AN ADAPTIVE CLINICAL TRIAL DESIGN FOR A SENSITIVE SUBGROUP EXAMINED IN THE MULTIPLE SCLEROSIS CONTEXT

by

Corinne Aileen Riddell

BMath, The University of Waterloo, 2009
BBA, Wilfrid Laurier University, 2009

A THESIS SUBMITTED IN PARTIAL FUFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in

The Faculty of Graduate Studies

(Statistics)

THE UNIVERSITY OF BRITISH COLUMBIA
(Vancouver)

April 2011

Abstract

Adaptive clinical trials are recently gaining more attention. In this thesis, generalizations to the Biomarker-Adaptive Threshold Design (BATD) are studied and applied in the multiple sclerosis (MS) context. The BATD was originally developed for survival outcomes for Phase III clinical trials and allows researchers to both study the efficacy of treatment in the overall group and to investigate the relationship between a hypothesized predictive biomarker and the treatment effect on the primary outcome. We first introduce the original methodology and replicate the authors' simulation studies to confirm their findings. Then, we generalize the methodology to accommodate count biomarkers and outcomes. Our interest in variables of this form is fuelled by the study of MS, where the number of relapses is a commonly used count outcome for patients with relapsing-remitting MS. Through simulation studies, we find that the BATD has increased power compared with a traditional fixed design under varying scenarios for which there exists a sensitive patient subgroup. As an illustrative example, we consider data from a previously completed trial and apply the methodology for two hypothesized markers: baseline lesion activity and the length of time that a patient has had MS. While we do not find a predictive biomarker relationship between baseline lesion activity and the number of relapses, MS duration does appear to have a predictive biomarker relationship for this dataset. In particular, we consider a randomly chosen subsample of the data for which the overall treatment effect on the outcome was insignificant. When the BATD is applied, a very significant treatment effect is detected and indicates that the effect is strongest for patients that have had MS for less than 7.8 years for this subsample. The methodology holds promise at preserving statistical power when the treatment effect is greatest in a sensitive patient subset.

Table of Contents

List of Tables

List of Figures

Acknowledgements

My time at UBC has been one of enormous personal and professional growth. Although I have spent less than two years in Vancouver, it has become my second home. My admiration for this place is influenced largely by the people I have met here, including my colleagues, advisors, and friends. However, I cannot forget to acknowledge the support I have constantly received which resonates from my home base in Ridgeway, Ontario – my family and friends there have remained true and have provided great advice and support over some of the more difficult times, and congratulations in times of success. I cannot be more thankful for you all.

I'd like to acknowledge my colleagues. Sky, you've been a great office mate. You are brilliant, hard-working, and hilarious. It is a testament to our friendship that we have to study in the office at different times in order to get any serious amount of work done! Eric, you were always brutally honest with me and taught me how to not sweat the small stuff – this is a very valuable lesson that is going to save me a lot of needless worry in the future. There are countless other colleagues to acknowledge, all of whom took time out of their busy schedules to help me with the more technical aspects of using software and the servers. Thank you all – the combination of these contributions had a great impact on my thesis.

I would also like to thank the support staff within the Department – both technical and administrative. Peggy, Elaine, and Andrea: thank you for your help with all the little things. Without your patience and care, this experience could have been a lot more stressful. The technical staff also took time out of their schedules to help me use the servers more efficiently. Both I and the other server users thank you!

I'd like to acknowledge the MS/MRI Research Group and the faculty whom I've worked with. The MS/MRI research group has been very welcoming and has watched me grow under their guidance and support. Through this collaboration, I've become a more confident researcher. I've had the great benefit of being co-supervised by Drs. John Petkau and Yinshan Zhao. Yinshan's patience and guidance made both my RA and thesis work that much more enjoyable. She herself is evidence of the high quality researchers that are the progeny of this department. I am very fortunate to have both been John's student and to have conducted research with him. He has been a great role model and embodies all the characteristics of the statistical consultant, researcher collaborator, and teacher that I aspire to be. Lastly, Dr. Paul Gustafson has been very supportive during my studies and has provided an additional source of guidance and support. Thank you, as well, for being my second reader and making valuable contributions to my thesis.

To my brother Justin

1 Introduction

Chapter 1 begins with an explanation of the motivation for adaptive designs. This is followed by a brief review of adaptive designs that have been introduced in the literature, with a focus on those which incorporate knowledge or hypotheses regarding predictive biomarkers. The chapter concludes by introducing the Biomarker Adaptive Threshold Design that is the focus of this thesis.

1.1 Motivation for Adaptive Designs

Clinical trials which have a fixed design and analysis plan are the standard used by pharmaceutical companies in the drug development process. However, the Food and Drug Administration (FDA) has recently begun to approve and offer guidance on adaptive designs, which they define as, "...a study that includes a prospectively planned opportunity for modification of one or more specified aspects of the study design..." (U.S. Department of Health and Human Services. Food and Drug Administration, 2010, p. 6). There is now an increasing need for the exploration of the statistical properties of adaptive designs.

Trials with a fixed design, although well-understood, have many practical concerns. For example, doctors face ethical concerns when they encourage their patients to enrol in a clinical trial if they have strong beliefs regarding the efficacy of one of the treatment arms, especially if they believe the alternative treatment may induce patient suffering. Clinical trials are also extremely expensive, often costing millions of dollars to administer. If only a portion of those who are treated receive benefit, then the trial may be overly expensive while causing unnecessary patient burden. Ethical and financial burdens are just two of the concerns with fixed designs. The goal of many adaptive designs is to improve the design of trials in order to address these types of concerns while maintaining desirable statistical properties.

Overall, the use of adaptive designs has many advantages over traditional or fixed designs. Cost savings often result with adaptive designs since they may require shorter study periods, thereby decreasing patient follow-up and associated costs. Some adaptive designs require fewer study participants compared with their traditional alternatives, again leading to a reduction in costs. As well, so-called seamless adaptive designs which combine clinical trial phases in one study can have enormous cost savings when done well.

Beyond the large cost savings to the companies funding the trials, the patient populations can also benefit from the use of adaptive trials. For patients already enrolled in a trial, follow-up time may be reduced, limiting the patient burden associated with the trial (such as adverse side effects of treatment, or uncomfortable aspects associated with follow-up visits). Adaptive designs which drop treatment arms that appear less effective, or modify enrolment to favour treatment arms which are performing better also benefit the patient. Done well, these trials can limit the chance of enrolling patients in the less effective treatment arms, which increases the chance that a patient will benefit directly (i.e. on an individual level) from the trial.

While large cost savings and direct patient benefits are desirable reasons to implement adaptive trials, they cannot serve as the only reasons. The primary objective of a clinical trial is to establish whether the treatment in question is beneficial to the patient population under investigation. To do this, the statistical framework underlying the trial must be well-understood and uphold desired properties, such as tolerable Type I and Type II error rates. Without such features, investigators increase the risk of accepting treatments which may not be truly effective or rejecting treatments which may indeed benefit the patient population.

1.2 A Brief Review of Adaptive Designs

There are many types of adaptive designs, some of which are commonly used, and others which are beginning to gain more recognition. In a review of the FDA Draft Guidance for Industry Report (U.S. Department of Health and Human Services. Food and Drug Administration, 2010), Cook and DeMets (2010) describe commonly used designs, including dose escalation studies and designs that allow for interim monitoring. Dose escalation studies are used during early phase clinical trials. They involve giving an initial group of patients a defined dose of a drug and observing some outcome which is indicative of toxicity. Based on the observed response, the researchers will decide whether the dose should be escalated for the next patient that enters the study. This process is continued until the maximum tolerated dose is found (Chang, 2008).

Designs which allow for interim monitoring have been well-incorporated into pivotal trials. Interim monitoring is typically conducted by a data and safety monitoring board (DSMB). One of the responsibilities of the DSMB is to ensure the safety of the patients enrolled in the trial. The DSMB analyses data from the trial at pre-specified time-points to assess if safety concerns are present. Based on this information, trials may be stopped early due to benefit, toxicity, or futility (Ellenberg, Fleming, & DeMets, 2003). In any of these cases, continuation of the trial to the pre-planned conclusion may be considered unethical.

Other types of designs are less well-known, understood, or used in practise. However, some types of adaptive designs have been gaining more attention for a variety of reasons. Firstly, government interest in adaptive designs, as shown by the FDA providing their draft guidance report (2010), is an indicator to researchers that these non-traditional types of trials are a valid avenue for research. Secondly, a shift towards providing personalized medicine has taken hold. Providing patients with an optimal treatment regime which is customized according to their particular attributes is very desirable. This trend goes hand-in-hand with the growing research area of genomics. In many diseases, particular genotypes or genomic sequences have been linked to predispositions for the diseases. Finding genetic or other types of markers of treatment effectiveness is another area of research. However, pinpointing these so-called predictive biomarkers requires a large amount of data and often uses data from already completed clinical trials. For all of these reasons, many new types of adaptive designs are being introduced, studied and considered for implementation.

Cook and DeMets (2010) draw attention to the importance of distinguishing between two types of goals: the goals of a clinical trial and the goals of adaptation. For example, the clinical trial goal of a Phase III trial, to establish treatment efficacy, will always be best met through the use of a randomized controlled trial (RCT). However, the traditional RCT design does not consider burdens associated with the trial, such as the financial or patient burdens discussed earlier. The goal of adaptation, for some adaptive designs, is to reduce these burdens. For other adaptive designs, the exploration and discovery of other types of scientific information are the adaptation goals.

In the cancer literature, Simon (2010) reviews designs which combine the goals of the clinical trial with some other scientific goal (the adaptive goal) . In particular, Simon considers adaptive designs which incorporate information on a biomarker, or a set of biomarkers into the trial design. The reviewed designs are wide-ranging in the amount of already known biomarker knowledge required to use the designs. Enrichment designs involve the most extensive level of advanced knowledge of the biomarker. This includes knowledge of a test based on the biomarker to subset patients into test positive patients who are considered very likely to benefit from the treatment and test negative patients, considered very unlikely to benefit from treatment. For these designs, only patients in the test positive group are randomized, while test negative patients are not studied further because they are believed unlikely to receive benefit from the treatment. Obviously, enrichment trials involve extensive previous information on the biomarker and its relationship to the treatment under consideration. At the other end of the spectrum, the Adaptive Signature Design (Freidlin, Jiang, & Simon, 2010) allows for the development of classifiers as the adaptation goal. Simon (2010) remarks on the strengths and weaknesses of each of the designs. One weakness of designs that have ambitious adaptation goals includes the limitations in power associated with detecting a treatment effect. Generally speaking, one must balance the two types of goals or there will be limited information to address the clinical trial goal, which should arguably remain the overarching objective.

