	0.79	0.80	0.82

procedure in assessing the relative efficacy of the two arms. The large sample sizes required by the method based on the disjunctive outcome measure when each outcome measure is dichotomized at its placebo mean indicate that this method is not competitive with the methods discussed in Chapter 2. Therefore, the comparison we make here is between the power achieved with 100 patients per arm by O'Brien's GLS and DCM^*. (Note that as we only examine the case of equally correlated outcome measures, GLS and OLS are equivalent.)

The results in Tables 2.1 and 3.16 yield the ratio of the power of DCM^* to that of GLS presented in Table 3.17. For a single outcome measure, DCM^* loses 25% of the power achieved by GLS. We first consider the special case of uncorrelated outcome measures. In Case A where only the first outcome is effective in comparing the two arms, the results in Table 3.17 show that there is a value of m below which the inclusion of an additional ineffective outcome results in an decreased ratio but above which the ratio increases with the number of ineffective outcomes included. There is a dramatic decreases

of the ratio for $m = 1$ to $m = 2$, but for other values of m, the change in the ratio is modest. In Case A, with two or more uncorrelated outcomes, DCM^* loses about 50% of the power achieved by GLS. For Case B with uncorrelated outcome measures, the ratio of the power of DCM^* to that of GLS gradually decreases as m increases. In Case C where the outcomes are all equally effective, GLS is a very powerful procedure; although DCM^* is also quite powerful, it is not comparable for small numbers of outcomes. However, the advantage of GLS over DCM^* decreases as m increases.

For Case A, positive common correlation among the multiple outcomes has a positive impact on the ratio of the powers: for a fixed value of m, the ratio increases as ρ increases. With more than 10 modestly correlated outcomes, DCM^* is competitive with GLS although both perform very poorly. Similarly for Case B, for a fixed value of m, the larger the common correlation, the more competitive DCM^* is with GLS; however, the advantage of GLS is still substantial. In Case C, while positive correlation has a negative impact on the procedure based on GLS, this negative impact is even more substantial on DCM^*. Consequently, the ratio of the power of DCM^* to that of GLS decreases as ρ increases.

No comparison of the disjunctive outcome measure to GLS are made for other patterns of the correlations among the different outcome measures since the message is already very clear: The disjunctive composite outcome measure with common cutoff points is substantially less powerful than GLS.

3.4 Unequal Cutoff Points for Uncorrelated Outcomes

The modest performance of the disjunctive composite outcome measure with common cutoff points described in the previous sections prompts us to briefly consider the extent of improvement over DCM^0 that is possible with unequal cutoff points. We consider

Table 3.18: For Case A with three uncorrelated outcomes: Power achieved by *DCM* with 100 patients per arm (c_j is expressed as a multiple of Δ^*)

c_1	c_2	c_3	power	(π_1, π_2)
0	0	0	.12	(.88, .84)
0	1	1	.19	(.79, .72)
0	1	2	.23	(.74, .66)
0	2	2	.29	(.69, .60)
0	4	4	.49	(.55, .42)
0	6	6	.58	(.51, .36)
1	0	0	.09	(.84, .80)
2	0	0	.07	(.80, .78)

only the case of three uncorrelated outcome measures.

Table 3.18 presents a few choices of cutoff points and the resulting power achieved with 100 patients per arm for Case A. The values of π_1 and π_2 are also provided. The results suggest that dichotomizing the ineffective outcome measures at values larger than the placebo mean results in more powerful performance. For instance, the power gained by dichotomizing the three outcome measures at $(c_1, c_2, c_3) = (0, 4\Delta^*, 4\Delta^*)$ instead of at the placebo means (that is, $(c_1, c_2, c_3) = (0, 0, 0)$) is quite substantial. This can be explained by the following reasoning: As the cutoff point increases, the contribution of the ineffective outcomes to the overall composite outcome decreases. This enables *DCM* to make better use of the information provided by the first outcome. Consider the choice of cutoff points $(0, 6\Delta^*, 6\Delta^*)$ as an example. When the ineffective outcomes are dichotomized at $6\Delta^*$, the probability that a patient has a significant clinical change on either of these outcomes is negligible for both treatment arms. Consequently, the probability of treatment failure is mainly determined by the first outcome. In this case, the resulting π_1 and π_2, (.51, .36), are close to the values realized with only the single effective outcome (.50, .35). On the contrary, the choices $(1\Delta^*, 0, 0)$ and $(2\Delta^*, 0, 0)$ result

Table 3.19: For Case B with three uncorrelated outcomes: Power achieved by DCM with 100 patients per arm (c_j is expressed as a multiple of Δ^*)

c_1	c_2	c_3	power	(π_1, π_2)
0.0	0.0	0.0	.36	(.88, .79)
0.0	0.5	1.0	.45	(.81, .70)
0.0	2.0	4.0	.59	(.63, .47)
0.0	3.0	4.5	.61	(.58, .42)
1.0	0.5	0.0	.35	(.81, .72)
2.0	1.0	0.0	.31	(.74, .65)
1.0	0.0	0.0	.33	(.84, .75)
4.0	0.0	0.0	.23	(.76, .69)

in decreased power relative to DCM^0 . For these two choices, the separation between π_1 and π_2 is small as the contribution of the first outcome to the overall composite outcome is modest.

The results in Table 3.19 reveal that for Case B, dichotomizing the less effective outcomes at larger cutoff points results in increased power. The same reasoning as above can be applied to Case B. When the less effective outcome is dichotomized at a cutoff point larger than the placebo mean, its contribution to the overall composite measure is smaller. The gain in power can be substantial; see the choice of cutoff points (0, $3\Delta^*$, $4.5\Delta^*$) for example. On the other hand, dichotomizing the more effective outcomes at cutoff points larger than the placebo mean and the least effective outcome at the placebo mean results in a decrease in power; the choice ($4\Delta^*$, 0, 0) is an example with a modest decrease in power.

The results for Case C are presented in Table 3.20. In this case, with uncorrelated outcomes, it seems reasonable to consider equal cutoff points as all outcomes are equally effective in comparing the two arms. The choice ($1.25\Delta^*$, $1.25\Delta^*$, $1.25\Delta^*$) is close to the optimal common cutoff point, c^*_{opt}, for three uncorrelated outcomes with 100 patients

Table 3.20: For Case C with three uncorrelated outcomes: Power achieved by DCM with 100 patients per arm (c_j is expressed as a multiple of Δ^*)

c_1	c_2	c_3	power	(π_1, π_2)
0.00	0.00	0.00	.78	(.88, .72)
1.25	1.25	1.25	.857	(.72, .51)
1.25	1.25	1.50	.8564	(.66, .44)
1.00	1.25	1.50	.8560	(.67, .46)
1.25	2.00	4.00	.79	(.49, .30)

per arm. The results in Table 3.20 suggest that dichotomizing the outcome measures at unequal cutoff points results in decreased power. The degree of loss in power depends upon the extent of deviation from the optimal common cutoff point. Slight deviation from c^*_{opt}, such as the choices ($1.25\Delta^*$, $1.25\Delta^*$, $1.5\Delta^*$) and ($1\Delta^*$, $1.25\Delta^*$, $1.5\Delta^*$), result in very mild decreases in power. The choice ($1.25\Delta^*$, $2\Delta^*$, $4\Delta^*$) results in a modest decrease in power as its deviation from c^*_{opt} is larger.

This brief discussion of the impact of unequal cutoff points for three uncorrelated outcome measures illustrates the potential improvement of DCM over DCM^0 for each of Cases A, B and C. For Cases A and B, this can happen as dichotomizing the ineffective outcomes (for Case A) or the less effective outcomes (for Case B) at cutoff points larger than the placebo means enables DCM to make better use of the information provided by the effective or more effective outcomes. This improvement can be substantial, particularly for Case A. For Case C, the results suggest that DCM the use of unequal cutoff points does not result in improved performance, but the equal cutoff points for DCM should not be at the placebo means.

But, most importantly from a practical point of view, the results also indicate that the choice of good cutoff points requires knowledge of the standardized differences between

the underlying population means. If the researcher believes that the standardized differences are similar as in Case C, then DCM with equal cutoff points should be considered. On the other hand, if the researcher believes that the standardized differences are similar to Case A or B, use of the best single outcome measure as the primary endpoint is indicated. Of course, it is exactly the inability to identify the best single outcome measure a priori that leads to the consideration of methods based on multiple outcome measures. The necessary information on the characteristics of outcome measures for target patient populations is typically not available. Hence, the information required to allow the best possible cutoff points for DCM is typically not available either.

3.5 Discussion

In this chapter, a different type of composite outcome measure, the disjunctive composite outcome measure, has been discussed. The approach based on this composite outcome measure converts responses on the individual outcome measures into a single overall binary response indicating treatment failure or success which is employed as the single primary endpoint. The comparison between the dichotomized test and the Z-test on one outcome variable indicates that the percent power loss can be dramatic when the cutoff point is removed from the placebo mean, particularly for small samples. Also, the ARE of the dichotomized test relative to the Z-test is only about 64% for normal populations.

For the method based on the disjunctive composite outcome, we considered mainly two possibilities: DCM^0 corresponding to the choice of cutoff points $c_1 = \cdots = c_m = 0$, and DCM^* corresponding to the choice $c_1 = \cdots = c_m = c^*_{opt}$. The choice $c_1 = \cdots = c_m = 0$ seems quite natural as the placebo means are used to identify significant clinical changes. DCM^* was considered mainly for purposes of illustration as it does not seem to be practical. The results in Table 3.16 indicate the potential improvement associated

with the latter choice; the improvement in power can be substantial. However, when DCM^* is compared to O'Brien's GLS statistic, the former is clearly quite inefficient.

We also briefly considered the disjunctive composite outcome with unequal cutoff points. The tabulated powers for three uncorrelated outcome measures when $n = 100$ indicate that for Case C, the choice of equal cutoff points suffices. The results for Cases A and B illustrate that substantial gains in power can be obtained by dichotomizing ineffective and less effective outcomes at cutoff points larger than the placebo means. However, if the researcher has knowledge that certain outcomes are ineffective or weakly effective, excluding these outcome measures would be a better strategy. As a result, the disjunctive composite outcome with unequal cutoff points does not seem very useful.

This particular composite outcome measure combines the evidence from the individual binary responses disjunctively. Other ways of converting the binary responses on the original outcome measures into a single overall binary response could be considered. For example, a different composite outcome measure of "treatment failure" could be defined as worsening of a designated amount on all of the m outcome measures. We briefly examined its statistical properties for the case of uncorrelated outcome measures when the individual outcomes are dichotomized at the placebo means. For all of Cases A, B, and C, the impact of the number of outcome measures included on this new composite outcome measure is similar to that on DCM. The main difference is that whereas for DCM, π_1 and π_2 approach 1 very quickly, with this new outcome measure, π_1 and π_2 approach 0 very quickly. The resulting small separation between π_1 and π_2 leads to poor performance of this procedure.

Although the simplicity of this type of composite outcome measure is its big attraction, its simplicity can result in the loss of a substantial amount of information and therefore poor statistical performance. The main difficulty associated with this type

of composite outcome is that there seems to be no obvious rules for constructing reliable pre-assigned cutoff values for individual outcome measures and for the best way of combining the individual binary responses. Constructing a clinical meaningful and statistically powerful disjunctive outcome measure requires a lengthy process of empirical assessment. This can be done only when high quality information on the outcome measures in current use in MS is available.

Chapter 4

Applications

In this chapter, the five procedures we investigated in Chapters 2 and 3, namely, Bonferroni adjustment, Hotelling's T^2, O'Brien's OLS and GLS, and the disjunctive composite outcome measure, will be applied to two MS clinical trial data sets. Our earlier discussion of these procedures was quite general in the sense that various patterns of correlations among the outcome measures and three configurations of the standardized differences in the underlying means were considered. The objective here is to provide a more focused comparison among these procedures using specific outcome measures observed in MS patient populations. In particular, the sample correlations among the outcome measures will guide our choice of the pattern of MS patient population correlations. Further, we will consider configurations of the standardized differences in the underlying means suggested by the treatment effects observed in these data set. As before, the comparisons among the procedures are based on power and sample size calculations.