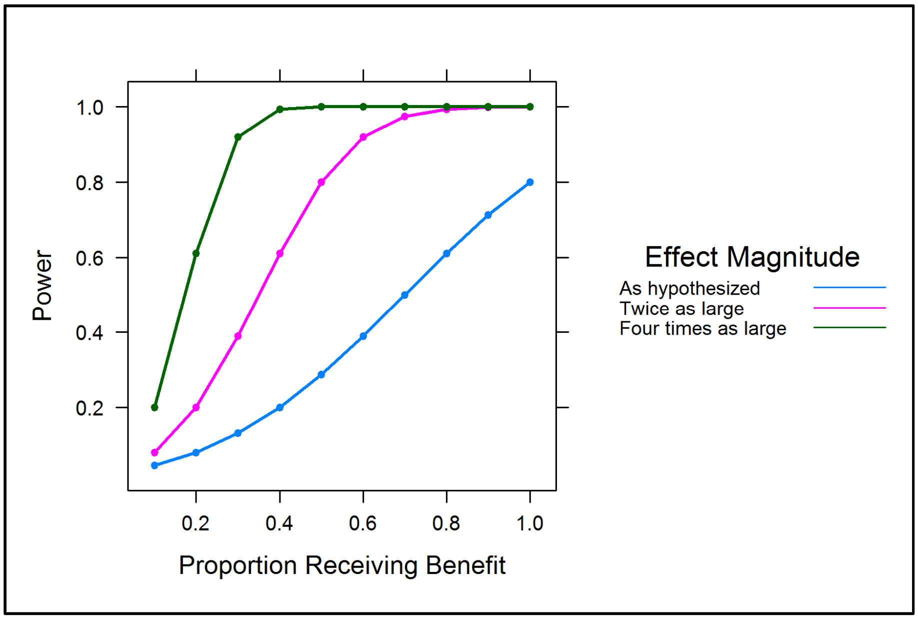
Also in cancer literature, Mandrekar and Sargent (2009) review the features and challenges of designs which specifically incorporate predictive biomarker validation. A predictive biomarker is a biological marker which provides information regarding a patient's response to a given treatment. Knowledge of predictive biomarkers is desirable as it allows treatment to be focused on the group of patients that are most likely to benefit. The objective of this thesis involves the investigation of one such design, namely the Biomarker-Adaptive Threshold Design first proposed by Jiang, Freidlin, and Simon (2007). Hereafter, this paper will be referred to as JFS for brevity.

In clinical trials in multiple sclerosis (MS), the more common designs are used, but there is a lack of research on novel methods as applied to this chronic disease. Multiple sclerosis is a very heterogeneous condition; the visible faces of MS cover a vast span of clinical symptoms and levels of disability, and the invisible faces as measured by silent indicators of disease using magnetic resonance images (MRIs) are highly variable as well. Because of this heterogeneity, it may well be that patients' responses to treatment differ based on their particular blend of symptoms and attributes. In a review on the development of biomarkers in MS, Bielekova and Martin (2004, pp. 1466-7) note that, "there is a strong need for developing new therapies in multiple sclerosis that are more process-specific and can be used in specific patient subpopulations...". This statement emphasizes the need for predictive biomarkers in MS and correspondingly for designs which are related to the development of such markers. Considering the Biomarker-Adaptive Threshold Design in the MS context will provide an introduction to MS clinicians of adaptive designs that incorporate biomarker validation.

1.3 Introduction to the Biomarker Adaptive Threshold Design

Unlike many types of adaptive designs which involve adaptation during the trial, the Biomarker-Adaptive Threshold Design (BATD) involves adaptation only during the analysis at the end of the study. It is a procedure for evaluating treatments which may provide greater benefit to a proportion of the population compared with the overall group. As an illustrative example, consider a clinical trial designed to detect a 10% reduction in the mean of a hypothetical outcome. This outcome is continuous and it is anticipated that untreated patients would have an average value of 1,000. Assuming a standard deviation of 300, a Type I error rate of 5% and a power of 80%, a two sample t-test with a two-sided hypothesis will be used to investigate the difference in the means between the two groups. Using this test with the stated target error rates, the sample size required per arm is 141 patients. If this 10% reduction is confined to a sensitive patient subset, then the realized power of the trial is reduced if only 282 patients are included in the study. The blue curve in Figure 1.1 illustrates how the power changes if the treatment effect is confined to a sensitive patient subset. It can be contrasted with the other curves depicting the power of the test for effect magnitudes larger than originally anticipated across different sizes of sensitive subsets.

Figure 1.1: Power of the t-test for treatment effect across varying effect magnitudes and proportions of the study population benefitting.



The statistical power drops dramatically as the treatment effect is confined to smaller subsets. This results in a higher probability that the treatment effect would be deemed insignificant and the trial would conclude that the treatment is ineffective for the studied population, whereas in fact a subset of the population did indeed benefit.

Now, if the researchers believed in advance that there exists a biological marker which defines a subset of patients that may benefit more from the treatment than the overall group, then the BATD incorporates this knowledge into the planning of the study's analysis. The BATD's analysis strategy is designed to identify if the treatment is effective for the group as a whole, and if not, to see if a treatment effect can be identified by looking for it in subsets of the data according to patients' values for the biomarker. As a secondary goal, this design estimates the level of the suggested biomarker at which the treatment becomes effective, if a sensitive patient subgroup appears to be present. In practise, it can provide a more efficient design by eliminating the need for multiple, expensive clinical trials, by combining the investigations into a single trial.

1.4 Goals of This Thesis

JFS present methodology and simulation results applied to a continuous biomarker and a time-to-event response. Chapter 2 of this thesis presents the JFS methodology and verifies their simulation findings. In Chapter 3, the JFS methodology is generalized so that the design can be applied to scenarios in which the biomarker and response are count variables. The statistical power of the adaptive method is investigated using a simulation study and contrasted with that of a traditional fixed analysis under varying circumstances. In Chapter 4, the generalization is then applied to a dataset from the multiple sclerosis literature to illustrate how the method could be applied and the potential benefits of its application in specific circumstances. Chapter 5 presents a brief discussion of further possible extensions to the BATD. Finally, Chapter 6 concludes this thesis.

2 The Biomarker Adaptive Threshold Design

The BATD was introduced by Jiang, Freidlin, and Simon (2007). In their paper, the authors develop methodology for a survival outcome and show results from a simulation study to compare the performance of their adaptive procedures and a traditional design. They apply their methodology to a clinical trial dataset in which the primary outcome was the survival time of patients with prostate cancer. The original study of this dataset found no treatment effect. However, when the BATD is used during the analysis, a significant treatment effect is found, suggesting the presence of a sensitive patient subgroup that is benefitting more from treatment than the studied population on the whole. In this chapter, we first describe the JFS methodology. Then, the results from my own simulation study are presented.

2.1 Methodology Employed by JFS

JFS consider three approaches to determine whether there is a treatment effect. The first approach, which they refer to as the Overall Effect Test (OET), is the traditional (non-adaptive) approach that would be used in the RCT setting in which there are only two arms, treatment and placebo, with the goal of estimating the treatment effect and finding its statistical significance. Here, the authors focus on primary endpoints that are survival outcomes. The adaptive approaches, called Procedure A and Procedure B, are modifications of the traditional approach which take into account the biomarker value of a patient. In this section, we describe each approach in turn, but first we present an overview of survival analysis and the most commonly used semi-parametric model of survival data.

2.1.1 Introduction to Survival Analysis and the Cox Proportional Hazards Model

Survival analysis refers to the analysis of time to event responses. In RCTs, examples of common survival responses include the time until death in cancer studies, or the time until a clinical relapse in multiple sclerosis studies. A common feature of survival data is that the exact survival time is not known for a proportion of individuals. This is due to a phenomenon known as censoring. In RCTs, censoring occurs for a variety of reasons, such as the study ending before a patient has the event of interest, or the patient being lost-to-follow-up while the study is still running. In both of these instances, we only know that the patient has “survived” up until the end of the study (in the first example) and until their last visit (in the second example). Thus, a survival response is either known exactly and is considered uncensored, or a minimum of the patient’s survival time is known and is considered right-censored. In some settings, the data can also be left-censored, but this type of data is not common in RCTs so it will not be discussed further here.

In survival analysis, two functions are commonly used to characterize the survival of the population of interest. These are the survival and hazard functions. The survival function, *S(t)*, denotes the probability that a patient will survive past the time *t*:

$$S(t) = P(T > t)\ . \tag{2.1}$$

The basic properties of the survival function include:

- It is non-increasing as a function of *t*
- $S(t=0)=1$
- $S(t=\infty)=0$

The hazard function, *h(t)*, denotes the instantaneous risk of death, given that the individual has already survived to time *t*:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \le T < t + \Delta t \mid T \ge t)}{\Delta t}\ . \tag{2.2}$$

While the survival function focuses on the chance that an individual will survive past a certain time, the hazard function is concerned with an individual's risk of death at an instance in time. As the hazard function is a function of time, one can imagine how this function would differ across individuals in different states of health. For example, if we are interested in the time until death, a very sick patient with a serious prognosis will likely have an increasing hazard function, a patient in recovery will likely have a decreasing hazard function, and a supposed "healthy" individual will have a flat hazard function, at least in the short-term (Kleinbaum & Klein, 2005).

Using a little bit of algebra the relationship between the survival and hazard function can be found:

$$
\begin{aligned}
h(t) &= \lim_{\Delta t \to 0} \frac{P(t \le T < t + \Delta t \,|\, T \ge t)}{\Delta t} \\
&= \lim_{\Delta t \to 0} \frac{P(t \le T < t + \Delta t \ \cap \ T \ge t)}{\Delta t} \cdot \frac{1}{P(T \ge t)} \\
&= \lim_{\Delta t \to 0} \frac{P(t \le T < t + \Delta t)}{\Delta t} \cdot \frac{1}{P(T \ge t)} \\
&= f(t) / S(t).
\end{aligned}
\tag{2.3}
$$

Further derivation shows the relationship between the hazard and survival functions another way:

$$
h(t) = \frac{-d\log\left(S(t)\right)}{dt} \quad \leftrightarrow \quad S(t) = \exp\left(-\int_0^t h(u)\right) du. \tag{2.4}
$$

With this foundation, we can introduce the Cox proportional hazards model (Cox, 1972). The Cox proportional hazards (PH) model is used to model the hazard as a function of its hypothesized predictors. It has two components, including its baseline hazard function, *$h_0(t)$*, which is independent of all covariates other than time, and an exponential quantity, which incorporates a vector of the *p* covariates of interest denoted by $\underset{\sim}{x}$. The model can be written as:

$$
h(t, \underset{\sim}{x}) = h_0(t) \cdot \exp\left(\sum_{i=1}^{p} \beta_i x_i \right). \tag{2.5}
$$

The PH model is considered semi-parametric because the baseline hazard function does not have a specified form. The main assumption of the model is that the ratio of any two patients' hazards is equal to a constant that is independent of time. This is referred to as the proportional hazards assumption and provides the basis for the name of the model.