4.1 Task Force Data

The first data set, which we will refer to as the Task Force data, was provided by the National Multiple Sclerosis Society's Task Force on Clinical Outcome Assessments in MS. This international Task Force was created to develop recommendations for optimal clinical assessment measures for use in future MS clinical trials. Its initial deliberations are reported in Rudick, Antel, Confavreux et al. (1996) and its recommendations are reported in Rudick, Antel, Confavreux et al. (1997). Data were provided for a total of

Table 4.20: Baseline information by treatment group

	Placebo (N = 219)		Treatment (N = 216)	
Dimension	Mean	SD	Mean	SD
Arm	0.14	1.01	0.18	0.99
Leg	0.12	0.99	0.18	0.96
Cognitive	−0.35	0.76	−0.32	0.74

429 patients: 216 in a placebo arm and 213 in a treatment arm. For the context of investigations being carried out by the Task Force, three major clinical dimensions have been identified for the outcomes which are available in this data set: Arm Function, Leg Function, and Cognitive Function. Each dimension is measured by a composite outcome measure. The data provided consist of the z-scores of the composite outcome measures corresponding to the individual clinical dimensions at Baseline, Year 1, and Year 2 for each patient. (The standardization employed to create the z-scores provided was based on all the baseline data in a larger data set available to the Task Force consisting data from several MS clinical trials.) For all three dimensions, the z-scores were constructed so that higher scores represent better functional performance. In other words, a negative difference in the mean change of the z-scores between the placebo arm and the treatment arm (placebo − treatment) corresponds to a beneficial effect of the treatment.

4.1.1 Data Description

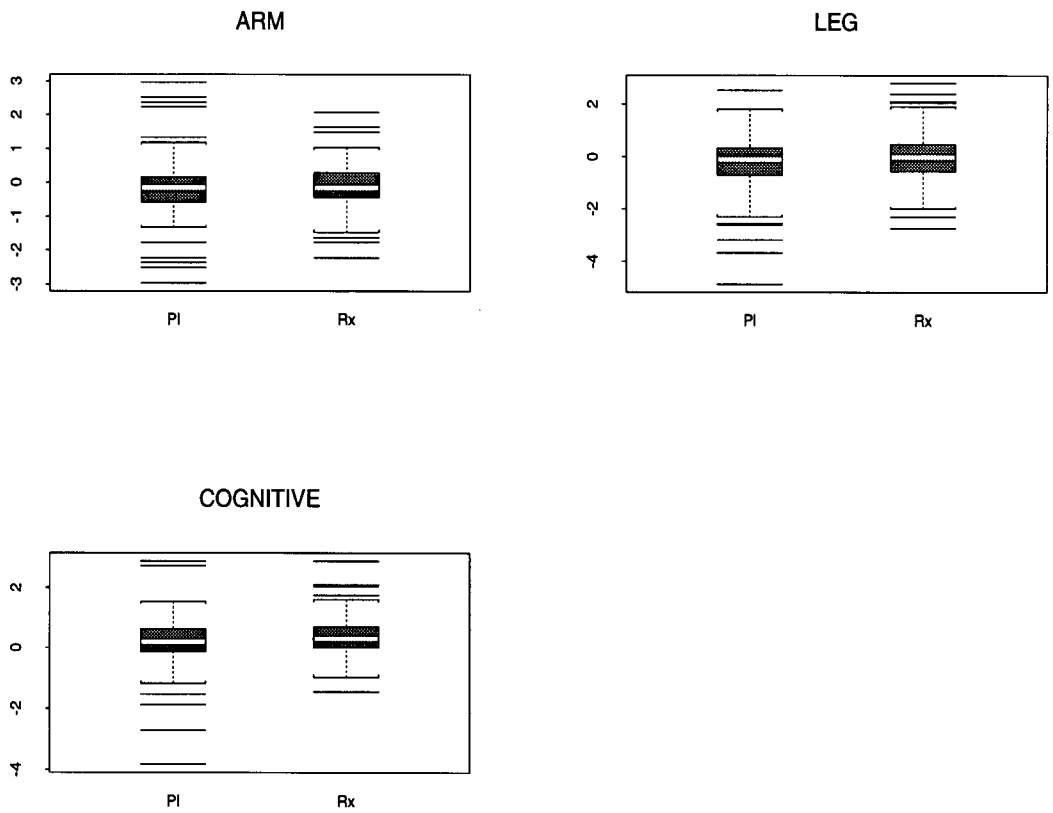
Table 4.20 summarizes the baseline information on the two arms. In addition to the summary statistics, the boxplots for each dimension (not presented) indicate that the patients on the two arms are comparable.

As we are interested in the change in the responses from the baseline to the end of the trial, we now turn to descriptive statistics for the changes from Baseline. As typically

Table 4.21: Summary of changes from Baseline to Year 2 by treatment group

	Placebo (N = 179)		Treatment (N = 152)	
Dimension	Mean	SD	Mean	SD
Arm	−0.15	0.78	−0.10	0.63
Leg	−0.24	0.96	0.01	0.89
Cognitive	0.26	0.75	0.33	0.62

Figure 4.3: Boxplots for the changes from Baseline to Year 2



often occurs in clinical trials, some patients did not provide data on one or more of the three dimensions at Year 2. For our purposes, we will focus on the changes from Baseline to Year 2 and regard those patients who did not provide complete data on all the three dimensions as dropouts. With this convention, approximately 24% of the patients are dropouts at Year 2, with a substantially higher percentage of dropouts occurring on the treatment arm.

Table 4.21 summarizes the changes from Baseline to Year 2 by treatment group and Figure 4.3 presents the boxplots of these changes. The summaries in Table 4.21 reveal that the changes in the responses are rather small on all three clinical dimensions, with Leg Function being more effective in comparing the two arms than the other clinical dimensions. The boxplots show that the data on the individual dimensions are roughly symmetrically distributed for both arms. There are quite a few outliers for the individual dimensions on both arms. In addition, the variability of the changes from Baseline to Year 2 on both arms are comparable (although the SD's on the treatment arm are slightly smaller on all three dimensions).

The correlation matrices of the changes from Baseline to Year 2 among the three dimensions, denoted as $\widehat{M}_{\rho_{Pl}}$ and $\widehat{M}_{\rho_{Rx}}$, are:

$$\widehat{M}_{\rho_{Pl}} = \begin{pmatrix} & Arm & Leg & Cog. \\ Arm & 1.00 & 0.34 & 0.20 \\ Leg & 0.34 & 1.00 & 0.28 \\ Cog. & 0.20 & 0.28 & 1.00 \end{pmatrix},$$

$$\widehat{M}_{\rho_{Rx}} = \begin{pmatrix} & Arm & Leg & Cog. \\ Arm & 1.00 & 0.33 & 0.16 \\ Leg & 0.33 & 1.00 & 0.21 \\ Cog. & 0.16 & 0.21 & 1.00 \end{pmatrix}.$$

The patterns of correlations for the two arms are similar: the three dimensions are only modestly correlated.

4.1.2 Results

Our objective is to investigate the appropriateness of the methods discussed in earlier chapters for a MS clinical trial involving therapies with similar characteristics; that is to say, a MS clinical trial using responses on these outcomes to assess the treatment efficacy based on similar patient populations. In addition, we hope to demonstrate the usefulness of the comparison among the methods in designing a clinical trial.

We will use the information from this Task Force data as the basis of our investigation. As already illustrated, the relationship of the changes from Baseline to Year 2 among the three clinical dimensions and the variability of these changes on the individual dimensions are similar on the two arms, so the assumption of a common variance-covariance matrix seems to be a reasonable approximation. The sample variance-covariance matrix of the changes from Baseline to Year 2 for the placebo arm will be taken to be the common variance-covariance matrix of these changes in the populations. In other words, the variance-covariance matrices for these changes are assumed to be known and common for both populations. Thus, the standard deviations are taken to be: $\sigma_{Arm} = 0.78$, $\sigma_{Leg} = 0.96$, and $\sigma_{Cog.} = 0.75$. (Note that use of the larger standard deviations provided by the data on the placebo-treated patients would be expected to lead to conservative results.) The correlations among the three clinical dimensions are taken to be: 0.34 between Arm and Leg, 0.20 between Arm and Cognitive, and 0.28 between Leg and Cognitive.

With this pattern of correlations, the average correlation is $\overline{\rho} = 0.27$, and the weights GLS assigns to Arm, Leg and Cognitive are 0.34, 0.30, and 0.36 respectively. Relating to our work in Chapters 2 and 3, we make two remarks on this particular pattern of

correlations. First, this pattern of correlations can be considered similar to the case of having three equally correlated outcomes with the common correlation of 0.3 (although the degree of correlation is slightly weaker here). Presumably, provided that the standardized differences between the underlying means are relevant, the comparisons among the methods should be similar to our work in the previous chapters. Second, under this correlation structure, as the correlations among the dimensions are roughly equal, the performance of O'Brien's GLS and OLS should be very similar. This is also indicated by the roughly equal weights GLS assigns to the three dimensions.

In addition, we will use the treatment effects observed to be indicative of the "true" treatment effects. The data suggest the standardized differences between the underlying population mean changes of: $\Delta_{Arm} = -0.05$, $\Delta_{Leg} = -0.27$, and $\Delta_{Cog.} = -0.10$. Note that for this particular data set, as higher scores correspond to better performance, a negative standardized difference between the mean changes (placebo − treatment) corresponds to a beneficial treatment effect. Under these presumed population characteristics, Leg is the most effective outcome measure in comparing the two arms, Cognitive is only modestly effective, and Arm is quite ineffective. This seems similar to our Case B where the outcome measures are of diminishing effectiveness although the rate of diminishing is faster here.

Before proceeding with the comparisons of the procedures, we emphasize that, the version of the disjunctive outcome measure used here is DCM^0, where each dimension is dichotomized at the placebo mean (i.e. $c_j = 0$ for all j).

Table 4.22 provides the power each of the five procedures achieves with 100 patients per arm. The sample sizes required to achieve a power of 0.80 are presented in Table 4.23. We first consider a rounded version (for simplicity) of the observed standardized differences; namely, $\Delta_{Arm} = -.05$, $\Delta_{Leg} = -.30$, and $\Delta_{Cog.} = -.10$. Bonferroni adjustment and Hotelling's T^2 are more powerful than the other procedures and are comparable;

Table 4.22: Power of procedures with 100 patients per arm

Configuration			Procedure				
Δ_{Arm}	Δ_{Leg}	$\Delta_{Cog.}$	Bon.	T^2	OLS	GLS	DCM^0
$-.05$	$-.30$	$-.10$	.42	.41	.31	.29	.14
.00	$-.30$	$-.10$	.41	.45	.26	.24	.12
.00	$-.30$	.00	.40	.47	.17	.14	.08
$--$	$-.30$	$-.10$	.47	.46	.42	.42	.23
$--$	$-.30$	$--$	.56	.56	.56	.56	.39
$-.30$	$-.30$	$-.30$	.70	.70	.84	.84	.48

Table 4.23: Sample size required to achieve power of 0.80

Configuration			Procedure				
Δ_{Arm}	Δ_{Leg}	$\Delta_{Cog.}$	Bon.	T^2	OLS	GLS	DCM^0
$-.05$	$-.30$	$-.10$	228	233	360	399	1030
.00	$-.30$	$-.10$	228	212	455	514	1400
.00	$-.30$	.00	232	203	809	1020	2960
$--$	$-.30$	$-.10$	206	213	251	251	526
$--$	$-.30$	$--$	174	174	174	174	277
$-.30$	$-.30$	$-.30$	125	125	90	90	213

with 100 patients per arm, both achieve power of around .40. OLS and GLS have power of around .30 and OLS has a small advantage. With power of only .14, DCM^0 is clearly inferior to the other procedures. Comparing the sample sizes required to achieve power of .80, Bonferroni adjustment and T^2 require only about $\frac{2}{3}$ as many patients as OLS and GLS and only about $\frac{1}{4}$ as many as DCM^0. With this particular correlation structure, the advantage of OLS over GLS is expected because GLS assigns less weight (.30 versus .33) to the most effective outcome measure Leg. But the difference in these weights is small, so as long as the effectiveness of Leg in comparing the two arms is not overwhelming, the advantage of OLS over GLS will be modest.

Now consider planning a clinical trial using these outcome measures to assess the treatment efficacy. Suppose that the researcher, who is willing to assume a common variance-covariance matrix for the populations, is convinced that the specified standardized differences between the underlying population means and the specified correlation structure are the most relevant values. If all three outcome measure are to be employed, the above power and sample size calculations indicate that the procedure based on the Bonferroni adjustment will provide the most sensitive evaluation of the results of the trial. These calculations suggest that about 230 patients per arm are required to detect the specified standardized differences with a probability of 0.80.

Following these basic calculations, the researcher may want to examine several further aspects. For example, because the above results are limited to the case of power of 0.80, the researcher may want to explore how the power relates to the sample size for each procedure. In addition, the researcher may be interested in how the relationship among the procedures changes with the sample size. This more detailed investigation helps the researcher to determine if one procedure is consistently most sensitive in the assessment of the treatment efficacy and hence truly is the one to be used in the design and analysis of the study. The power of the procedures based on Bonferroni adjustment, Hotelling's

Figure 4.4: Power of Bonferroni adjustment, Hotelling's T^2, OLS, and GLS as a function of n when $(\Delta_{Arm}, \Delta_{Leg}, \Delta_{Cog.}) = (-.05, -.30, -.10)$

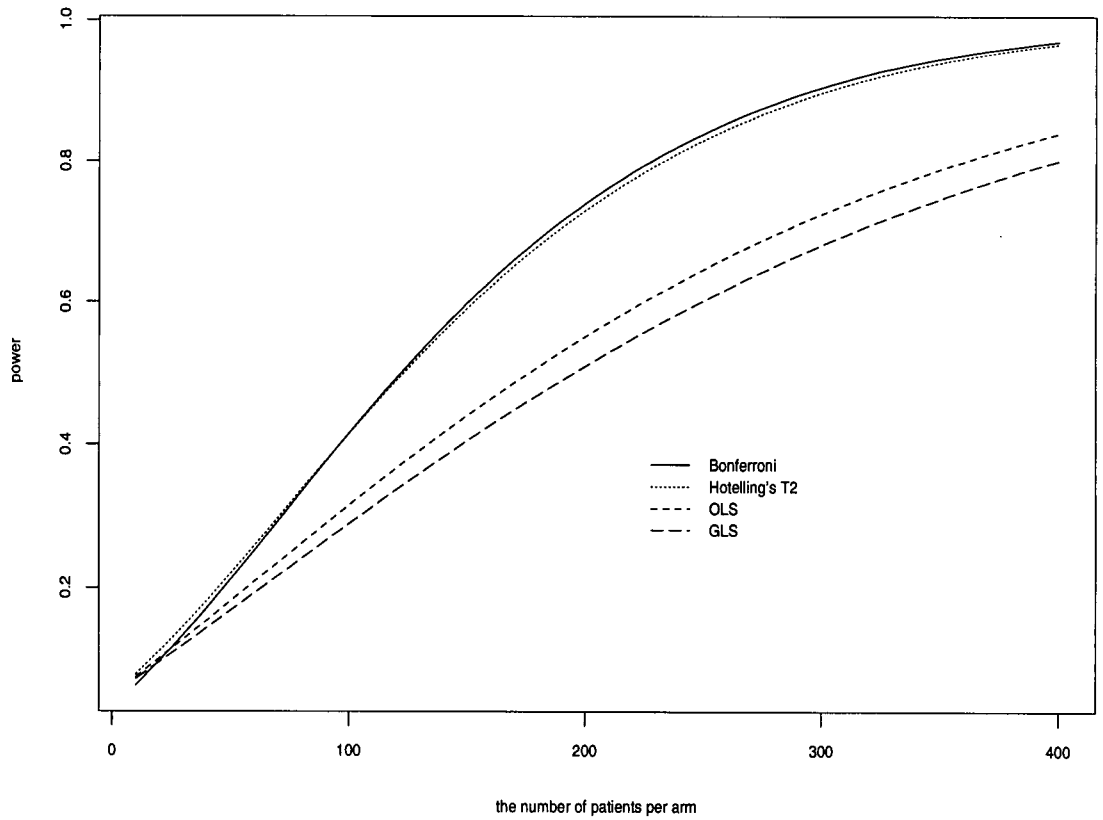
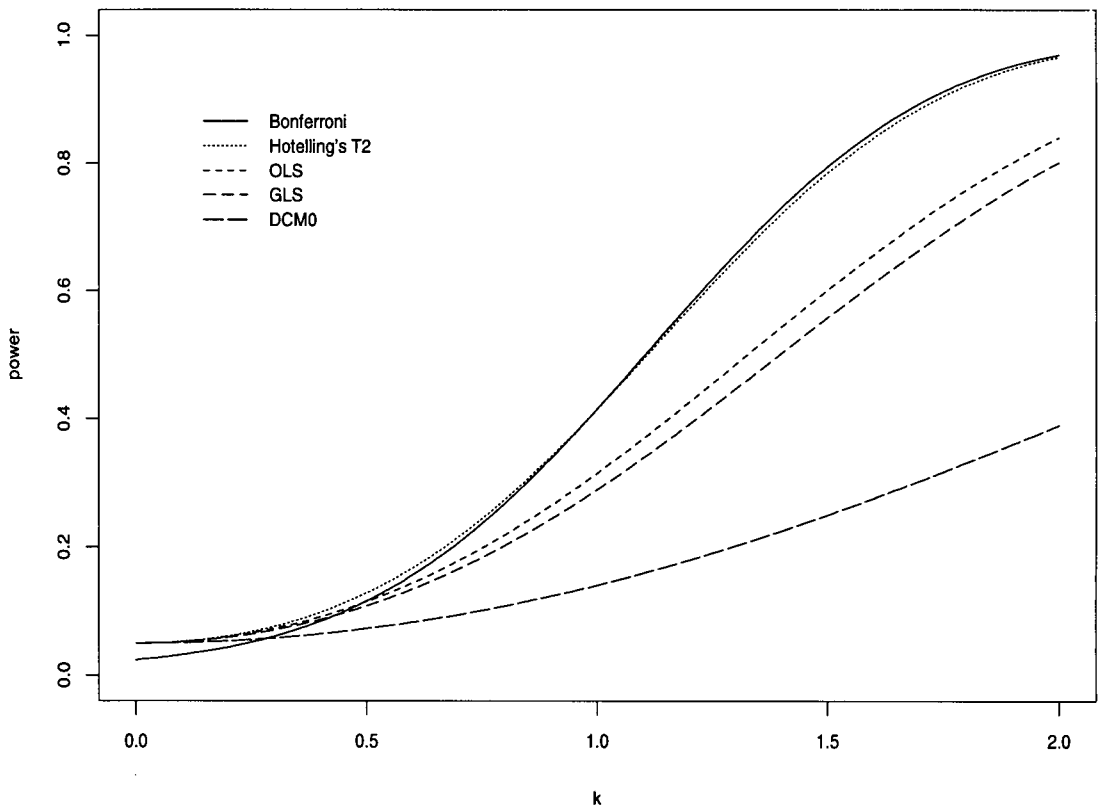


Figure 4.5: Power of procedures with 100 patients per arm when $\boldsymbol{\Delta} = k \cdot \boldsymbol{\Delta}_{base}$, where $\boldsymbol{\Delta}_{base} = (-.05, -.30, -.10)$



T^2, OLS and GLS as a function of the sample size per arm is presented in Figure 4.4. (Note that DCM^0 is not considered because of its clear inferiority.) This plot shows that Bonferroni adjustment and Hotelling's T^2 are competitive; however, the former is slightly more powerful when there are more than about 50 patients per arm. OLS is consistently more powerful than GLS for the values of n considered. Figure 4.4 reveals that the relationship among the procedures is quite similar for different values of power; therefore, Bonferroni adjustment should be used in this situation.

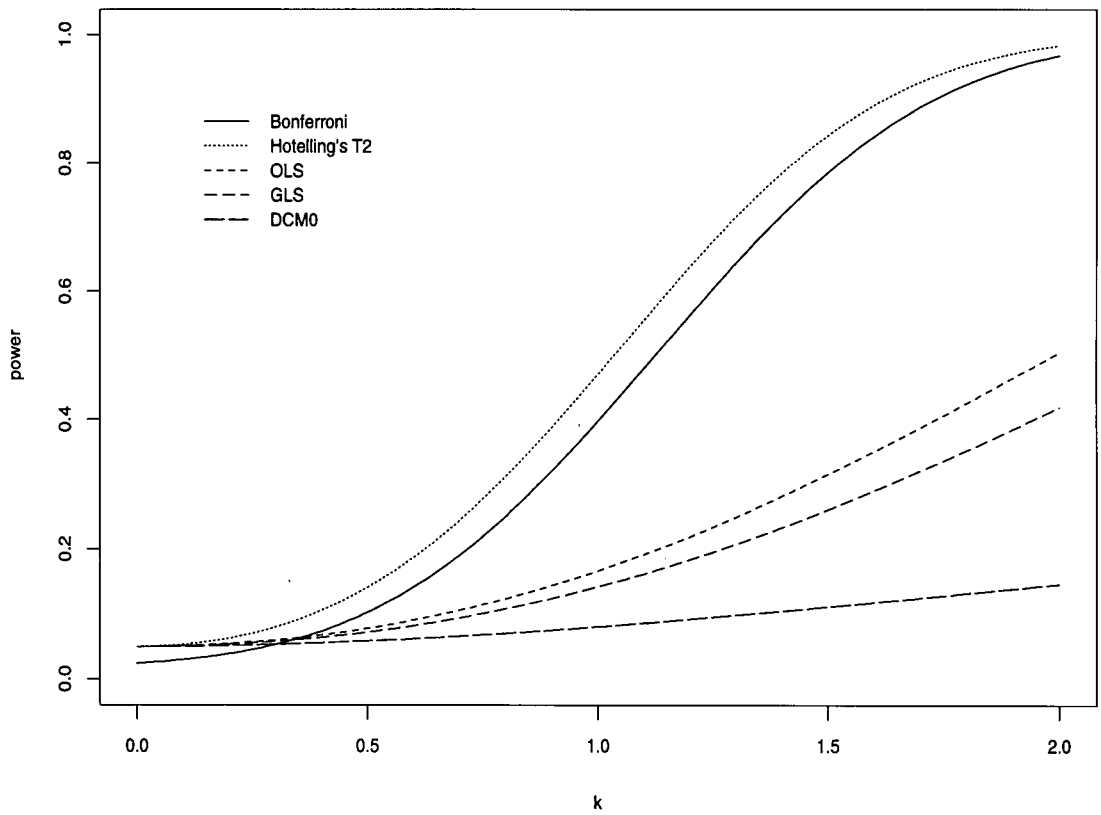
We next examine the performance of these procedures under a few other configurations

relevant to the specified one. We will denote the specified standardized differences (Δ_{Arm}, Δ_{Leg}, $\Delta_{Cog.}$) as $\boldsymbol{\Delta}_{base}$. Suppose the true treatment effects are a multiple of the specified treatment effects, i.e. $\boldsymbol{\Delta} = k \cdot \boldsymbol{\Delta}_{base}$. (The configuration of $\boldsymbol{\Delta}_{base}$ corresponds to the case $k = 1$.) With $\boldsymbol{\Delta}_{base} = (-0.05, -0.30\ -0.10)$, Figure 4.5 shows how the power with 100 patients per arm changes as k ranges from 0 to 2. This figure shows that when the true treatment effects are less than one-half of the specified ones ($k < 0.5$), none of the procedures are sensitive in the assessment of the treatment efficacy. Due to the inclusion of two almost ineffective outcome measures and only one modestly effective outcome measure, all five procedures perform poorly. Figure 4.5 shows that for $0.5 < k \leq 2$, the procedures based on Bonferroni adjustment and Hotelling's T^2 are competitive and have a clear advantage over the other procedures. When the treatment effects are 50% larger than the specified treatment effects, i.e. $k = 1.5$ and $\boldsymbol{\Delta} = (\Delta_{Arm}, \Delta_{Leg}, \Delta_{Cog.}) = (-.75, -.45, -.10)$, the procedures based on the Bonferroni adjustment and Hotelling's T^2 have reasonable sensitivity with 100 patients per arm.

As the data suggest that Arm is only modestly effective in comparing the two arms, we next consider the more extreme configuration where Arm is an ineffective outcome measure: $\boldsymbol{\Delta} = (.00, -.30\ , -.10)$. Comparing to the configuration $\boldsymbol{\Delta}_{base}$, the power achieved by all procedures except Hotelling's T^2 decreases slightly for this configuration. The inclusion of an ineffective outcome measure instead of a weakly effective outcome results in a slight improvement in the performance of the procedure based on Hotelling's T^2. This illustrates the complicated nature of this procedure. For this configuration, with power of .45, Hotelling's T^2 performs slightly better than Bonferroni adjustment (power of .41), moderately better than OLS and GLS (power around .25) and substantially better than DCM^0 (power of .13). Both OLS and GLS require more than twice as many patients as T^2.