Letting Y_i denote the survival time for individual *i*, and assuming that no two patients can die at the same time, the log of the so-called partial likelihood which was proposed by Cox (1972) is:

$$
l(\underset{\sim}{\beta}) = \sum_{Y_i \ uncensored} \left\{ \underset{\sim}{x}_i \cdot \underset{\sim}{\beta} - \log\left[\sum_{Y_j \ge Y_i} \exp(\underset{\sim}{x}_j \cdot \underset{\sim}{\beta}) \right] \right\}. \tag{2.6}
$$

Essentially, we can treat the partial log-likelihood as we would treat a usual log-likelihood and use it for inference. In particular, we will use the partial log-likelihood to perform a likelihood ratio test to determine if treatment has a statistically significant effect on the survival outcome.

2.1.2 The Overall Effect Test

For the overall test, we do not consider the biomarker value of a patient; the only covariate in the model is the binary treatment indicator, *trt*. Then the model for the hazard function, $h(t,\underset{\sim}{x})$, is simplified from the form shown in (2.5) to:

$$h(t,trt)=h_0(t)\exp(trt\cdot\beta), \tag{2.7}$$

where $trt=1$ for treated patients, and $trt=0$ for patients in the control arm.

We perform the standard test for a treatment effect by testing the null hypothesis that the coefficient (β) corresponding to the treatment indicator in the model is equal to 0. To test this hypothesis we use the likelihood ratio test to compare model 2.7 (the full model) with a reduced model that contains the baseline hazard as its only term.

To perform the likelihood ratio test, we calculate twice the difference between the maximum of the partial log-likelihood of the full model and that of the reduced model. Letting Y_i denote the survival time for individual *i*, the maximum of the partial log-likelihood for the full model is:

$$l(\hat{\beta})=\sum_{Y_i\ uncensored}\left\{trt_i\cdot\hat{\beta}-\log\left[\sum_{Y_j\geq Y_i}\exp(trt_j\cdot\hat{\beta})\right]\right\}, \tag{2.8}$$

where $\hat{\beta}$ is the maximum partial likelihood estimate. The partial log-likelihood for the reduced model is:

$$l(0)=\sum_{Y_i\ uncensored}-\log(r_i), \tag{2.9}$$

where r_i is the number of individuals in the risk set at the time of death for individual *i*. That is, $r_i=\sum_j I(Y_j\geq Y_i)$. The likelihood ratio statistic, $D=-2\left[l(0)-l(\hat{\beta})\right]$, where $D\sim\chi_1^2$ under the null hypothesis H_0 (asymptotically), is compared to the appropriate cut-off for a significance level of 0.05. In this simplified setting, for which the model contains only one binary covariate, the likelihood ratio statistic essentially reduces to the log-rank test for comparing two Kaplan Meier survival curves under the null hypothesis that the two curves are statistically equivalent.

2.1.3 Procedure A

Procedure A is a two-stage procedure. In Stage I, the overall test is conducted using the method described above. Since this procedure has two stages, the overall Type I error (i.e.: across the two stages) must be controlled. JFS suggest partitioning the overall error of 0.05 by allocating some of this error to Stage I and the remainder to Stage II. They use a significance level of α_1 = 0.04 during the first stage and the remaining 0.01 as the significance level for the second stage.

If the test in the first stage is not significant at the α_1 level, then the method calls for proceeding to Stage II which takes into consideration the patients' biomarker values. The purpose of Stage II is to assess the significance of a treatment effect in the presence of a sensitive subgroup of patients. To do so, Cox PH models that include the treatment indicator as the only covariate are fit on subsets of the data, where these subsets are created according to patients' biomarker values. Assuming that biomarker values lie uniformly between 0 and 1, the four models that are fit include:

- A model for patients with biomarker values greater than or equal to 0.6
- A model for patients with biomarker values greater than or equal to 0.7
- A model for patients with biomarker values greater than or equal to 0.8
- A model for patients with biomarker values greater than or equal to 0.9

For each of the models, the partial log-likelihood is maximized to provide an estimate of the treatment effect. Then, the likelihood ratio statistic for each cut-off value can be calculated. The test statistic, T_A, is taken to be the maximum of these likelihood ratio statistics. Notationally,

$$T_A = \max\left(D(0.6),\ D(0.7),\ D(0.8),\ D(0.9)\right), \qquad (2.10)$$

where $D(c)$ represents the log partial likelihood ratio statistic corresponding to the model for patients with biomarker values greater than or equal to *c*.

Unlike the likelihood ratio statistic which (asymptotically) follows a chi-square distribution under the null hypothesis, the distribution of T_A does not follow a chi-square distribution. This is because of the well-known multiple testing problem, as the procedure calls for the computation of four partial likelihood ratio statistics and defines the test statistic as their maximum. To find a valid p-value, an approximation to an exact test is based on the permutation distribution of the test statistic. The p-value is approximated using the following algorithm:

1. Create P permutation datasets where each dataset corresponds to the original dataset but with the treatment codes randomly permuted.
2. For each permutation dataset, re-run Stage II of Procedure A. That is, fit four Cox proportional hazards models for subsets of the dataset above the given biomarker cut-offs and calculate the permutation test statistic, T*, as defined by the maximum partial likelihood ratio statistic across the four models.
3. The permutation p-value is defined as $\frac{1+\sum_{i=1}^{P} I(T_i^* > T_A)}{1+P}$

To maintain the overall Type I error at the 5% level, the Stage II p-value is compared to a significance level of 1%, since Stage I was conducted at a 4% level.

The use of Procedure A requires two implicit assumptions. Firstly, the subset tests assumed that the biomarker values lie uniformly between 0 and 1. For biomarkers in which this is not the case, the percentiles corresponding to the values are used, instead of the actual values. As we will see in Chapter 4, this is not always the most sensible approach to partitioning patients, especially with count markers that have highly asymmetric distributions.

Secondly, the procedure has an inherent assumption that higher values of the biomarker are more likely to be associated with a treatment effect, or with a treatment effect of greater magnitude than lower values. However, if the biomarker relationship occurs in the opposite direction (i.e.: patients with low biomarker values are expected to benefit most from treatment) the procedure can be easily modified to conform to this relationship. The researcher must hypothesize in advance the direction of the relationship between the biomarker and treatment effect and use the methodology accordingly.

The partition of the Type I error across the stages of Procedure A can be modified to the researcher's preference. In the original methodology, Stage I in Procedure A was conducted at a 4% significance level, which forced Stage II to be conducted at a 1% level to maintain an overall Type I error rate of 5%. One could modify the partition of the error rate based on the objectives of the analysis. For example, if Stage I was conducted at a 1% level this would decrease the chance of finding a treatment effect in the overall group. However, Stage II would then be conducted at a 4% level which would increase the chance of finding a treatment effect in a biomarker-defined subset. Thus, the specification of how the Type I error is be partitioned should be defined *a priori* and will likely reflect the researchers' belief regarding how the treatment effect will behave across biomarker levels.

2.1.4 Procedure B

Similar to Procedure A, Procedure B fits multiple Cox PH models across biomarker-defined patient subsets. Ten models are fit in total; the first model includes all patients, and subsequent models look at subsets of patients with biomarker values larger than a specified cut-off. The cut-offs are 0.1, 0.2, and so on to 0.9. For each model, the treatment effect is estimated by maximizing the partial log likelihood. The value of the test statistic is computed as:

$$T_B = \max\left(D(0) + 2.2,\ \{D(c) \mid 0 < c < 1\}\right). \tag{2.11}$$

Note that when calculating T_B, an adjustment is made by adding 2.2 to the partial likelihood ratio statistic from the overall model. This increases the chance that the treatment effect would be deemed significant for the overall group, rather than in just a subset of patients, in the event that the log likelihood ratio statistics are very close between the overall test and one of the subset-based tests. JFS apply such a modification because it is preferable that the treatment effect be detected in the overall group, rather than in a smaller subset, if one is in fact present.

The permutation distribution is used to find the permutation p-value for T_B as described earlier for Procedure A. This p-value is compared to a significance level of 5% as Procedure B only has one stage.

2.1.5 Estimating the Cut-off That Defines the Sensitive Subset

JFS define the point estimate of the cut-off which defines the sensitive subset of patients, $\hat{c}$, as:

$$\hat{c} = \arg\max_{c}\left(D(c)\right). \tag{2.12}$$

To get a confidence interval (CI) for this point estimate, JFS suggest use of bootstrapping. This is done using the following algorithm:

1. Create B bootstraps. Each bootstrap sample is created by sampling rows from the dataset with replacement until a sample of the same size as the original dataset is chosen.
2. For bootstrap i, find the point estimate for the biomarker cut-off value. Call this estimate c_i^*.
3. The distribution of the cut-off value can be estimated using the empirical distribution of c^*. Percentiles of the empirical distribution can be used to find a confidence interval for the cut-off value. For a 95% CI we consider the 2.5th and 97.5th percentiles.

2.2 Simulation Study

JFS present simulation results comparing the empirical power of the three procedures. In this thesis, I present results based on my own simulations to: i) confirm my understanding of the JFS methodology and ii) verify the empirical power results presented by the authors. I first describe the 19 simulation scenarios which are evaluated in the original paper and which I also chose to evaluate. This is followed by a description of the models used to generate the simulated datasets and a description of how the empirical power was calculated for each procedure. Lastly, a brief interpretation of the results from the simulation is presented.

2.2.1 Simulation Scenarios

Two general scenarios are investigated:

- Scenario 1: Treatment benefit is confined to a biomarker-defined patient subset. The magnitude of treatment benefit is constant within the group of sensitive patients and the value for which benefit begins is referred to as the biomarker cut-off.
- Scenario 2: Treatment benefit increases as a function of the biomarker values.

In Scenario 1 we consider five sub-scenarios in which the cut-offs are placed at: i) 0 (corresponding to the entire treated group receiving benefit), ii) 0.25, iii) 0.50, iv) 0.75, and v) 0.90. Without loss of generality, we again assume that the biomarker values lie uniformly between 0 and 1.

For Scenario 2 we consider two sub-scenarios. In Scenario 2.i), the log hazard ratio decreases linearly over the entire range of biomarker values from a hazard ratio of 1 (corresponding to no benefit) to a pre-specified reduction level (corresponding to the maximum reduction in hazard). For Scenario 2.ii), patients with biomarker values less than 0.5 have a hazard ratio of 1, and those subjects with biomarker values between 0.5 and 1 experience a linear decrease in the log hazard ratio over this restricted range. For each of the sub-scenarios, 2-3 levels of hazard reduction are considered to explore how the empirical power of the procedures vary across relatively small to relatively large treatment effects.

2.2.2 Dataset Creation

The samples from the simulated population had the following properties:

1. 100 patients each were assigned to the treatment and placebo arms, for a total of 200 patients in the sample.
2. For each patient, a biomarker value, *bmk*, was generated from a uniform(0,1) distribution:

   $$bmk \sim U(0,1)$$
3. Entry times were staggered and generated from a uniform(0, 0.5) distribution:

   $$entry\ time \sim U(0,0.5)$$

   The total study length was fixed at a length of 3. If we think of these units of time as years, then the entry time can be interpreted to mean that patients are allowed to enter the study anytime during the first six months of the study's total length of three years, after which time study enrolment was closed.
4. The generation of patients' lifetimes is dependent on the simulation scenario.
   i. Scenario 1: A patient's lifetime was generated from an exponential distribution, where the survival rate was dependent on the subject's treatment status and biomarker value. If the patient was in the control arm, then a hazard rate of 1 was used (independent of the patient's biomarker value). If the subject was in the treatment arm and had a biomarker value greater than the cut-off defined by the simulation scenario, a reduced hazard rate, *rr*, was used. This reduced rate corresponded to that defined by the simulation scenario. Thus, only those in the treatment arm who have a biomarker value higher than the level specified for a given simulation have responses generated from an exponential distribution with the reduced hazard rate, *rr*. In detail:

$$survival\ time \sim \begin{cases} \exp(1), \textit{ if the patient is in the control arm} \\ \exp(1), \textit{ if the patient is in the treated arm but has bmk} \leq \textit{cut-off} \\ \exp(rr), \textit{ if the patient is in the treated arm with bmk > cut-off} \end{cases}$$

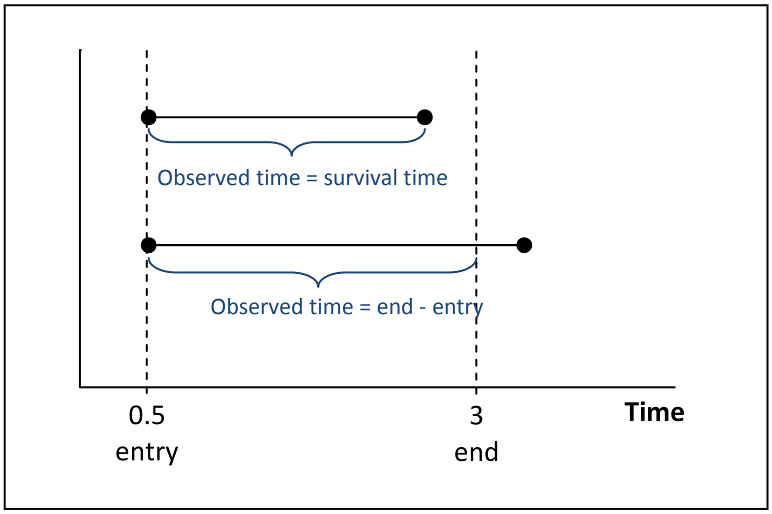
ii. Scenario 2: Here, treatment effect increases linearly in the biomarker. The reduced hazard rate is calculated according to the description provided in Section 2.2.1 for the two sub-scenarios under Scenario 2.

5. All subjects were assumed to have complete follow-up. As a result, right censoring was introduced into the dataset only for patients that reach the end of the study before they've had events. The end of the study was chosen to occur at time 3, and the calendar time of a subject's event is the sum of the subject's entry time and lifetime. The authors restricted censoring to lie between 10 and 20 percent, so if the proportion of censoring falls outside this range the response times were re-generated until a censoring rate within these limits was reached.
6. Each patient's observed response depends on whether their generated event occurs before or after the study ends. If the event occurs before the study ends, then their complete survival time is observed. If it occurs after the study ends, then we only observe the difference between the time of the end of the study and the time that the patient entered the study:

$$observed\ response = \begin{cases} survival\ time,\ if\ patient\ has\ an\ event\ before\ the\ study\ ends \\ 3 - entry\ time,\ if\ patient\ has\ an\ event\ after\ the\ study\ ends \end{cases}$$

To elucidate, consider two patients who enrol in the study at time equal to 0.5, where the first patient has a survival time of 2.1, and the second patient has a survival time of 2.6. Then, we observed the first patient's full lifetime, since it occurs before the end of the study. For the second patient, the lifetime is censored at 2.5 (i.e.: 3-0.5), since the event will occur after the study finished. This example is illustrated in Figure 2.1.

Figure 2.1: Illustration of different observed responses



2.2.3 Calculation of Empirical Power

For each procedure the following outlines the method used to evaluate the empirical power:

1. Overall Effect Test: For each sample in the simulation, a Cox PH model containing the treatment indicator as the only covariate is fit. The empirical power is the proportion of samples in the simulation for which the p-value from the likelihood ratio test is less than the specified significance level of 0.05.
2. Procedure A: For each sample in the simulation, we first determine if the OET is significant in Stage I by testing against a reduced significance level of $\alpha_1 = 0.04$. If it is, then the procedure stops and the treatment effect is deemed significant. If it is not, then the test statistic T_A is computed and its p-value is calculated using an approximate permutation test. This p-value is compared to a significance level of $\alpha_2 = 0.01$. If this Stage II result is significant, then the treatment effect is deemed significant. The empirical power is the proportion of samples in the simulation that have a significant result, either at Stage I or Stage II.
3. Procedure B: Since there is only one stage for this procedure, calculation of empirical power is straightforward. The p-value for the test statistic T_B for each sample is calculated using an approximate permutation test and the empirical power is defined as the proportion of samples for which this p-value is less than 0.05.

2.2.4 Simulation Results

Using the methods described above, I conducted simulations to validate the results presented by JFS. Ten thousand samples from the simulated population were drawn on which each of the procedures was performed. The permutation distribution required by Procedures A and B used 1,000 permutations for each of these samples.

My results were very similar to those presented by JFS, with deviations in the expected range. Table 2.1 presents the empirical power of the three methods under the varying simulation scenarios. Figure 2.2 and Figure 2.3 present this information graphically.

Table 2.1: Simulation results for a survival outcome

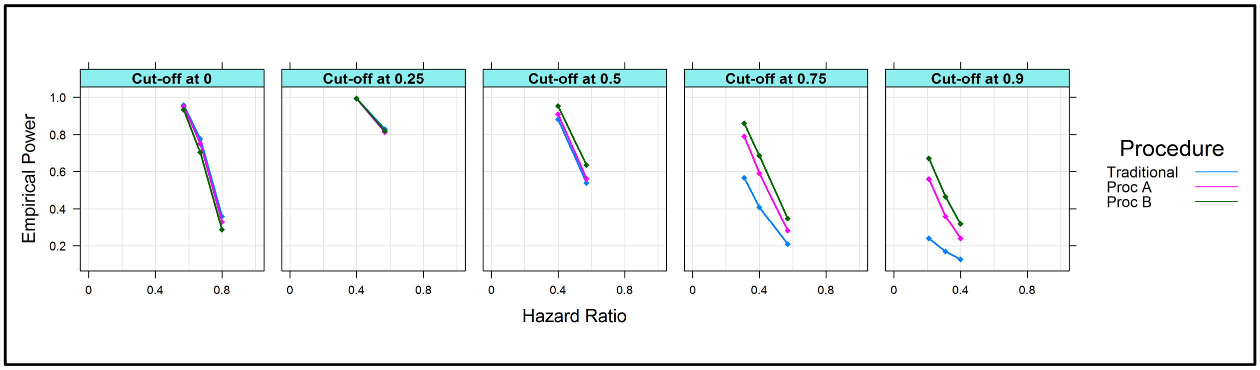
Scenario		Model	% Reduction in hazard (hazard ratio)Ŧ	Overall Power*	Procedure A Power◊	Procedure B Power◊
1.	i)	Everybody benefits from new therapy	20 (0.80)	0.36	0.33	0.29
			33 (0.67)	0.78	0.75	0.70
			43 (0.57)	0.96	0.95	0.93
	ii)	Only patients with biomarker values > 0.25 benefit from new therapy	43 (0.57)	0.83	0.81	0.82
			60 (0.40)	0.99	0.99	0.99
	iii)	Only patients with biomarker values > 0.50 benefit from new therapy	43 (0.57)	0.54	0.56	0.63
			60 (0.40)	0.88	0.91	0.95
	iv)	Only patients with biomarker values > 0.75 benefit from new therapy	43 (0.57)	0.21	0.28	0.35
			60 (0.40)	0.41	0.59	0.68
			69 (0.31)	0.57	0.79	0.86
	v)	Only patients with biomarker values > 0.90 benefit from new therapy	60 (0.40)	0.13	0.24	0.32
			69 (0.31)	0.17	0.36	0.47
			79 (0.21)	0.24	0.56	0.67
2.	i)	Linear decrease in log hazard ratio	43 (0.57)	0.52	0.51	0.54
			60 (0.40)	0.87	0.87	0.89
			69 (0.31)	0.97	0.98	0.98
	ii)	Linear decrease in log hazard ratio for patients with biomarker values > 0.5	43 (0.57)	0.19	0.23	0.27
			60 (0.40)	0.41	0.51	0.60
			69 (0.31)	0.57	0.71	0.79

Ŧ For scenario 1, the hazard ratio corresponds to the survival rate of the treated group for the biomarker-defined subset. For scenario 2, the log hazard ratio is a linear function of the biomarker value, so the table indicates the minimal hazard ratio (maximal benefit) corresponding to those treated patients with a biomarker value of 1.

*Overall Test based on 10,000 samples

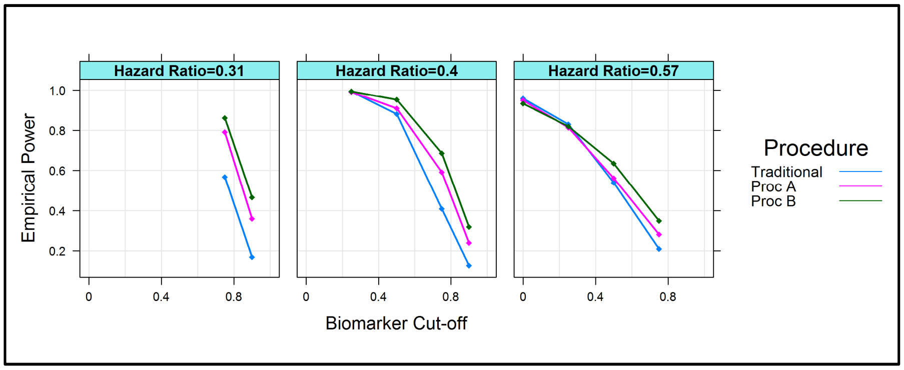
◊Procedures A and B based on 10,000 samples. The approximation of the permutation distribution was based on 1,000 permutations of the data and was used to calculate the p-value and find the empirical power for Procedures A and B.

Figure 2.2: Empirical power as a function of hazard ratio across biomarker cut-off values



Note: This figure is based on data from Scenario 1 only (and does not include the data from the scenario in which the treatment effect (hazard ratio) is a linear function of the biomarker).

Figure 2.3: Empirical power as a function of the biomarker cut-off for a range of hazards



Note: This figure is based on data from Scenario 1 which consider hazard ratios between 0.31 and 0.57 inclusively.