We next consider the even more extreme situation where Cognitive is also ineffective

Figure 4.6: Power of procedures with 100 patients per arm when $\boldsymbol{\Delta} = k \cdot \boldsymbol{\Delta}_{base}$, where $\boldsymbol{\Delta}_{base} = (.00, -.30, .00)$

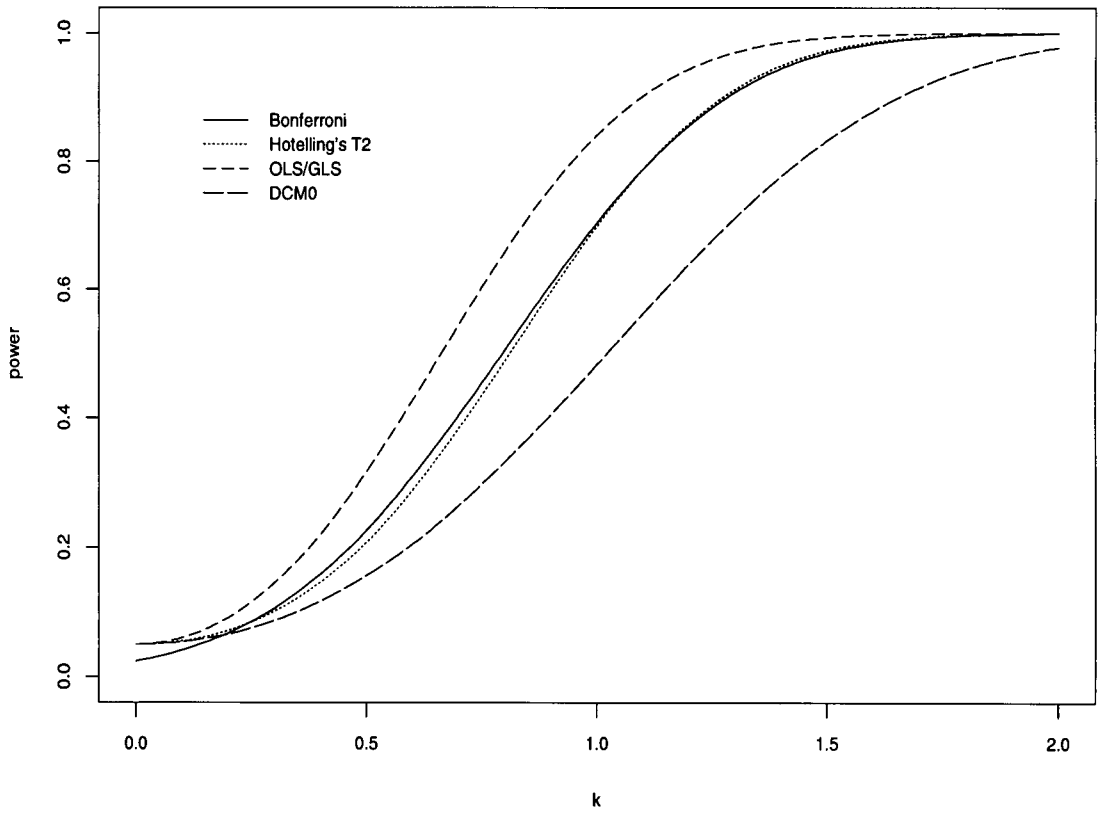


(i.e. $\boldsymbol{\Delta} = (.00, -.30, .00)$) which is analogous to our Case A. Comparing to the previous configuration, again, all procedures except Hotelling's T^2 are less powerful for this configuration. It is interesting to observe that T^2 actually performs slightly better when both Arm and Cognitive are ineffective. The performance of Bonferroni adjustment is not much affected. For both GLS and OLS, the penalty for including an ineffective outcome measure instead of a mildly effective outcome is quite substantial. Taking $\boldsymbol{\Delta} = (.00, -.30, .00)$ as $\boldsymbol{\Delta}_{base}$, Figure 4.6 shows how the power of the procedures with 100 patients per arm is affected by the magnitude of the effectiveness of the single effective outcome measure. This figure indicates the advantage of Hotelling's T^2. Bonferroni adjustment is the only procedure which is competitive with T^2.

We then examine the impact of excluding less effective outcome measures. Based on the observed standardized differences, Arm is the least effective outcome. Suppose the dimension Arm is deleted; the configuration considered is: $\boldsymbol{\Delta} = (--, -.30, -.10)$. The results in Tables 4.22 and 4.23 indicate that all procedures benefit from the exclusion of the least effective outcome. This exclusion has a great impact on the performance of DCM^0: it now requires only half as many patients to achieve a power of 0.80. The impact of excluding the least effective outcome measure on OLS and GLS is moderate; this impact on Bonferroni adjustment and T^2 is only mild.

Suppose now only the most effective outcome is included: $\boldsymbol{\Delta} = (--, -.30, --)$. For this configuration, only a single outcome measure is included so Bonferroni adjustment, T^2, OLS, and GLS procedures are identical provided two-sided tests are carried out in each case. Comparing to the results for the configurations considered earlier, we note that for this pattern of correlations, the inclusion of the other nearly ineffective or weakly effective outcome measures has a detrimental effect on all of the procedures. This should deliver a clear message to the researcher that the choice of outcome measures to be used is crucial.

Figure 4.7: Power of procedures with 100 patients per arm when $\boldsymbol{\Delta} = k \cdot \boldsymbol{\Delta}_{base}$, where $\boldsymbol{\Delta}_{base} = (-.30, -.30, -.30)$



Finally, we consider a more optimistic configuration analogous to our Case C, where Arm and Cognitive are as effective as Leg: $\boldsymbol{\Delta} = (-.30, -.30, -.30)$. GLS and OLS are expected to perform more powerfully as indicated by the results for three equally correlated outcome measures with common correlation of 0.30 in Chapter 2. Moreover, GLS should have a small advantage over OLS as it assigns slightly more weight Cognitive. Again, we consider different magnitudes of effectiveness, taking $\boldsymbol{\Delta} = k \cdot \boldsymbol{\Delta}_{base}$, where $\boldsymbol{\Delta}_{base} = (-.30, -.30, -.30)$. As the difference in power between GLS and OLS is negligible, only the power of GLS is displayed on Figure 4.7. For this pattern of correlations, when the three outcomes are equally effective in comparing the two arms, GLS is most powerful. We also notice the substantial improvement in the performance of DCM^0 due to the increased sensitivity on each dimension. Comparing the configuration $\boldsymbol{\Delta} = (-.05, -.30, -.10)$ to the configuration $\boldsymbol{\Delta} = (-.30, -.30, -.30)$, the results in Table 4.23 show that the latter requires only about 20% as many patients to achieve adequate sensitivity.

Summary

Suppose a researcher is planning a MS clinical trial with therapies having similar characteristics as those investigated in the study which led to the Task Force data. Suppose also that s/he is willing to assume a common variance-covariance matrix for both populations and is convinced that the observed standardized differences and the sample correlation structure of the placebo arm are the most relevant values. Our power and sample size calculations indicate that if the researcher intends to use one of these procedures with all three outcome measures, the procedure based on Bonferroni adjustment is the best way to proceed. However, even when the researcher has a reasonably good knowledge of the correlation structure, s/he is still unlikely to know the configuration of the standardized differences. Consequently, it is still very difficult to conclude which procedure performs

better under the specified pattern of correlation structure. However, the results in Tables 4.22 and 4.23 and Figures 4.5 to 4.7 provide several clear messages for a clinical trial with three outcome measures and for the specific pattern of correlation structure considered where all three outcomes are modestly correlated. First, DCM^0 with each dimension dichotomized at the placebo mean is not comparable to the other procedures. Second, when all dimensions are equally effective in comparing the two arms, O'Brien's GLS is most powerful no matter the magnitude of effectiveness. Third, when only a single outcome measure is effective, T^2 is most powerful. Finally, when the situation is intermediate; for example, when all three dimensions are effective but with unequal effectiveness, Bonferroni adjustment and T^2 are competitive and perform better than the other procedures.

The results in Tables 4.22 and 4.23 also demonstrate the detrimental effect of including less effective outcome measures on the performance of all procedures under the specified correlation structure. The researcher who is planning such a clinical trial should definitely consider including only the most effective outcome provided s/he is convinced that the magnitude of observed standardized differences are most relevant.

4.2 Oral Methotrexate Data

The second MS clinical trial data set, which we will refer to as the Oral Methotrexate data, originated with the randomized, placebo-controlled, double-blind clinical trial of oral methotrexate in chronic progressive MS (Goodkin et al. 1989) and was provided by Dr. D. Goodkin. A total of 60 patients were involved in this study: 29 in the placebo arm and 31 in the treatment arm. The data consist of six outcome measures: EDSS, Ambulation Index (AMB), the Box and Block Test on the left arm (LBB), and the Box and Block Test on the right arm (RBB), the Nine Hole Peg Test on the left arm (L9HP),

the Nine Hole Peg Test on the right arm (R9HP). For this two-year study, responses were obtained monthly, but our data set contains these responses at only baseline, Year 1 and Year 2. The original analysis of the data for this clinical trial was based on the monthly data and employed a single primary endpoint, the proportion of patients experiencing treatment failure. This endpoint was a disjunctive composite outcome measure which will be described in Section 4.2.3.

EDSS, an ordinal scale taking values from 0.0 to 10 in steps of 0.5, measures the degree of neurologic impairment on nine functions which are believed to be most relevant to MS. The Ambulation Index, a 10-step ordinal scale, is an assessment of the time required to walk 25 feet. Although both EDSS and AMB are ordinal variables, for the sake of simplicity, we will treat them as continuous variables in what follows. The response on the Box and Block Test, a timed test given separately for the left and right arms, is the total number of blocks a patient puts into a box within 60 seconds (a higher score represents better performance). The response on the Nine Hole Peg Test, another timed test given separately for the left and right arms, is the time (in seconds) a patient takes to put nine pegs into pre-specified holes (a lower score corresponds to better performance). For those patients who failed to complete this test, a score of 777 seconds was assigned to indicate the failure to complete the task and to differentiate these responses from the missing values for those who did not take the test. (We do not know exactly why 777 was chosen. The largest score for patients completing the task was 342.8 seconds.) For the Nine Hole Peg Test and the Box and Block Test, instead of using the left hand and right hand scores separately, we will use the average scores. We create a new outcome measure, BB, which represents the average number of blocks a patient puts into a box within 60 seconds. Similarly, 9HP represents the average time (in minutes) a patient takes to put nine pegs into the pre-specified holes. Since 9HP is a timed measure, its reciprocal, I9HP $= 1/9$HP, represents the rate at which the task is completed. To be

Table 4.24: Baseline information by treatment group

	Placebo (N = 29)		Treatment (N = 31)	
Response	Mean	SD	Mean	SD
EDSS	5.27	1.45	5.48	1.26
AMB	4.14	1.72	4.03	1.47
NBB	−46.63	8.12	−49.77	10.63
NI9HP	−1.67	0.58	−2.07	0.58

consistent with EDSS and AMB for which lower scores represent better performance, BB and I9HP need to be transformed so that lower scores also represent better performance. We will use NBB = −BB, and NI9HP = −I9HP in what follows. Therefore, for the four outcome measures considered, EDSS, AMB, NBB, and NI9HP, a positive difference of the mean changes between the placebo arm and the treatment arm (placebo − treatment) indicates a beneficial treatment effect.

4.2.1 Data Description

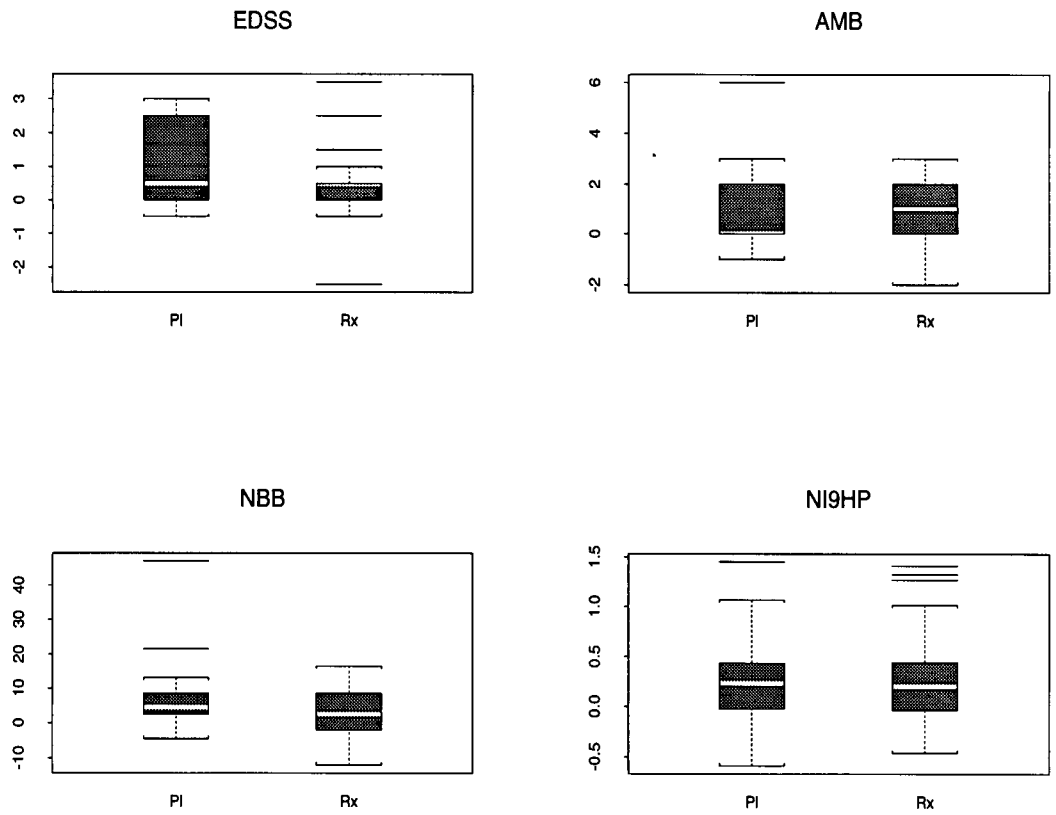
Table 4.24 provides the baseline summary statistics for the two treatment groups. The patients on the two arms are quite comparable at baseline.