Table 2.1 and Figure 2.2 illustrate that, as one expects, power decreases as a function of the hazard ratio for all methods. That is, it is harder to detect a significant treatment effect as the risk differential between the treatment and placebo group is decreased. As well, if most patients are benefitting from treatment (corresponding to biomarker cut-off values at 0 or 0.25) then the procedures' performances are very similar across the different magnitudes of treatment effects considered. However, if the treatment effect is confined to less than half of the treated population, then the adaptive procedures outperform the traditional method, with the most marked increase in benefit for the smallest hazard ratios. In terms of the adaptive procedures, Procedure B tends to have the same power or outperform Procedure A across the scenarios considered. In our simulations, it is only when the whole treatment group benefits from treatment that the adaptive methods underperform the traditional method, and even then the loss in power of the adaptive methods is very modest.

Figure 2.3 provides additional insight. Power decreases as a function of the biomarker cut-off value; as the size of the patient subset which benefits from treatment decreases, the ability to detect the treatment effect is reduced in turn. However, the rate of decrease is slower for the adaptive procedures compared to the traditional procedure.

The scenarios in which the treatment effect is linearly related to the biomarker value define a different type of biomarker-treatment relationship than that considered by the scenarios in which the treatment effect is confined to a range defined by a cut-off value. From Table 2.1, we see that the traditional and adaptive procedures perform similarly across different magnitudes of treatment benefit if the treatment benefit starts to become effective for relatively low biomarker values. On the other hand, if treatment only becomes effective for patients with biomarker values above 0.5, then the adaptive procedures outperform the traditional method.

We can also compare the performance of the methods across Scenarios 1 and 2 under equivalent treatment effects. We firstly can compare Scenario 1.i) and Scenario 2.i) for a hazard ratio of 0.57. In Scenario 1.i), all patients in the treated group can benefit and in this formulation the traditional test for the treatment effect has slightly higher power than the adaptive method. However, in Scenario 2.i), benefit increases as a function of the biomarker value. In this case, power across the methods is reduced to almost half of that in Scenario 1.i), but Procedure B maintains a slightly higher power, showing that it is somewhat better at detecting a treatment effect for this type of treatment-biomarker relationship. Similarly, we can compare Scenario 1.iii) to Scenario 2.ii) for hazard ratios of 0.57 and 0.40. Both of these scenarios correspond to benefit commencing at a biomarker cut-off value of 0.5, with the first scenario having a constant treatment effect and the second scenario having the treatment effect as a function of the biomarker value. In both cases, the adaptive procedures have better performance than the traditional test. However, the relative increase in performance is larger for Scenario 2.ii). This conclusion is sensible because the treatment effect for Scenario 2.ii) is less detectable in the overall group than the treatment effect in Scenario 1.iii). Overall, the adaptive procedures are good at detecting treatment effects that are characterized by both types of treatment-biomarker relationships.

The JFS methodology was motivated by situations in which the effectiveness of treatment may vary according to some genetic biomarker. The primary endpoint considered was a time-to-event variable, as these types of variables are commonly used in cancer studies. In the next chapter of this thesis, we explore generalizations of the BATD to allow for the different types of endpoints that are more common in clinical trials for treatment of multiple sclerosis (MS).

3 Extensions to the Biomarker Adaptive Threshold Design

We begin this chapter with a brief background of MS and a discussion of typical endpoints used in MS clinical trials. Following this, methodology is developed to extend the BATD to facilitate count endpoints and biomarkers and properties of the developed adaptive methods are studied via simulation.

3.1 Endpoints in Multiple Sclerosis Clinical Trials

Multiple sclerosis is a chronic disease of the central nervous system. It affects between 180 and 350 out of every 100,000 people in Canada, varying across geographic region (Beck, Metz, Svenson, & Patten, 2005). Research in the area of MS is plentiful, as the aetiology is still unclear and there is no known cure for MS at this time. There are four different types of MS, including relapse-remitting (RR), primary-progressive (PP), secondary-progressive (SP), and progressive-relapsing (PR) MS. According to the National MS Society (2010a), a large proportion of patients have RR MS at initial diagnosis, and half of these patients will progress to SP MS within 10-20 years of diagnosis. While RR MS patients have temporary relapses followed by a return to their usual levels of function, SP MS experience a steadily worsening disease course and may have only minor or no relapses (National MS Society, 2010a). Currently, there are eight FDA-approved disease modifying therapies available for the treatment of MS (National MS Society, 2010b). Each of these drugs was approved for treatment only after undergoing rigorous testing during the stages of clinical trials which are required for all pharmaceutical products before they can be used by the general public. With the advent of new treatments, further clinical trials will be needed to assess treatment effectiveness.

One of the defining features of RR MS is the periods of relapse and remission that patients undergo during these stages of the disease. During relapse, MS symptoms are exacerbated and the relapse may induce an increase in the patient’s overall level of disability which persists through future periods of remission. The primary outcome in many Phase III (or pivotal) clinical trials in RR MS is often the annual relapse rate. However, the annual relapse rate is often less than one relapse per year in many patient populations, necessitating large studies (i.e.: in the way of long follow-up or a large sample size) in order to accurately determine whether a significant reduction in the relapse rate has occurred. Thus, in early phase clinical trials, other endpoints that generate more frequent events are commonly used.

MRI of the brain and spinal cord have been used to monitor MS patients as patients typically experience lesions in these areas of their central nervous system. Lesions are considered silent indicators of inflammatory activity (Yong, Chabot, Olaf, & Williams, 1998). There are different MRI techniques to measure lesion activity, including T1-weighted, T2-weighted and FLAIR images. Lesions are considered enhancing if they appear as bright spots on the T1 MRI image after injection of a contrast-enhancing agent such as Gadolinium. These bright spots indicate a breakdown of the blood-brain barrier at that location. Contrast enhanced T1-weighted images are generally considered reflective of current disease activity, while T2-weighted images are more reflective of the accumulation of disease as measured by lesion load. In this thesis, we consider contrast enhancing lesions on T1-weighted images as a potential predictive biomarker. Generally, the methodology developed herein for count outcomes can apply to lesion count activity, independent of the imaging technique used. Lastly, the numbers of MRI images taken during trials vary. Frequent MRI is a phrase used denote when MRI images are taken frequently (i.e.: monthly) and characterizes some of the trials that are discussed in this thesis.

It is generally accepted that MRI lesion activity is indicative of the level of disease activity, although research is still ongoing to investigate the properties of this relationship. Most recently, the literature has focused on investigating whether a surrogacy relationship exists between new MRI lesions and relapses on the trial level (Sormani, Bonzano, Roccatagliata, Cutter, Mancardi, & Bruzzi, 2009). In this research, the authors investigate how predictive a treatment effect on MRI lesion activity is of the treatment effect on relapse activity using meta-analysis. They found a strong predictive relationship, suggesting the notion that MRI lesion activity is predictive of relapse activity is a feasible one – at least at the level of patient groups. Since MRI lesion activity is related to disease, many early phase clinical trials have used imaging outcomes as the primary outcome, as imaging provides more sensitive outcomes than the annualized relapse rate.

Both clinical (relapse rate) and imaging (lesion activity) outcomes have been used during different phases of clinical trials as the primary outcome to assess treatment effectiveness. For clinical endpoints, the number of relapses, or the relapse rate has been a commonly used endpoint and has been considered one of the most suitable endpoints (Whitaker, McFarland, Rudge, & Reingold, 1995). The expanded disability status scale (EDSS) score, which is a measure of disability, is also a suitable endpoint but we will not consider it in this thesis. Often, the choice of clinical endpoint is related to the disease course of MS being studied. For RR MS, relapse-related endpoints are commonly used, while for SP MS, endpoints related to EDSS are the norm. In terms of imaging endpoints, there are many options, such as the frequency or area of lesions, the appearance of new lesions, or the enhancement of existing lesions (Whitaker, McFarland, Rudge, & Reingold, 1995).

3.2 Alternative Setting: Going from Genetic Biomarkers in Cancer to Pre-Study Information in Multiple Sclerosis

The strength of the BATD lies in its ability to conserve statistical power in cases where a sensitive subgroup of patients benefits from treatment more than the overall group. For some conditions or diseases, the sensitive subgroup may be defined by a genetic predisposition which was the motivating scenario behind the development of the BATD. The best indicator of a sensitive subset, however, will vary according to the disease-treatment combination. In this section, we consider a specific type of treatment, Interferon β (IFNβ), and hypothesize some predictive biomarkers for this specific treatment in the MS context.

Interferon β has been shown to significantly reduce the relapse rate across many MS patient cohorts. The mechanism of action of IFNβ is complex, where part of the mechanism includes its anti-inflammatory properties (Yong, Chabot, Olaf, & Williams, 1998). Due to this property of IFNβ, it may be the case that patients with more inflammation experience higher levels of benefit from these treatments.

3.2.1 Disease Severity as a Possible Indicator of Treatment Effectiveness

An indicator of treatment responsiveness may be disease severity. As MRI lesions are a silent indicator of the inflammation process which is associated with the disease it could be that the number of lesions at baseline is predictive of the effectiveness of IFNβ treatment.

MS patient populations exhibit a high variability of symptoms. The majority of patients will have no lesions or very few at times of remission, although some patients will have many. For example, in a cohort of patients enrolled in a clinical trial to test the safety and effectiveness of IFNβ, 105 of the 158 patients (66%) with baseline measures of lesion count had no newly active lesions at baseline, while 6% had more than five (The North American Study Group on Interferon beta-1b in Secondary Progressive MS, 2004). If a treatment is only effective in patients with a higher severity of disease as defined by baseline lesion count (BLC), then performing the OET may not detect this effect (especially if it is powered to detect the difference in the overall group). Here, it becomes even more critical to consider patient subsets in addition to the overall group.

3.2.2 MS Duration as a Possible Indicator of Treatment Effectiveness

In MS, it has been found that inflammation has a negative relationship with the length of time that a patient has had the disease (Huitinga, Erkut, van Beurden, & Swaab, 2004). The longer a patient has MS, the lower the amount of expected inflammation. Thus, another potential indicator of treatment effectiveness is MS duration, where a shorter duration may be indicative of greater benefit from treatments that target the suppression of inflammatory activities.

The length of MS duration varies across the patient cohort enrolled in clinical trials. In many trials, participants must have MS for at least two years in order to be eligible to participate in the trial. Based on the hypothesized negative relationship between duration and treatment effectiveness, one can imagine that if a trial was conducted on a cohort that includes few patients with short duration, the estimated treatment effect may not be statistically significant. The BATD may be useful in this setting as well.

3.3 Generalization of Methodology for a Count Biomarker and Outcome

The preceding section provides examples in the MS framework for which the BATD may be useful in detecting treatment effects that may be larger in a sensitive patient subgroup, or confined to the subgroup altogether. In the first example, the hypothesized patient subset was related to number of lesions, which is a count variable. In the second example, the subset was defined by a continuous variable; the duration of time a patient has had MS. While the JFS methodology is applicable to the continuous marker, modification to this methodology is required for the count marker.

Here, we consider a count biomarker and response, both of which have distributions that are asymmetric and highly skewed to the right. Although the methodology developed in this section can be applied to any count variables, we are specifically interested in the number of newly enhancing lesions as the hypothesized predictive marker and the number of relapses during study follow-up as the measured outcome. Then the annual relapse rate (ARR), which is equal to the number of relapses for each patient divided by their time on study in years, will be used as the indicator of treatment effectiveness.

We first investigate which parametric distributions are most appropriate to model these counts. Then, aspects of the original methodology are generalized and elements of the simulation study are revised according to the structure of the data investigated. We also discuss characteristics of the simulated population to illustrate that it is reflective of traits which could reasonably be exhibited in real populations of interest.

3.3.1 Choosing the Most Appropriate Distributions

Arguably the most common way to model count outcomes is using the Poisson distribution. It is most appropriate for modelling data in which the variance is roughly equal to the mean. Both lesion activity and relapse counts tend to be overdispersed relative to the Poisson distribution. That is, the variability found in the data tends to be much larger than the mean. One reason for this is that patients typically exhibit different average levels of activity, both in terms of lesion and relapse activity. For lesion activity, over a fixed period of time some patients will be highly active and exhibit higher lesions counts than most other patients, while other patients will have much less activity over the period (Wang, Meyerson, Tang, & Qian, 2009). The means of the underlying distributions of the processes which are responsible for generating lesion activity differ across patients. Thus, the assumption of identical rates of activity required by the Poisson model is violated.

For data which exhibits overdispersion researchers often use negative binomial or quasi-Poisson models. For both of these models, the variance is a function of the mean but this function differs according to the model. For quasi-Poisson models, the variance is proportional to the mean, such that $Var(Y) = \phi \cdot E(Y)$ where ϕ denotes an arbitrary positive parameter which allows for overdispersion relative to the Poisson distribution. For the negative binomial model, the variance is a quadratic function of the mean taking the form $Var(Y) = E(Y) + \theta \cdot E^2(Y)$. For the remainder of this thesis, we will refer to θ as the dispersion parameter, such that values of θ which are larger than zero will denote those distributions with variances larger than their mean.

The negative binomial distribution can be derived as a Poisson-Gamma mixture model. First, we define our model notation (Table 3.1).

Table 3.1: Properties of the distributions of interest

Notation and parameters	Mean	Variance
$X \sim Gamma(\alpha,\beta)$ $\alpha > 0$ $\beta > 0$	$\alpha\beta$	$\alpha\beta^2$
$Y \sim NB(\mu,a)$ $\mu > 0$ $a > 0$	μ	$\mu + a \cdot \mu^2$

If we let y denote the response, x denote a vector of covariates which influences y, and v a factor which allows for heterogeneity among patients' response rates, then we have the following result (Lawless, 1987):

$$If \begin{cases} Y \mid V = v, X = x \sim Pois(v \cdot \mu(x)) \\ V \sim Gamma(1/\theta,\ \theta) \end{cases},\ then\ Y \mid X = x \sim NB(\mu(x),\ \theta).$$

By the properties shown in Table 3.1, the distribution of the response conditional on the covariates has the variance $Var(Y \mid X = x) = \mu(x) + \theta \cdot \mu^2(x)$ where $E(Y \mid X = x) = \mu(x)$.

Derived in this way, the model is easy to interpret in the setting of lesion or relapse counts. For example, let $\mu(x)$ define the mean rate of relapse for the group of patients with covariates defined by x. Then v allows for heterogeneity among patients' relapse rates within this group. Specifically, those patients with $v > 1$ are more apt to relapse at any given time than other patients with the same characteristics as defined by x.

The negative binomial model has been recommended and subsequently accepted in the MS literature as an adequate model for lesion count data (Sormani, Bruzzi, Miller, Gasperini, Barkhof, & Filippi, 1999). In terms of relapse counts, there has been less research into the adequacy of competing models, with a notable exception in which the negative binomial model was recommended (Wang, Meyerson, Tang, & Qian, 2009). In this thesis, we will model relapse count also using the negative binomial distribution. One benefit of using such a model is that the limiting distribution of a negative binomial model when θ, the dispersion parameter, approaches 0 is a Poisson model. In the event that a negative binomial model is wrongly used to model Poisson data, the inferences based on the model will still be correct.

3.3.2 Generalizing the Methodology

Three prominent aspects of the JFS methodology need to be revised to accommodate the count distributions of the lesion activity and the relapse counts. We address each of these aspects in turn.

Firstly, JFS assume that the biomarker of interest can be easily converted to a uniform scale by using the percentiles of the biomarker's distribution. Procedures A and B then fit models using data from these percentile-based subsets. With over-dispersed count data, the use of percentiles no longer applies as it did with biomarkers that lie on a continuous range. Instead we consider patients according to their level of the count variable. That is, we first consider all patients, then patients with biomarker counts greater than 0, then patients with counts greater than 1, and so forth. As we continue to subset the data, the amount of patients in each subset quickly diminishes by virtue of characteristics of count distributions with low averages and right-skewed distributions (as is common with lesion counts). Thus, we will limit the number of subsets to be assessed such that subsets are only created if they will contain at least 10% of the patients.

Secondly, a related implication is how to revise the form of the test statistic for Procedures A and B. Recall that when the subsets were created according to percentile-cuts of the biomarker distribution, the test statistics for Procedures A and B are:

$$T_A = \max\left(D(0.6), D(0.7), D(0.8), D(0.9)\right),$$

$$T_B = \max\left(D(0) + 2.2, \left\{D(c) \mid 0 < c < 1\right\}\right),$$

where $D(c)$ represents the log-likelihood ratio statistic corresponding to the model for patients with biomarker percentile values greater than or equal to *c*. For Procedure B then, it is straightforward to update *c* here to denote the level of a patient's baseline lesion count, where the range of *c* corresponds to subsets which contain at least 10% of the patients. For Procedure A we also need to decide which subsets the test will be based on. Since it is commonplace that more than half of the patients will have no baseline lesions, the 60th percentile usually occurs at either 0 or 1 lesions, depending on the population under consideration. In correspondence with the original methodology then, the test statistic for procedure A should include all subsets starting with patients with 1 or more lesions. Note that defining the subsets in this way makes the form of the test statistics for Procedures A and B more similar than in the JFS methodology. The proposed test statistics become:

$$T_A = \max\left(D(1), D(2), \ldots\right), \tag{3.1}$$

$$T_B = \max\left(D(0) + 2.2, \left\{D(c) \mid c \in \left[1,\ 2,\ \ldots\right]\right\}\right). \tag{3.2}$$

Thirdly, the test for the treatment effect will be based on the negative binomial distribution. JFS used an unadjusted test for the treatment effect (i.e.: the only covariate specified in their model is the treatment indicator). This may be inappropriate in the MS framework, as there has been evidence that new lesion activity is a surrogate endpoint for relapse counts (Sormani & Filippi, 2007). Part of the criterion of surrogacy is the existence of a significant correlation between the two variables, thus it is reasonable to adjust the test for treatment effect for baseline lesion activity. For other biomarkers, there may be a relationship between the biomarker and primary outcome, in which case adjustment for the biomarker would also make sense.

As was done by JFS, the test for treatment effect will use the likelihood ratio statistic. A negative binomial regression model is used to relate the mean ARR for each patient *i*, μ_i, to the covariates in the model where this relationship depends on the regression parameters denoted by $\underset{\sim}{\beta}$. For a negative binomial distribution with regression coefficients $\underset{\sim}{\beta}$ and dispersion parameter denoted by θ, the log-likelihood can be expressed as follows:

$$l(\underset{\sim}{\beta}, \theta) = \sum_{i=1}^{n} \log\left(\frac{\Gamma(y_i + \theta^{-1})}{y_i!\,\Gamma(\theta^{-1})} \left(\frac{\theta\mu_i}{1+\theta\mu_i}\right)^{y_i} \left(\frac{1}{1+\theta\mu_i}\right)^{\theta^{-1}} \right), \tag{3.3}$$

where $\mu_i = \exp(\underset{\sim}{\beta}'\underset{\sim}{x}_i) = \exp\left(\beta_0 + \beta_1 \cdot f(BLC) + \beta_2 \cdot trt\right)$, and BLC denotes the baseline lesion count. Discussion regarding the form of the relationship between the average relapse rate and BLC is considered later in this chapter.

The likelihood for the reduced model (without treatment) is the same as the likelihood for the full model, except here: $\mu_i = \exp(\underset{\sim}{\beta}'\underset{\sim}{x}_i) = \exp\left(\beta_0 + \beta_1 \cdot f(BLC)\right)$.

Then the likelihood ratio statistic, $D = -2\left[l(\hat{\beta}_R) - l(\hat{\beta}_F) \right]$ where $\hat{\beta}_R$ and $\hat{\beta}_F$ denote the vector of maximum likelihood estimates of the model parameters for the reduced and full models, respectively. $D \sim \chi_1^2$ is compared to a significance level of 0.05 for the overall effect test used in the non-adaptive method. This quantity, *D*, is known as the deviance for generalized linear models.

3.4 Simulation Study in the MS Context

We first discuss attributes of the simulated population and provide reasoning for why these attributes were chosen. Then, the simulation scenarios that were investigated are described and results from the simulations are presented and interpreted.

3.4.1 Attributes of the Simulated Patient Population

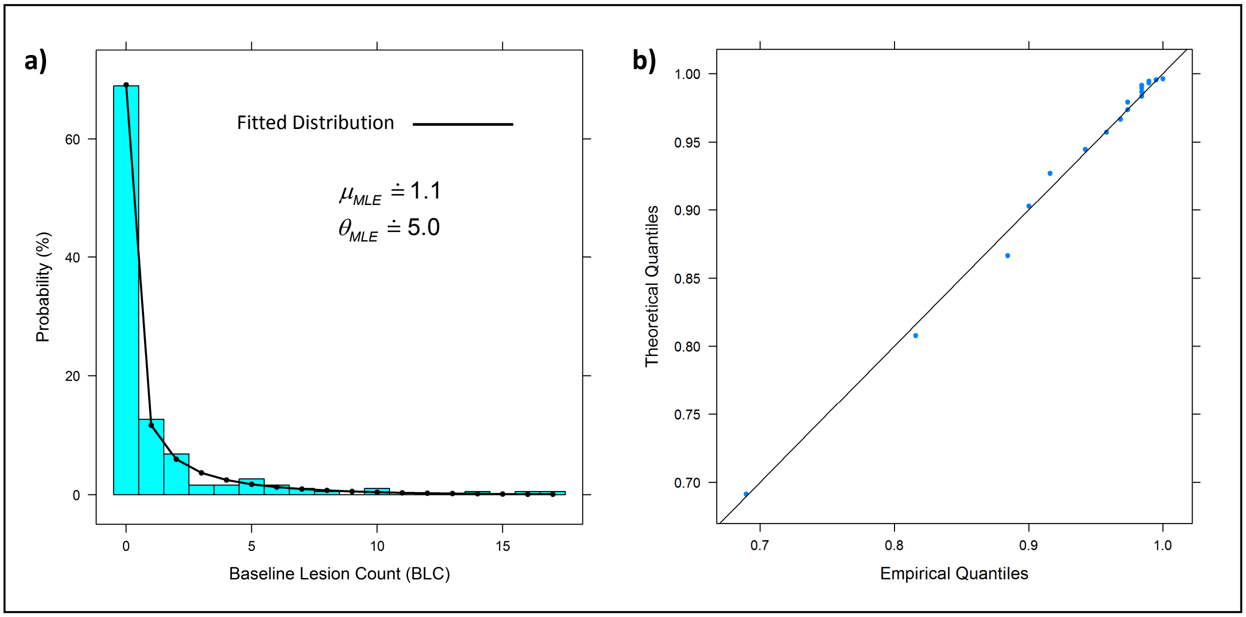
Characteristics of the simulated population were motivated by the investigation of attributes of two populations. Cohort I included SP MS patients that were enrolled in the Canadian Phase II Micellar Paclitaxel study which failed to meet its primary efficacy outcome. Zhao et al. (2010) present further details about this study. Cohort II consists of patients from the Phase III Tysabri (Natalizumab) MS trial. Wang et al. (2009) provide more detail regarding this study. We had access to the dataset corresponding to Cohort I, but only summary results from Cohort II. Because of this, we primarily investigated Cohort I and supported this investigation using the known attributes of Cohort II.