We now examine some descriptive statistics for the changes from baseline. As for the previous application, we focus on the changes from Baseline to Year 2. Table 4.25 provides the summary of the changes from Baseline to Year 2 by treatment group. Figure 4.8 presents the boxplots of these changes for the individual outcome measures. The summaries in Table 4.25 reveal that the treatment appears to have beneficial effects on EDSS, AMB and NBB but not on NI9HP. The boxplots indicate some departures from normality. For example, the collection of changes in EDSS on the placebo arm is heavily

Table 4.25: Summary of changes from Baseline to Year 2 by treatment group

	Placebo (N = 22)		Treatment (N = 23)	
Response	Mean	SD	Mean	SD
EDSS	1.02	1.24	0.41	1.09
AMB	1.00	1.66	0.78	1.17
NBB	7.18	10.50	3.07	6.91
NI9HP	0.29	0.45	0.33	0.50

Figure 4.8: Boxplots for the changes from Baseline to Year 2



skewed to the right. Also, there are a few outliers in AMB and NBB on the placebo arm and in EDSS and NI9HP on the treatment arm. The boxplots indicate the variability of these changes in the two populations are reasonably comparable although Table 4.25 indicates the standard deviations are somewhat smaller in the treatment arm except for NI9HP.

The sample correlations of the changes from Baseline to Year 2 among the four outcome measures are:

$$\widehat{M}_{\rho Pl} = \begin{pmatrix} & EDSS & AMB & NBB & NI9HP \\ EDSS & 1.00 & 0.72 & 0.36 & 0.34 \\ AMB & 0.72 & 1.00 & 0.80 & 0.61 \\ NBB & 0.36 & 0.80 & 1.00 & 0.75 \\ NI9HP & 0.34 & 0.61 & 0.75 & 1.00 \end{pmatrix},$$

$$\widehat{M}_{\rho Rx} = \begin{pmatrix} & EDSS & AMB & NBB & NI9HP \\ EDSS & 1.00 & 0.82 & 0.27 & 0.48 \\ AMB & 0.82 & 1.00 & 0.15 & 0.36 \\ NBB & 0.27 & 0.15 & 1.00 & 0.47 \\ NI9HP & 0.48 & 0.36 & 0.47 & 1.00 \end{pmatrix}.$$

The correlations among EDSS and the other outcome measures show a similar pattern for both arms. On the other hand, the pattern of correlations among AMB, NBB, and I9HP differs substantially between the two arms: all three correlations are considerably stronger on the placebo arm than on the treatment arm.

4.2.2 Results

The objective of our investigation in this subsection is to illustrate how the comparisons among the methods can assist the researcher in planning a study. Our focus is on MS clinical trials with treatment having characteristics similar to those investigated in the study which led to the Oral Methotrexate data.

The information from the Oral Methotrexate data will be the basis of our investigation. Assuming the variabilities of the changes from Baseline to Year 2 are equal in both populations, we will take the standard deviations of these changes on the placebo arm as the standard deviations of these changes in the populations, σ_{EDSS}, σ_{AMB}, σ_{NBB} and σ_{NI9HP}. (Because the standard deviations for EDSS, AMB, and NBB are larger on the placebo arm, our results might be conservative.) The data suggest that the assumption of equal variability for the populations is reasonable as a rough approximation.

We are sometimes in a situation where the researcher has the knowledge of the correlation structure only for the placebo population (because data for placebo patients are often available from previous trials but that for treated patients is not). Suppose that the researcher is willing to assume that the correlation structures are common for the populations. Under such a situation, the best one can do is to take $\widehat{\boldsymbol{M}}_{\rho Pl}$ as a guide for the pattern of the population correlations among the outcome measures. This is how we will proceed in specifying the pattern of correlations among the four outcome measures (despite the substantial differences in the observed correlation structures between the two arms). Guided by $\widehat{\boldsymbol{M}}_{\rho Pl}$, we notice that the correlations between EDSS and NBB and EDSS and NI9HP are about the same (average = 0.35) and the remaining correlations, while considerably stronger, are also similar (average = 0.72). For simplicity, we will take the respective average values to represent the common correlation structure for both populations and we have:

$$M_\rho = \begin{pmatrix} & EDSS & AMB & NBB & NI9HP \\ EDSS & 1.00 & 0.72 & 0.35 & 0.35 \\ AMB & 0.72 & 1.00 & 0.72 & 0.72 \\ NBB & 0.35 & 0.72 & 1.00 & 0.72 \\ NI9HP & 0.35 & 0.72 & 0.72 & 1.00 \end{pmatrix}$$

With this particular structure, we have three highly correlated outcome measures: AMB, NBB, and NI9HP. EDSS is highly correlated with AMB but only modestly correlated with NBB and NI9HP. The average of the correlations in this structure is $\overline{\rho} = 0.60$. GLS is expected to assign EDSS the most weight and AMB the least weight, with equal and moderate weights assigned to NBB and NI9HP. The weights GLS assigns to EDSS, AMB, NBB, and NI9HP are 0.63, -0.43, 0.40, and 0.40 respectively.

The observed treatment effect suggests standardized differences between the underlying mean changes of the populations of: $\Delta_{EDSS} = .49$, $\Delta_{AMB} = .13$, $\Delta_{NBB} = .39$ and $\Delta_{NI9HP} = -.09$. EDSS is the most effective outcome measure in comparing the two arms, NBB is moderately effective, AMB is modestly effective, and NI9HP is nearly ineffective. As indicated earlier, NI9PH shows a detrimental treatment effect, so the directions of the treatment effects on the individual outcome measures are not consistent.

Treating the pattern of the correlations in the populations to be known and common, we examine a few configurations of the standardized differences. Tables 4.26 and 4.27 present the power achieved with 100 patients per arm and the sample size required to achieved power of 0.80 for the five procedures, Bonferroni adjustment, Hotelling's T^2, O'Brien's OLS and GLS, and DCM^0.

We first consider a configuration of standardized differences suggested by the data; for simplicity, a rounded version $\boldsymbol{\Delta} = (\Delta_{EDSS}, \Delta_{AMB}, \Delta_{NBB}, \Delta_{NI9HP}) = (.50, .10,$

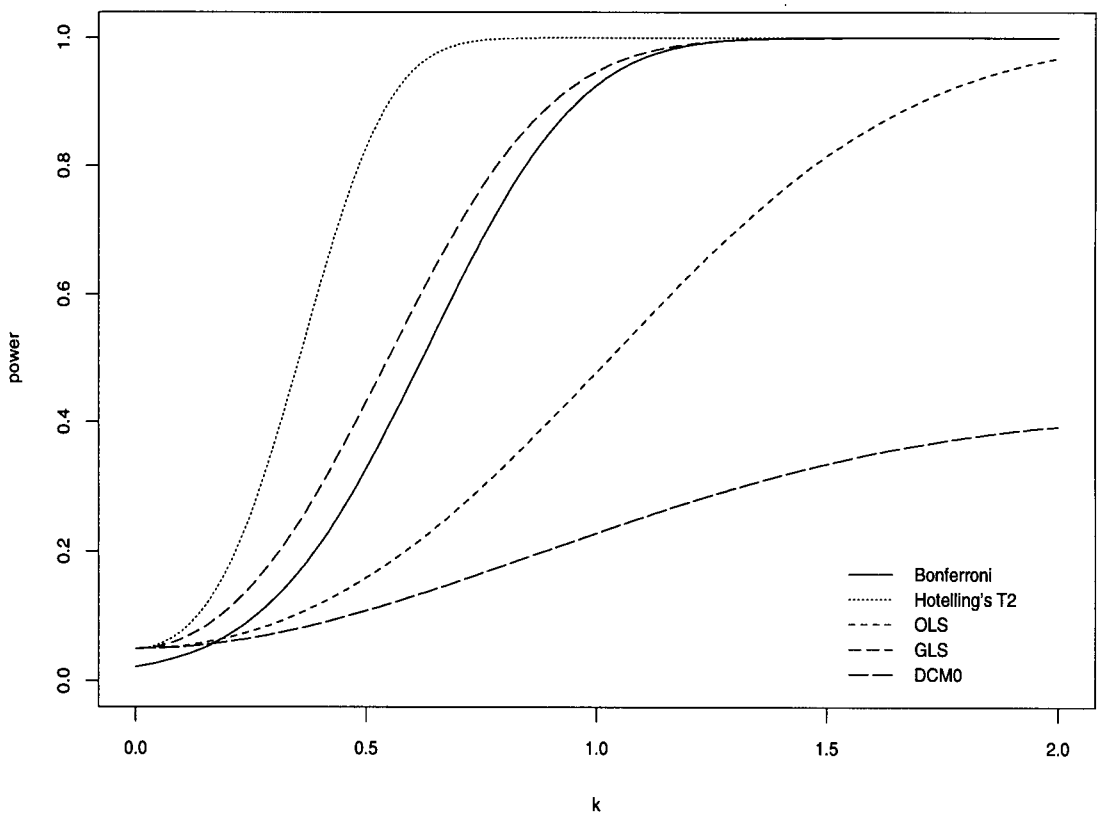
Table 4.26: Power of procedures with 100 patients per arm

Configuration				Procedure				
Δ_{EDSS}	Δ_{AMB}	Δ_{NBB}	Δ_{NI9HP}	Bon.	T^2	OLS	GLS	DCM^0
.50	.10	.40	−.10	.93	1.0000	.48	.95	.23
.50	.10	.40	.10	.93	1.0000	.64	.991	.68
.50	.10	.40	−−	.93	1.0000	.79	.99	.54
.50	−−	.40	−−	.95	.95	.97	.97	.80
.50	−−	−−	−−	.94	.94	.94	.94	.79
.50	.50	.50	.50	.97	.97	.99	.995	.87

Table 4.27: Sample size required to achieve power of 0.80

Configuration				Procedure				
Δ_{EDSS}	Δ_{AMB}	Δ_{NBB}	Δ_{NI9HP}	Bon.	T^2	OLS	GLS	DCM^0
.50	.10	.40	−.10	72	23	216	61	530
.50	.10	.40	.10	71	26	145	42	134
.50	.10	.40	−−	69	23	103	39	184
.50	−−	.40	−−	62	63	52	52	99
.50	−−	−−	−−	63	63	63	63	102
.50	.50	.50	.50	61	57	44	38	82

Figure 4.9: Power of procedures with 100 patients per arm when $\boldsymbol{\Delta} = k \cdot \boldsymbol{\Delta}_{base}$, where $\boldsymbol{\Delta}_{base} = (.50, .10, .40, -.10)$



.40, −.10) is used. Hotelling's T^2 is most powerful and requires substantially fewer patients to achieve an adequate power than other procedures; DCM^0 and OLS perform particularly poorly. GLS has a small advantage over Bonferroni adjustment. GLS is expected to perform more powerfully than OLS as the most heavily weighted outcome measure, EDSS, is most effective in comparing the two arms. In fact, the advantage of GLS over OLS is substantial as the weight GLS assigns to EDSS is more than twice that assigned by OLS. GLS requires less than $\frac{1}{3}$ as many patients as OLS to achieve a power of 0.80. With a power of .23, DCM^0 is not comparable.

We next consider the power of these procedures for $\boldsymbol{\Delta} = k \cdot \boldsymbol{\Delta}_{base}$, where $\boldsymbol{\Delta}_{base} =$ (.50, .10, .40, −.10). Figure 4.9 shows that the procedure based on Hotelling's T^2 is most powerful although for k greater than about 1.3, GLS and Bonferroni adjustment are comparable to T^2. For k less than about 1.3, T^2 has a modest advantage over GLS and Bonferroni adjustment and a substantial advantage over OLS and DCM^0. While neither OLS nor DCM^0 is competitive, the former has substantial advantage.