Firstly, we consider characteristics of the distribution of lesions at baseline. This is followed by a discussion of the relationship between on-study relapse activity and BLC and characteristics of the distribution of on-study relapses. While we consider attributes of the patients across treatment groups from the two studied cohorts, our focus is on determining reasonable parameters to characterize the simulated placebo patients' baseline lesion and on-study relapse distributions. The characteristics for the treated patients' distributions will then behave largely in the same way, only with treatment effects that will change the shape of the on-study relapse distribution accordingly.

Baseline Lesion Count

To investigate the distribution of newly enhancing lesions on T1-weighted MRI at baseline, model parameters for both the Poisson and negative binomial distributions were estimated using the method of maximum likelihood with the Cohort I data. This was done using the *fitdistr* function from the MASS package (Venables & Ripley, 2002). The fits of the distributions to the data were tested using the chi-squared test as performed by the *goodfit* function from the VCD package (Meyer, Zeileis, & Hornik, 2010). The Poisson distribution provides a very bad fit to the data (p-value < 0.001), however the negative binomial distribution provides an adequate fit (p-value = 0.059). A histogram of the lesion data from Cohort I is shown here (Figure 3.1a), with an overlaid probability distribution function of the fitted negative binomial model. As well, Figure 3.1b shows the corresponding quantile-quantile plot.

Figure 3.1: Histogram and quantile-quantile plot of the lesion data and the corresponding NB probability distribution function



Using the above visual aids, the negative binomial model with the chosen parameters appears to adequately capture the trends in this dataset. This fit provided the simulation parameters for the lesion distribution (Table 3.2).

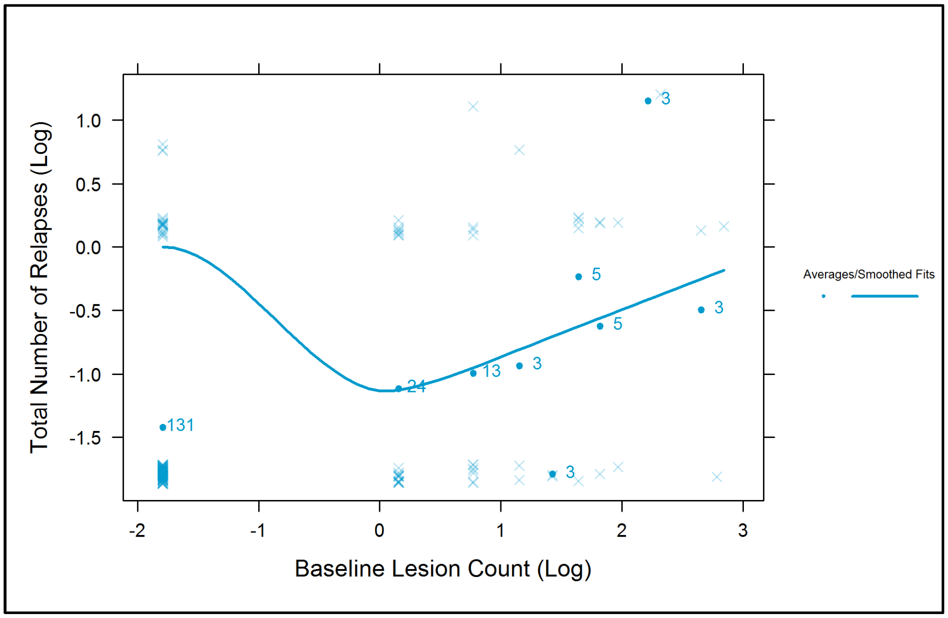
Table 3.2: Simulation parameters for the lesion distribution

Parameter	Value
Mean	1.1
Dispersion	5.0

Distribution of Relapses

Using Cohort I, we can assess the relationship between BLC and the total number of relapses descriptively (Figure 3.2); we ignore the treatment groups in this modelling as no treatment effect was identified in this trial.

Figure 3.2: Scatter plot of relapse vs. lesion count (on the log-scale)



Note: The LOESS curve uses a span of 2/3 of the data to fit first degree local polynomials to these data subsets

Figure 3.2 displays the raw data from Cohort I overlaid with averages and a curve based on locally weighted regression (LOESS). For baseline lesion counts less than 6, the averages are shown (alongside with the number of patients within each group) for each level of BLC. Since there are so few patients with more than 6 baseline lesions, the averages are taken over patients grouped by BLC and are plotted at the midpoints of the groups. From this plot, we see that there is some evidence of a positive relationship between BLC and the relapse count. This agrees with the findings of Wang et al. (2009) on Cohort II, who specify that relapse rates are higher for patients with lesions, and the demonstration that lesion activity is an adequate surrogate endpoint for relapse activity (Sormani & Filippi, 2007).

Negative binomial regression was used on Cohort I to model the mean relapse rate as a function of BLC. To perform this regression, the *glm.nb* function from the MASS package was used (Venables & Ripley, 2002). The canonical log link was utilized to relate the mean relapse rate to the predictor. The predictor was included in the model, first on an untransformed scale, and then on a log-transformed scale[1]. Both models support a positive relationship between lesion and relapse activity (p-values < 0.001), however the model using a log-transformed predictor has better model fit according to the Akaike information criterion (AIC). In practise, putting BLC on a log-scale may be more sensible because there are often large outliers in terms of lesion activity, but not as large outliers in terms of relapse activity. If kept untransformed, a large lesion count will be associated with an increase in the estimated mean relapse rate at an exponential rate, however the increase is on a linear scale when the log transformation is applied.

For Cohort I, the summary of the model fit indicates that a Poisson distribution provides an adequate fit to the relapse data, as θ, the estimated dispersion parameter for a negative binomial model is essentially 0, implying that the estimated variance is very close to the estimated mean. Oftentimes, overdispersion (i.e.: $\theta > 0$) relative to the Poisson is found for relapse activity in clinical populations, so we will instead use the estimate of the dispersion parameter based on Cohort II which is thought to be more representative of clinical populations in general. For Cohort II, the estimate of the dispersion parameter is approximately 1 (Wang, Meyerson, Tang, & Qian, 2009).

The model coefficients defining the relationship between mean relapse rate and log-transformed lesion activity also need to be specified for the placebo patients. As established above, this relationship is positive. We defined a system of equations such that those patients with 0 baseline lesions will have a relapse rate of 0.6 and those with 30 baseline lesions will have a relapse rate of 4. The solution to this

[1] Since lesion counts can equal to zero, we added a small arbitrary constant, 1/6, to the BLC in order that the log of this variable would be defined when BLC is equal to zero.

system results in an intercept parameter of 0.14 and a slope parameter of 0.36. Table 3.3 summarizes the parameters used in the negative binomial models to characterize the relapse distribution for placebo patients, while Table 3.4 provides some attributes of the relapse distribution for the population under the described parameterization of relapse activity.

Table 3.3: Simulation parameters for the relapse distribution of placebo patients

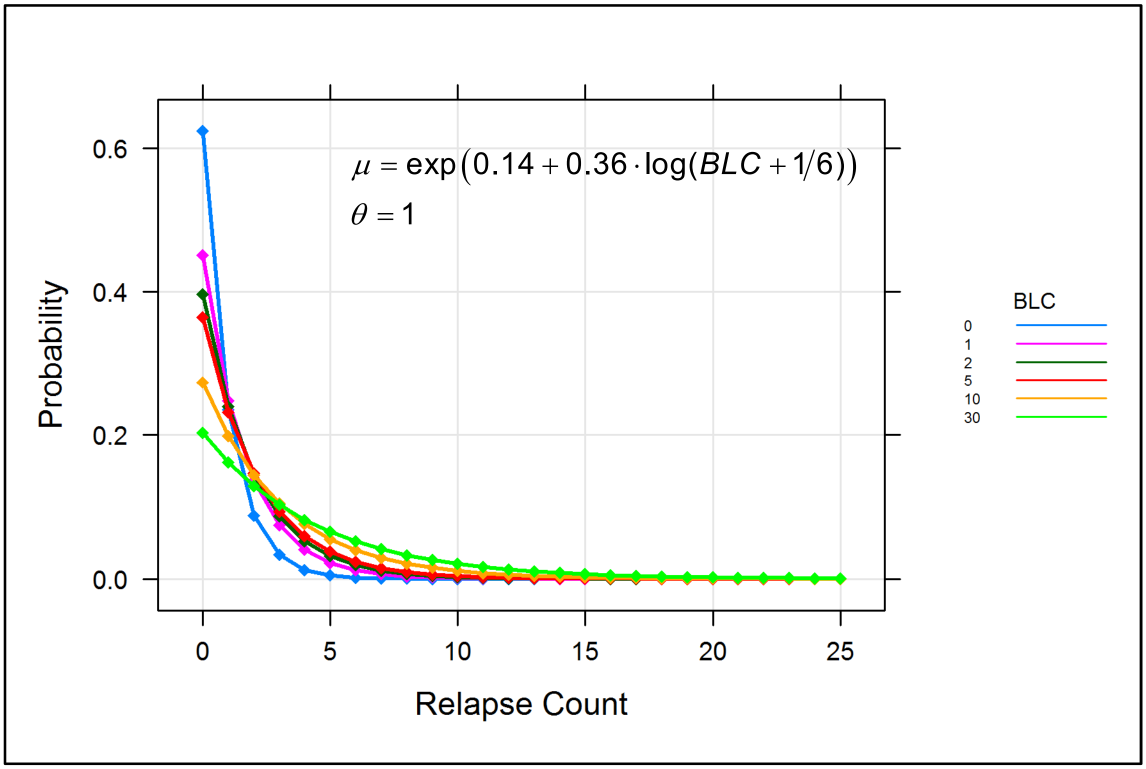
Parameter	Value
Mean	$\exp\left(0.14 + 0.36 \cdot \log(BLC + 1/6)\right)$
Dispersion	1.0

Table 3.4: Attributes of the relapse distribution for placebo patients

BLC	Mean Relapse	SD of Relapse
0	0.60	0.98
1	1.22	1.64
2	1.52	1.97
...	...	...
10	2.69	3.15
30	4.00	4.47

Figure 3.3 depicts the distribution of relapses across varying levels of baseline lesion activity using the specified parameters.