The directions of the standardized differences on the individual outcome measures in the configurations considered in Figure 4.9 are not consistent as NI9PH shows a detrimental treatment effect while the other outcomes show beneficial treatment effects. In Chapter 2, we noted the main limitation of Hotelling's T^2 is that it does take the direction of the treatment effects into account. Consequently, the advantage of Hotelling's T^2 shown in Figure 4.9 deserves some further examination. We want to examine if this advantage is a result of its limitation and therefore consider the outcome measure NI9HP with a beneficial treatment effect. The configuration to be considered is $\boldsymbol{\Delta} =$ (.50, .10, .40, .10) and the results are presented in Tables 4.26 and 4.27. Comparing to $\boldsymbol{\Delta} =$ (.50, .10, .40, −.10), T^2 is extremely sensitive for both configurations but it requires slightly fewer patients when $\boldsymbol{\Delta} =$ (.50, .10, .40, −.10). This illustrates our concern with the limitation of T^2. In contrast, OLS, GLS, and DCM^0 improve substantially when the

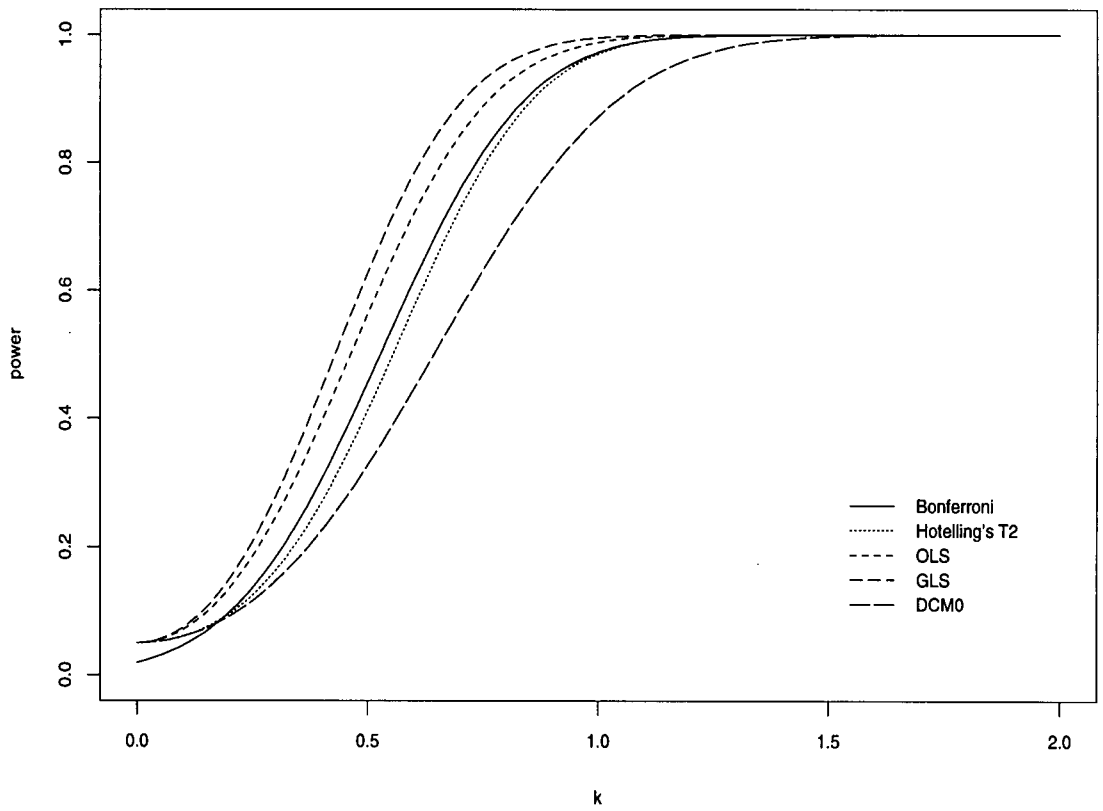
direction of the treatment effects are consistent. In this case, Bonferroni adjustment is only very little affected. It is worth pointing out that although Bonferroni adjustment also addresses the question of whether there is a difference between the two arms as Hotelling's T^2, the former requires one to assess the difference between the two arms for the individual outcomes and hence the direction of the difference on each outcome will be apparent when the analysis is carried out.

We next consider excluding the outcome measure NI9HP: $\boldsymbol{\Delta} = (.50, .10, .40, --)$. Comparing to $\boldsymbol{\Delta} = (.50, .10, .40, -.10)$, the results for this configuration show that dropping the outcome measure with detrimental treatment effects improves OLS and DCM^0 substantially, improves GLS slightly, and has essentially no impact on T^2.

Suppose now only the two most effective outcomes are included in the study; the configuration to be considered is $\boldsymbol{\Delta} = (.50, --, .40, --)$. Comparing to the configuration $\boldsymbol{\Delta} = (.50, .10, .40, -.10)$, all procedures except T^2 improve their performance upon excluding the two least effective outcomes. Note the dramatic improvement of the performance of DCM^0: it now requires less than 20% as many patients to achieve an adequate sensitivity (power of 0.80). Consequently, the choice of outcome measures also has a great impact on DCM^0. The improvement of the performance of OLS is also substantial. Comparing to the configuration of $\boldsymbol{\Delta} = (.50, .10, .40, --)$, both OLS and DCM^0 improve their performance substantially upon the exclusion of the weakly effective outcome.

We next consider the configuration where only the most effective outcome is included: $\boldsymbol{\Delta} = (.50, --, --, --)$. Comparing to the case where the two most effective outcomes are included, the results indicate that dropping a weakly correlated but reasonably effective outcome measure has a very small negative effect on all procedures. The correlation between EDSS and NBB is 0.35 and the two outcomes are reasonably effective; this is similar to Case C with two mildly correlated outcome measures considered in Chapter 2.

Figure 4.10: Power of procedures with 100 patients per arm when $\boldsymbol{\Delta} = k \cdot \boldsymbol{\Delta}_{base}$, where $\boldsymbol{\Delta}_{base} = (.50, .50, .50, .50)$



As illustrated by the results in Tables 2.1 and 2.2, using two weakly correlated outcome measures with equal effectiveness, is more effective than using only one of these outcomes. However, here we see an example where addition of a reasonably effective outcomes leads to only limited gain in sensitivity.

Finally, we consider a more optimistic configuration, where the four outcome measures are equally effective. Regarding $\boldsymbol{\Delta} = (.50, .50, .50, .50)$ as $\boldsymbol{\Delta}_{base}$, Figure 4.10 shows that when the magnitude of the effectiveness is large, say $k > 0.70$, all five procedures perform well but DCM^0 is still not comparable. When lesser magnitudes are considered, GLS

and OLS are clearly most powerful with the former having a slight advantage.

Summary

Imagine a clinical investigator planning a MS clinical trial with therapies having similar characteristics as those investigated in this study. Suppose further that s/he is willing to assume a common variance-covariance matrix for both populations, and is convinced that the observed standardized differences and the sample correlation structure of the placebo arm are the most relevant values. Our calculations show that the procedure based on T^2 is most powerful. However, one should be aware of the limitation of the procedure based on Hotelling's T^2. For the Oral Methotrexate data, the directions of the standardized differences are not consistent: three of the outcomes show beneficial treatment effects and the remaining outcome shows a detrimental treatment effect. Our example illustrates the limitation of Hotelling's T^2.

Suppose the researcher intends to use one of the five procedures in the design and analysis of a MS trial with four outcome measures. For the specified pattern of correlations among the four outcome measures, when all the outcomes are equally effective, GLS is most sensitive in the assessment of the relative efficacy of the two arms.

In addition, the results in Tables 4.26 and 4.27 illustrate the importance of the selection of outcome measures to be included in designing a study. Not surprisingly, the inclusion of an outcome measure with a detrimental treatment effect has a negative effect on the procedures (except T^2) although Bonferroni adjustment is not much affected. The inclusion of a weakly effective outcome measure can also have a negative impact on the performance of the procedures. Also, the gain in sensitivity from the addition of reasonably effective outcome measures can sometimes be quite limited. These results should encourage researchers to attempt to identify the best single outcome measure as the primary outcome measure for the design and analysis of clinical trials.

4.2.3 Another Disjunctive Composite Outcome Measure

So far, our discussion on the disjunctive composite outcome measure in this chapter has focused on DCM^0. We found that this approach is not as competitive with the others considered. In this subsection, we want to examine another definition of treatment failure which is related to the definition used in the original analysis of this data set. We first provide this definition of treatment failure; see Goodkin et al. (1992):

Definition 4.1 *Patients could meet treatment failure requirements for the disjunctive composite outcome measure in any of the following ways:*

1. *Worsening of the entry EDSS score by $\geq$1.0 point for patients with an entry score of 3.0-5.0 or by $\geq$0.5 point for those patients with an entry score of 5.5-6.5;*
2. *Worsening of the entry AMB score of 2-6 by $\geq$1.0 point;*
3. *Worsening of $\geq$20% from the baseline value on the best performance of two successive Box and Block or Nine Hole Peg test scores obtained with either hand.*

Changes on any the four components of this composite outcome measure had to be sustained for $\geq$2 months to be designated as treatment failure.

Note that the original definition of treatment failure also contained: the appearance of new or enlarged lesions on annual serial magnetic resonance imagine (MRI) scans. However, early in the study, it was decided to remove this dimension form the definition of treatment failure due to concerns regarding the potential contribution of measurement and repositioning error to what was assumed to represent disease activity.

As we have only the baseline and annual scores for each of these outcomes, we modify this definition of treatment failure for our purposes. The requirement that changes had to be sustained for $\geq$ 2 months is dropped. Second, the evaluation of successive scores

Table 4.28: Treatment failure rates based on DCM^D

Failure parameter	Placebo	Treatment
EDSS	.57	.39
Ambulation Index (AMB)	.35	.52
Box and Block Test (BB)	.39	.44
Nine Hole Peg Test (9HP)	.61	.44
(EDSS, AMB, BB, 9HP)	.87	.74
(EDSS, BB, 9HP)	.87	.65
(EDSS, 9HP)	.78	.57

on the Box and Block and Nine Hole Peg tests is dropped. In other words, for each of the Box and Block and Nine Hole Peg tests, the requirement becomes: worsening of $\geq$ 20% from the baseline value on the scores obtained with either hand. We will refer to the resulting procedure as DCM^D in what follows.

Table 4.28 presents the treatment failure rates for each of the outcome measures. According to our definition of treatment failure, 87% of the patients on the placebo arm and 74% on the treatment arm experienced treatment failure. (These compare to 83% and 52% according to the original definition based on the monthly data.)

We now take the sample treatment failure rates as the population treatment failure rates and evaluate the power and the required sample size for this disjunctive composite outcome measure. With $\pi_1 = .87$ and $\pi_2 = .74$, we find that with 100 patients per arm, the power of the procedure based on this composite outcome measure is 0.64 and the required sample size to achieve a power of 0.80 is 144 patients per arm.

We wish to compare the performance of DCM^D to other procedures. As the results of DCM^D were based directly on the data, it seems most reasonable to compare to the performance of the other procedures under the configuration most relevant to the

Table 4.29: Treatment failure rates based on DCM^0

Failure parameter	Placebo	Treatment
EDSS	.50	.31
AMB	.50	.46
NBB	.50	.34
NI9HP	.50	.54
(EDSS, AMB, NBB, NI9HP)	.76	.69
(EDSS, AMB, NBB)	.72	.58
(EDSS, NBB)	.69	.50

observed standardized differences, i.e. $\boldsymbol{\Delta} = (.50, .10, .40, -.10)$. First consider DCM^0. The results in Tables 4.26 and 4.27 show that DCM^D provides substantial improvement in performance over DCM^0. The treatment failure rates on the individual outcomes for DCM^0 are presented in Table 4.29. The results in Tables 4.28 and 4.29 show that the differences in the failure rates between the placebo and the treatment arms on Ambulation Index and Nine Hole Peg Test for DCM^D are substantially larger than for DCM^0, this difference on NBB for DCM^D is considerably smaller than for DCM^0, and the difference on EDSS is about the same for both procedures. Note that the directions of the differences in the failure rates are not consistent for either DCM^D or DCM^0. For DCM^D, the failure rate on AMB is considerably larger for the patients on the treatment arm and that on BB is slightly larger on the treatment arm whereas for DCM^0, the failure rate on NI9HP is slightly larger on the treatment arm. The net result is a moderately larger difference in the failure rates on the composite outcome measure DCM^D which leads to its substantially better performance. Comparing DCM^D to the other four procedures in Tables 4.26 and 4.27, we find that DCM^D has a clear advantage over OLS; DCM^D requires about 65% as many patients as OLS to achieve a power of 0.80. However,

DCM^D is still not competitive with Bonferroni adjustment, GLS and T^2.

Suppose we consider dropping less effective outcomes from DCM^D and DCM^0. First consider excluding the least effective outcome; that is, dropping AMB from DCM^D and NI9HP from DCM^0. Tables 4.28 and 4.29 present the failure rates of these new composite outcomes. For DCM^D, with $\pi_1 = .87$ and $\pi_2 = .65$, the power achieved with 100 patients per arm is substantially improved to 0.96 and it now requires only 59 patients per arm to achieve a power of 0.80. For DCM^0, with the exclusion of NI9HP, the power and the required sample size are now 0.54 and 184 respectively (with $\pi_1 = .72$ and $\pi_2 = .58$). Consequently, both procedures gain substantially from deleting the least effective outcome. Suppose now only the two most effective outcomes are included. DCM^D is negatively affected as its power decreases to 0.91 and n increases to 72 whereas DCM^0 improves as power = 0.80 and $n = 99$. This detrimental effect on DCM^D by dropping an outcome with a negative treatment effect is unexpected; an explanation requires further investigation. These results illustrate the detrimental effect on disjunctive outcome measures resulting from the inclusion of weakly effective outcomes and indicate the importance of the choice of outcomes in the design and analysis of a study. The potential of this type of outcome measure is revealed as well.

4.3 Discussion

In this Chapter, the five procedures discussed in the Chapters 2 and 3 were applied to two data sets from MS clinical trials. For the Task Force data, the three outcome measures are modestly and roughly equally correlated on both arms. The results in Tables 4.22 and 4.23 indicate that with this particular pattern of correlation structure, the performance of the procedures depends heavily on the configuration of the standardized differences. This confirms the findings based on idealized scenarios in Chapters 2 and 3 that the

anticipated configuration of standardized differences should play an important role in the selection of the procedure to be used for the design and analysis of the trial. For example, when all the outcome measures are equally effective in comparing the two arms, O'Brien's GLS is the best way to proceed. OLS has almost identical performance, but the other procedures are clearly inferior. On the other hand, when only a single outcome is effective, T^2 is most powerful. Bonferroni adjustment is reasonably competitive but the other procedures are clearly less sensitive. For intermediate cases, Bonferroni adjustment performs better. Therefore, it is essential for the clinical investigator to obtain as much information as possible on the characteristics of the outcome measures for the patient population to be studied. Without adequate knowledge, it is impossible to decide which of these statistical approaches to multiple outcome measures is most appropriate for the MS clinical trial being planned.

The Oral Methotrexate data provided several interesting features. The directions of the observed standardized differences on the individual outcome measures are not consistent as one outcome shows a detrimental treatment effect whereas the rest show beneficial treatment effects. The least correlated outcome measure is most effective in comparing the two arms. The results in Tables 4.26 and 4.27 illustrate the limitation of the procedure based on Hotelling's T^2 resulting from the fact that it does not address the question of whether one arm is better than the other. Also, the results indicate that for the specified correlation structure and configuration of the standardized differences, the procedure based on O'Brien's GLS is most appropriate.

For both data sets, DCM^0 is not competitive with the other procedures. However, to some extent this is due to the definition of treatment failure underlying DCM^0. For example, the alternate disjunctive outcome measure based on a definition of treatment failure related to that used in the original analysis of this data performs substantially better than DCM^0. This suggests the potential of the procedure based on this type of

composite outcome measure. It also indicates a difficulty of using this type of composite outcome measure: its performance depends heavily on the definition of treatment failure employed, but in most circumstances the most appropriate definition will not be obvious.

For both applications, we also considered several configurations to illustrate the effect of the exclusion of weakly effective or ineffective outcomes. The results indicate that when planning a study, researchers should pay particular attention to the selection of the outcome measures to be included as the inclusion of outcomes of little effectiveness or no effectiveness can have considerable negative impact on the sensitivity of these procedures for the assessment of the relative efficacy of the two arms. Also, addition of even reasonably effective outcomes sometimes adds little to the sensitivity. These results demonstrate the importance of effort in identifying the best single outcome measure when planning a study as the primary outcome measure for the design and analysis of clinical trials. If several primary endpoints must be included because it is not possible to identify the single best outcome, then, these results stress the extreme importance of identifying equally effective outcome measures for the assessment of each clinical dimension judged to be of importance in the clinical trial under consideration.

Chapter 5

Conclusion

In this thesis, five statistical methods for the design and analysis of clinical trials where the efficacy of a therapy is assessed by multiple outcome measures were compared. The results presented allow several general remarks.

First, the inclusion of ineffective or weakly effective outcome measures can result in a substantial penalty. Consequently, the selection of outcome measures to be used is very important. The results for equally correlated outcome measures show that the inclusion of ineffective outcome measures leads to detrimental effects on all the procedures. In this situation, identifying the best single outcome becomes essential. However, when it is not clear which outcome is effective, the results suggest that Bonferroni adjustment should be used as the impact of including ineffective outcomes on this procedure is smallest.

Second, our examples presented in Section 4.2.2 illustrate that results obtained using the procedure based on Hotelling's T^2 can be misleading as the inclusion of an outcome measure with a detrimental treatment effect leads to a smaller required sample size. Because T^2 does not take into account the directions of the treatment effects, it is not an appropriate procedure for the clinical trials context.

Third, procedures which combine the evidence provided by individual outcomes can be quite sensitive in the assessment of the relative efficacy of the two arms. The procedure based on O'Brien's GLS statistic shows its superiority in many of the settings considered. In particular, when several outcomes with roughly equal effectiveness are included, GLS is very sensitive.

On the other hand, there is potential danger in using the GLS procedure: Depending upon the correlation structure among the outcome measures, it is possible for GLS to perform very well or very poorly. Therefore, to determine the appropriateness of a particular procedure relies heavily on the researcher's knowledge of the outcome measures to be used. Without high quality information on the outcomes to be used, providing a specific recommendation on the most appropriate procedure for a particular MS clinical trial is impossible.

Although our results suggest that DCM is not comparable to the other procedures, this may be due to the limited scope of DCM considered. The results in Section 4.2.3 illustrate the potential of this method. The main advantage associated with DCM is its ease of handling longitudinal data as using the longitudinal data would presumably add sensitivity in the assessment of treatment efficacy. Its main difficulty is that there seems to be no obvious rules of constructing reliable pre-assigned cutoff values for the individual outcome measures. Constructing a clinical meaningful and statistically powerful disjunctive outcome measure requires the researcher to provide detailed and reliable information on the outcomes to be used. Overall, perhaps the most important message is that more empirical work on high quality information is essential to provide a better understanding of the properties of outcome measures in current use and the relationships among these outcome measures.

The discussion in this thesis has focused on the case of continuous and normal responses. For the Hotelling's T^2, OLS and GLS procedures, as long as the joint distribution of the vector of Z-statistics can be reasonably approximately by the multivariate normal distribution, these procedures can be applied and our numerical results are relevant.

Another limitation of our investigation is that we have assumed that the data to be analyzed are the changes in the responses from the baseline to the end of the trial. However, quite often outcome measures are recorded regularly throughout the period

of the study. Using the procedures based on Bonferroni adjustment, Hotelling's T^2, O'Brien's OLS and GLS to analyze such longitudinal data involve first summarizing the data by a suitable univariate descriptor. In other words, some of the information collected in the study is not used. In contrast, longitudinal data can easily be analyzed by DCM. This appears to be the main reason DCM was proposed and used in the original analysis of the Oral Methothexate data. Our results have illustrated the potential of O'Brien's GLS statistic in providing a sensitive assessment of treatment efficacy, so a procedure analogous to GLS but for longitudinal data certainly deserves future work.

Appendix A

The non-centrality parameter, denoted by λ^2, for the Hotelling's T^2 statistic for testing $H_0 : \boldsymbol{\Delta} = \mathbf{0}$ against $H_a : \boldsymbol{\Delta} = \boldsymbol{\Delta}^*$ is:

$$\lambda^2 = \frac{n}{2}(\boldsymbol{\Delta}^*)' \boldsymbol{M}_\rho^{-1} (\boldsymbol{\Delta}^*)$$

For the special case of equally correlated outcome measures, the correlation matrix is of the form:

$$\boldsymbol{M}_\rho = \begin{pmatrix} 1 & \rho & \rho & \cdots & \rho \\ \rho & 1 & \rho & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \rho & \rho & \cdots & 1 & \rho \\ \rho & \rho & \cdots & \rho & 1 \end{pmatrix} = (1-\rho)\boldsymbol{I} + \rho \boldsymbol{J},$$

where ρ is the common correlation among the outcomes. To simplify the expression for λ^2, re-express $\boldsymbol{M}_\rho$ as:

$$\boldsymbol{M}_\rho = (1-\rho)\left(\boldsymbol{I} + \frac{\rho}{1-\rho}\boldsymbol{J}\right)$$

We will need the following lemma before proceeding further:

Lemma A.1 *Let the $p \times p$ matrix $\boldsymbol{W}$ have the form*

$$\boldsymbol{W} = \boldsymbol{I} + a\boldsymbol{J}.$$

Then,

$$\boldsymbol{W}^{-1} = \boldsymbol{I} - \frac{a}{1+ap}\boldsymbol{J}.$$

Applying this result yields

$$\begin{aligned} \boldsymbol{M}_\rho{}^{-1} &= \frac{1}{(1-\rho)}\left(\boldsymbol{I} - \frac{\left(\frac{\rho}{1-\rho}\right)}{1+\left(\frac{\rho}{1-\rho}\right)m}\boldsymbol{J}\right) \\ &= \frac{1}{(1-\rho)}\left(\boldsymbol{I} - \frac{\rho}{1-\rho+m\rho}\boldsymbol{J}\right) \\ &= \frac{1}{(1-\rho)}\left(\boldsymbol{I} - \frac{\rho}{1+(m-1)\rho}\boldsymbol{J}\right), \end{aligned}$$

and we obtain:

$$\begin{aligned} \lambda^2 &= \frac{n}{2}(\boldsymbol{\Delta}^*)'\boldsymbol{M}_\rho{}^{-1}(\boldsymbol{\Delta}^*) \\ &= \frac{n}{2(1-\rho)}(\boldsymbol{\Delta}^*)'\left(\boldsymbol{I} - \frac{\rho}{1+(m-1)\rho}\boldsymbol{J}\right)(\boldsymbol{\Delta}^*) \end{aligned}$$

Appendix B

For a simple two-armed clinical trial with n patients on each arm, suppose the parameter of interest is the difference in the population means, $\mu_1 - \mu_2 = \delta$ say, and the common population variance, σ^2 say, is known. We would like to test $H_0 : \delta = 0$ against $H_a : \delta \neq 0$. At the end of the study, we estimate this parameter by the difference in the sample means, $\hat{\delta} = \overline{X}_1 - \overline{X}_2$. The expectation and variance of this estimator is:

$$E(\hat{\delta}) = \delta, \; Var(\hat{\delta}) = \frac{2}{n}\sigma^2.$$

By the Central Limit Theorem, for large n, the distribution of $\hat{\delta}$ can be approximated as $N(\delta, \frac{2}{n}\sigma^2)$. Therefore, the distribution of

$$Z = \frac{\hat{\delta} - \delta}{\sqrt{2\sigma^2/n}}$$

can be approximated as standard normal. To produce an approximate level α test, H_0 is rejected if $\mid \hat{\delta} \mid > z_{1-\alpha/2}\sqrt{2\sigma^2/n}$. The power of this test evaluated at $H_a : \delta = \delta^*$ is:

$$\begin{aligned}
Power_{\delta=\delta^*} &= P_{\delta=\delta^*}\left(\mid \hat{\delta} \mid > z_{1-\alpha/2}\sqrt{2\sigma^2/n}\right) \\
&= 1 - P_{\delta=\delta^*}\left(-z_{1-\alpha/2}\sqrt{2\sigma^2/n} \leq \hat{\delta} \leq z_{1-\alpha/2}\sqrt{2\sigma^2/n}\right) \\
&= 1 - P_{\delta=\delta^*}\left(\frac{-z_{1-\alpha/2}\sqrt{2\sigma^2 n} - \delta^*}{\sqrt{2\sigma^2/n}} \leq \frac{\hat{\delta} - \delta^*}{\sqrt{2\sigma^2/n}} \leq \frac{z_{1-\alpha/2}\sqrt{2\sigma^2 n} - \delta^*}{\sqrt{2\sigma^2/n}}\right) \\
&= 1 - P\left(-z_{1-\alpha/2} - \sqrt{\frac{n}{2}}\frac{\delta^*}{\sigma} \leq Z \leq z_{1-\alpha/2} - \sqrt{\frac{n}{2}}\frac{\delta^*}{\sigma}\right) \\
&\cong 1 - \Phi\left(z_{1-\alpha/2} - \sqrt{\frac{n}{2}}\frac{\delta^*}{\sigma}\right) + \Phi\left(-z_{1-\alpha/2} - \sqrt{\frac{n}{2}}\frac{\delta^*}{\sigma}\right)
\end{aligned}$$

If $\delta^* > 0$, then provided n is large,

$$\Phi\left(-z_{1-\alpha/2}-\sqrt{\frac{n}{2}}\frac{\delta^*}{\sigma}\right)\approx 0.$$

Therefore, we can approximate $Power_{\delta=\delta^*}$ by the upper tail probability only; that is,

$$Power_{\delta=\delta^*}\approx 1-\Phi\left(z_{1-\alpha/2}-\sqrt{\frac{n}{2}}\frac{\delta^*}{\sigma}\right).$$

The approximate sample size required to achieve power $1-\beta$ can be obtained by solving

$$1-\Phi\left(z_{1-\alpha/2}-\sqrt{\frac{n}{2}}\frac{\delta^*}{\sigma}\right)\approx 1-\beta$$

for n. This is equivalent to solving $z_{1-\alpha/2}-\sqrt{\frac{n}{2}}\frac{\delta^*}{\sigma}\approx z_\beta$ for n. This calculation yields:

$$n\approx\frac{2\sigma^2(z_{1-\alpha/2}-z_\beta)^2}{(\delta^*)^2}.$$

With $\Delta=\frac{\delta}{\sigma}$, the standardized difference of the population means, we can re-express the approximate power and the required sample size as:

$$\begin{aligned} Power_{\Delta=\Delta^*} &\cong 1-\Phi\left(z_{1-\alpha/2}-\sqrt{\frac{n}{2}}\Delta^*\right)+\Phi\left(-z_{1-\alpha/2}-\sqrt{\frac{n}{2}}\Delta^*\right) \\ &\approx 1-\Phi\left(z_{1-\alpha/2}-\sqrt{\frac{n}{2}}\Delta^*\right) \quad for\ \delta^*>0, \end{aligned}$$

and

$$n\approx\frac{2(z_{1-\alpha/2}-z_\beta)^2}{(\Delta^*)^2}.$$

Appendix C

Here, we want to show that when the outcome measures are equally correlated, O'Brien's OLS and GLS statistics are equivalent. As already shown in Appendix A, for equally correlated outcome measures, the correlation matrix has the form:

$$\boldsymbol{M}_\rho = (1-\rho)\boldsymbol{I} + \rho\boldsymbol{J},$$

and

$$\boldsymbol{M}_\rho{}^{-1} = \frac{1}{(1-\rho)}\left(\boldsymbol{I} - \frac{\rho}{1+(m-1)\rho}\boldsymbol{J}\right).$$

With this expression, we can proceed to show the equivalence of $\widehat{\boldsymbol{\beta}}_{OLS}$ and $\widehat{\boldsymbol{\beta}}_{GLS}$:

$$\widehat{\boldsymbol{\beta}}_{OLS} = \frac{Y_1 + Y_2 + \ldots + Y_m}{m},$$

$$\widehat{\boldsymbol{\beta}}_{GLS} = (\mathbf{1}'\boldsymbol{M}_\rho{}^{-1}\mathbf{1})^{-1}\mathbf{1}'\boldsymbol{M}_\rho{}^{-1}\boldsymbol{Y}.$$

Let $a = \frac{\rho}{1+(m-1)\rho}$ for simplification. Then,

$$\mathbf{1}'\boldsymbol{M}_\rho{}^{-1}\mathbf{1} = \frac{m(1-ma)}{(1-\rho)},$$

and

$$\mathbf{1}'\boldsymbol{M}_\rho{}^{-1}\boldsymbol{Y} = \frac{(1-ma)}{(1-\rho)}\sum_{j=1}^{m} Y_j.$$

Therefore,

$$\widehat{\boldsymbol{\beta}}_{GLS} = \frac{(1-\rho)}{m(1-ma)}\frac{(1-ma)}{(1-\rho)}\sum_{j=1}^{m} Y_j = \frac{\sum_{j=1}^{m} Y_j}{m} = \widehat{\boldsymbol{\beta}}_{OLS}.$$

Appendix D

In many clinical trials, the parameter of interest is the difference between two population proportions, $\pi_1 - \pi_2 = \theta$ say. Suppose that one has available independent binomial samples of size n with probability of success π_1 for the placebo arm and π_2 for the treated arm. We would like to test $H_0 : \pi_1 = \pi_2 = \pi$ say, against $H_a : \pi_1 \neq \pi_2$. We estimate π_1 and π_2 by the sample proportions, p_1 and p_2, and

$$E(p_1) \doteq \pi_1,\ Var(p_1) = \frac{\pi_1(1-\pi_1)}{n},$$

$$E(p_2) = \pi_2,\ Var(p_2) = \frac{\pi_2(1-\pi_2)}{n}.$$

By the normal approximation, for large n,

under H_0: $\pi_1 - \pi_2 = 0,\ \ p_1 - p_2 \approx N(0, \frac{2\pi(1-\pi)}{n})$.

under H_a: $\pi_1 - \pi_2 = \theta,\ p_1 - p_2 \approx N(\theta, \frac{\pi_1(1-\pi_1)+\pi_2(1-\pi_2)}{n})$.

Estimating π by $\overline{p} = \frac{p_1+p_2}{2}$, the Z-statistic for this test is

$$\frac{p_1 - p_2}{\sqrt{2\overline{p}(1-\overline{p})/n}}.$$

To produce an approximate level α test, H_0 is rejected if $\mid \frac{p_1-p_2}{\sqrt{2\overline{p}(1-\overline{p})/n}} \mid > z_{1-\alpha/2}$.

The power of this test can be evaluated as follows:

$$\begin{aligned} Power_{\pi_1-\pi_2=\theta} &= P_{\pi_1-\pi_2=\theta}\left(\mid p_1 - p_2 \mid > z_{1-\alpha/2}\sqrt{2\overline{p}(1-\overline{p})/n}\right) \\ &= 1 - P_{\pi_1-\pi_2=\theta}\left(-z_{1-\alpha/2}\sqrt{2\overline{p}(1-\overline{p})/n} \leq p_1 - p_2 \leq z_{1-\alpha/2}\sqrt{2\overline{p}(1-\overline{p})/n}\right) \\ &= 1 - P\left(a \leq \frac{(p_1-p_2)-\theta}{\sqrt{\pi_1(1-\pi_1)+\pi_2(1-\pi_2)/n}} \leq b\right) \end{aligned}$$

where $a = \frac{-z_{1-\alpha/2}\sqrt{2\overline{p}(1-\overline{p})/n}-\theta}{\sqrt{\pi_1(1-\pi_1)+\pi_2(1-\pi_2)/n}}$ and $b = \frac{z_{1-\alpha/2}\sqrt{2\overline{p}(1-\overline{p})/n}-\theta}{\sqrt{\pi_1(1-\pi_1)+\pi_2(1-\pi_2)/n}}$.

If n is large, $\frac{(p_1-p_2)-\theta}{\sqrt{\pi_1(1-\pi_1)+\pi_2(1-\pi_2)/n}}$ approximately follows the standard normal distribution. Also for large n, $\overline{p}$ approaches $\frac{\pi_1+\pi_2}{2} = \overline{\pi}$ in probability. It follows that $Power_{\pi_1-\pi_2=\theta}$ can be approximated by:

$$\begin{aligned} Power_{\pi_1-\pi_2=\theta} \approx\ & 1-\Phi\left(z_{1-\alpha/2}\sqrt{\frac{2\overline{\pi}(1-\overline{\pi})}{\pi_1(1-\pi_1)+\pi_2(1-\pi_2)}}-\frac{\sqrt{n}\theta}{\sqrt{\pi_1(1-\pi_1)+\pi_2(1-\pi_2)}}\right) \\ & +\Phi\left(-z_{1-\alpha/2}\sqrt{\frac{2\overline{\pi}(1-\overline{\pi})}{\pi_1(1-\pi_1)+\pi_2(1-\pi_2)}}-\frac{\sqrt{n}\theta}{\sqrt{\pi_1(1-\pi_1)+\pi_2(1-\pi_2)}}\right). \end{aligned}$$

Because $2\overline{\pi}(1-\overline{\pi}) = \pi_1(1-\pi_1)+\pi_2(1-\pi_2)+\frac{\theta^2}{2}$, this can be re-expressed as:

$$\begin{aligned} Power_{\pi_1-\pi_2=\theta} \approx\ & 1-\Phi\left(z_{1-\alpha/2}\sqrt{\frac{2\overline{\pi}(1-\overline{\pi})}{2\overline{\pi}(1-\overline{\pi})-\theta^2/2}}-\frac{\sqrt{n}\theta}{\sqrt{2\overline{\pi}(1-\overline{\pi})-\theta^2/2}}\right) \\ & +\Phi\left(-z_{1-\alpha/2}\sqrt{\frac{2\overline{\pi}(1-\overline{\pi})}{2\overline{\pi}(1-\overline{\pi})-\theta^2/2}}-\frac{\sqrt{n}\theta}{\sqrt{2\overline{\pi}(1-\overline{\pi})-\theta^2/2}}\right). \end{aligned}$$

If $\theta > 0$, then for large n, $\Phi\left(-z_{1-\alpha/2}\sqrt{\frac{2\overline{\pi}(1-\overline{\pi})}{2\overline{\pi}(1-\overline{\pi})-\theta^2/2}}-\frac{\sqrt{n}\theta}{\sqrt{2\overline{\pi}(1-\overline{\pi})-\theta^2/2}}\right) \approx 0$. We can therefore approximate $Power_{\pi_1-\pi_2=\theta}$ by the upper tail probability only:

$$Power_{\pi_1-\pi_2=\theta} \approx 1-\Phi\left(z_{1-\alpha/2}\sqrt{\frac{2\overline{\pi}(1-\overline{\pi})}{2\overline{\pi}(1-\overline{\pi})-\theta^2/2}}-\frac{\sqrt{n}\theta}{\sqrt{2\overline{\pi}(1-\overline{\pi})-\theta^2/2}}\right).$$

Solving the equation

$$z_{1-\alpha/2}\sqrt{\frac{2\overline{\pi}(1-\overline{\pi})}{2\overline{\pi}(1-\overline{\pi})-\theta^2/2}}-\frac{\sqrt{n}\theta}{\sqrt{2\overline{\pi}(1-\overline{\pi})-\theta^2/2}} \approx z_\beta$$

for n yields the approximate required sample size:

$$n \approx \frac{\left(z_{1-\alpha/2}\sqrt{2\overline{\pi}(1-\overline{\pi})}-z_\beta\sqrt{2\overline{\pi}(1-\overline{\pi})-\frac{1}{2}\theta^2}\ \right)^2}{\theta^2}.$$

Bibliography

[1] Follmann, D. (1995). Multivariate tests for multiple endpoints in clinical trials. *Statistics in Medicine* **14**, 1163-1175.

[2] Gibbons, J.D. (1971). *Nonparametric Statistical Inference.* McGraw-Hill, New York.

[3] Goodkin, D.E. and Rudick, R.A. (Eds)(1996). *Multiple Sclerosis: Advances in Clinical Trial Design, Treatment and Future Perspectives.* Springer-Verlag, London.

[4] Goodkin, D.E., Rudick, R.A., VanderBrug, M.S. et al. (1992). Low-dose (7.5 mg) oral methotrexate for chronic progressive multiple sclerosis: design of a randomized, placebo-controlled trial with sample size benefits from a composite outcome variable including preliminary data on toxicity. *Online Journal of Current Clinical Trials* [serial online] Document No. 19.

[5] Goodkin, D.E., Rudick, R.A., VanderBrug, M.S. et al. (1995). Low-dose (7.5 mg) oral methotrexate reduces the rate of progression in chronic progressive multiple sclerosis. *Annals of Neurology* **37**, 30-40.

[6] Kurtzke, J.F. (1983). Rating neurologic impairment in multiple sclerosis: an expanded disability scale (EDSS). *Neurology* **33**, 1444-1452.

[7] Johnson, N.L. and Kotz, S. (1972). *Distributions in Statistics: Continuous Multivariate Distributions.* John Wiley and Sons, New York.

[8] Joe, H. (1995). Approximations to multivariate normal rectangle probabilities based on conditional expectations. *Journal of the American Statistical Association* **90**, 957-964.

[9] Johnson, R.A. and Wichern, D.W. (1982). *Applied Multivariate Statistical Analysis.* Prentice Hall, New Jersey.

[10] Miller, R.G. (1981). *Simultaneous Statistical Inference (2nd edition).* Springer-Verlag, New York.

[11] O'Brien, P.C. (1984). Procedures for comparing samples with multiple endpoints. *Biometrics* **40**, 1079-1087.

[12] Petkau, A.J. (1996). Statistical and design considerations for multiple sclerosis clinical trials. Chapter 4 in: *Multiple Sclerosis: Advances in Clinical Trial Design, Treatment and Future Perspectives.* Goodkin, D.E. and Rudick, R.A. (Eds) Springer-Verlag, London, 63-103.

[13] Pocock, S.J., Geller, N.L. and Tsiatis, A.A. (1987). The analysis of multiple endpoints in clinical trials. *Biometrics* **43**, 487-498.

[14] Rudick, R.A., Antel, J, Confavreux, C. et al. (1996). Clinical outcomes assessment in multiple sclerosis. *Annals of Neurology* **40**, 469-497.

[15] Rudick, R.A., Antel, J, Confavreux, C. et al. (1997). Recommendations from the National Multiple Sclerosis Society Clinical Outcomes Assessment Task Force. *Annals of Neurology* **42**, 379-382.

[16] Tang, D., Geller, N.L. and Pocock, S.J. (1993). On the design and analysis of randomized clinical trials with multiple endpoints. *Biometrics* **49**, 23-30.