Figure 3.3: Distribution of relapse count for placebo patients with different baseline lesion counts



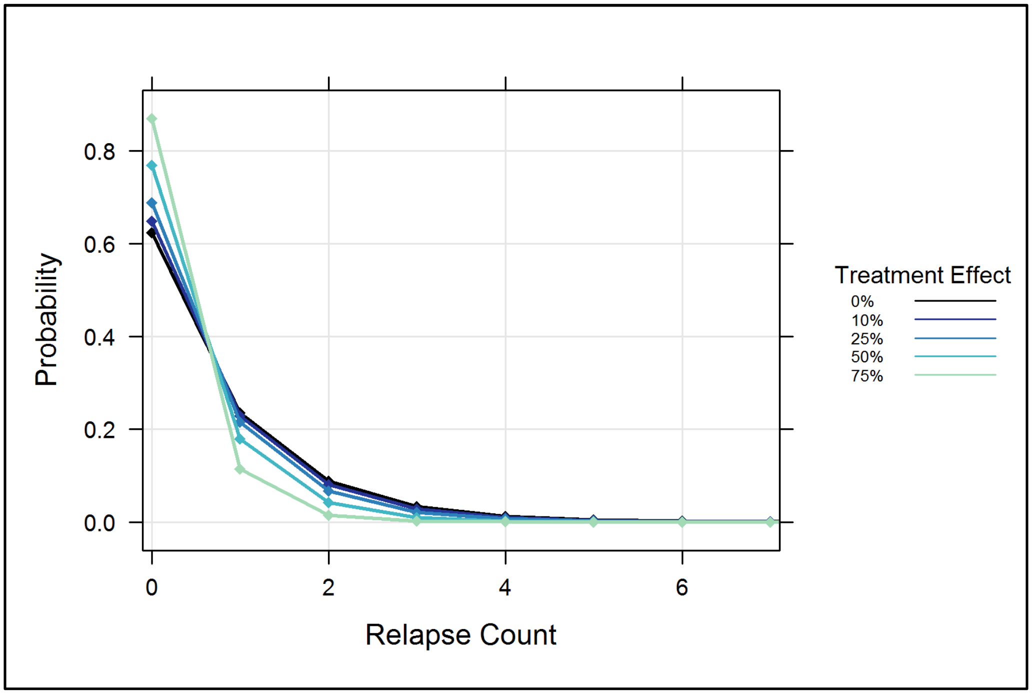
3.4.2 Simulation Scenarios Considered

Five general scenarios are considered, where each scenario corresponds to a different size of the patient subset within the treatment arm that receives benefit from treatment. This benefit is through a reduction in the mean relapse rate by some factor. These general scenarios include:

- Scenario 1: Treatment benefit is received by all patients in the treatment arm
- Scenario 2: Only the treated subset with 1 or more baseline lesions benefit
- Scenario 3: Only the treated subset with 2 or more baseline lesions benefit
- Scenario 4: Only the treated subset with 3 or more baseline lesions benefit
- Scenario 5: Only the treated subset with 4 or more baseline lesions benefit

Within each general scenario there are four sub-scenarios corresponding to treatment effects of varying magnitude. The magnitude of the treatment effects include: i) 0% reduction in the relapse rate of the treated group (i.e.: no treatment effect), ii) 10% reduction, iii) 25% reduction, iv) 50% reduction, and v) 75% reduction. The distribution of relapses for patients across different treatment effects for Scenario 1 is shown in Figure 3.4.

Figure 3.4: Distribution of relapses for treated patients for Scenario 1



3.4.3 Simulation Overview

The flow of the simulation is very similar to that used by JFS, which was described in detail in Section 2.2. For completeness, we will describe the main elements of the design of the simulation study here.

1. Each simulated population consists of 1,000 patients, with 500 patients each assigned to the control and treatment arms.
2. A baseline lesion count is generated for each patient from a negative binomial distribution with parameters as shown in Table 3.2:

$$BLC \sim NB(\mu = 1.1,\ \theta = 5.0)$$

3. For placebo patients, the generating distribution of the relapse rate is a negative binomial distribution with parameters as given in Table 3.3. For the treated patients who receive benefit (i.e.: have BLCs above the cut-off for the specified scenario), the generating distribution is the same as for placebo patients, except that the mean parameter is multiplied by a reduction factor, rr , where this reduction factor is 1 minus the treatment effect (TE):

$$\textit{Number of Relapses} \sim \begin{cases} NB\left(\mu_{Pl}, \theta = 1.0\right), \textit{ if the patient is in the control arm} \\ NB\left(\mu_{Pl}, \theta = 1.0\right), \textit{ if the patient is in the treated arm but has BLC < BLC cut-off} \\ NB\left(\mu_{Trt} = \mu_{Pl} \cdot rr, \theta = 1.0\right), \textit{ if the patient is in the treated arm with BLC} \geq \textit{BLC cut-off} \end{cases},$$

where $\mu_{Pl} = \exp\left(0.14 + 0.36 \cdot \log(BLC + 1/6)\right)$.

3.4.4 Results

Five thousand samples from the simulated population were generated. For each sample the OET and the adaptive procedures were performed to test the significance of the treatment effect as described in Section 3.3.2. The p-values for the adaptive procedures were calculated using an approximation to a permutation test in which 1,000 permutations of each sample were used to approximate the distribution of the test statistics under the null hypothesis of no treatment effect. Table 3.5 displays the empirical power results for each of the simulation scenarios presented in Section 3.4.2. This information is displayed graphically in Figure 3.5.

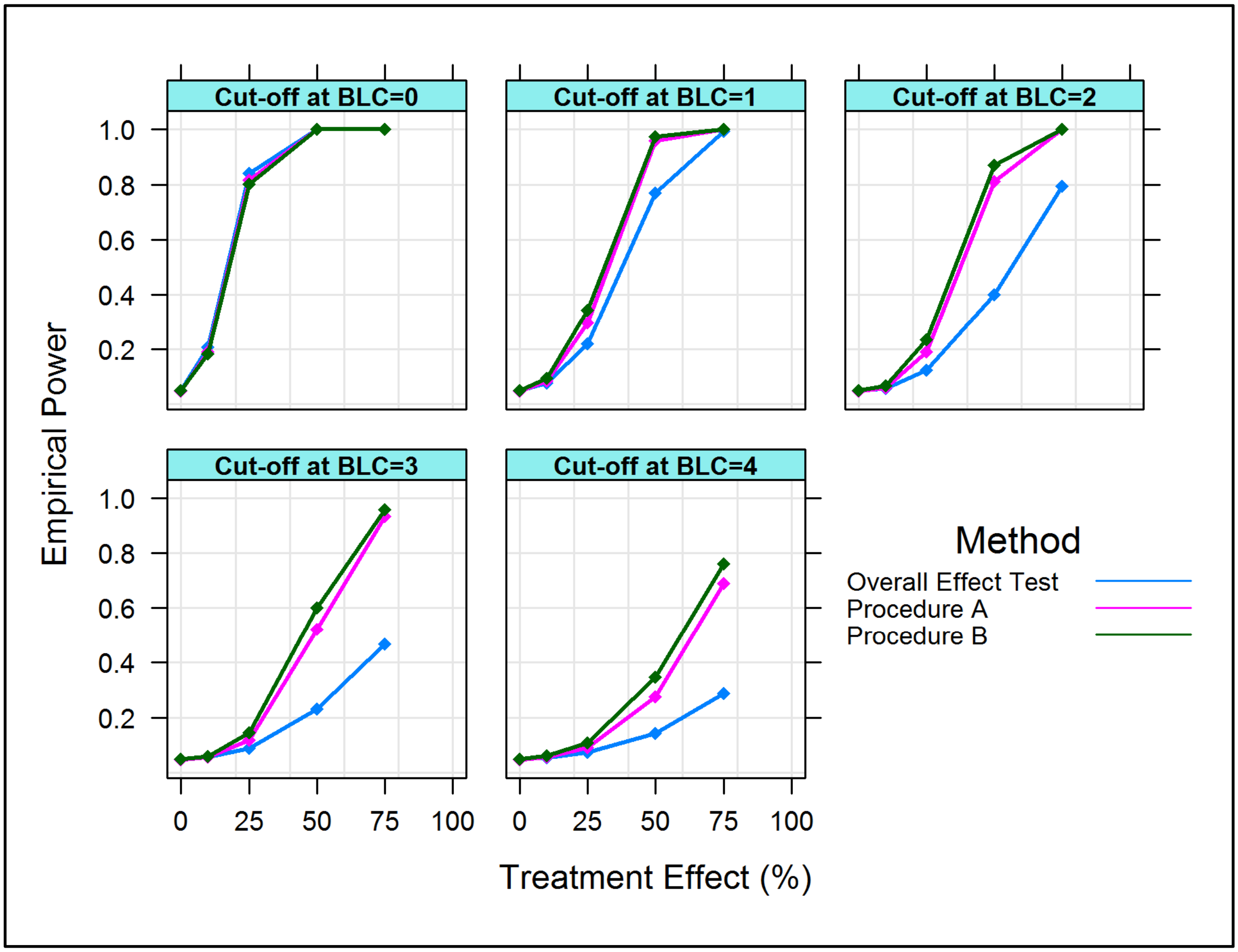
Table 3.5: Simulation results for a count outcome

Scenario		Model	Treatment Effect (%)	Overall Power*	Procedure A Power◊	Procedure B Power◊
1.	i)	Everyone treated benefits from new therapy	0	0.05	0.05	0.05
	ii)		10	0.21	0.19	0.18
	iii)		25	0.84	0.82	0.80
	iv)		50	1.00	1.00	1.00
	v)		75	1.00	1.00	1.00
2.	i)	Only patients with BLC ≥ 1 benefit from new therapy	0	0.05	0.05	0.05
	ii)		10	0.08	0.08	0.09
	iii)		25	0.22	0.30	0.34
	iv)		50	0.77	0.96	0.97
	v)		75	0.99	1.00	1.00
3.	i)	Only patients with BLC ≥ 2 benefit from new therapy	0	0.05	0.05	0.05
	ii)		10	0.06	0.06	0.07
	iii)		25	0.12	0.19	0.23
	iv)		50	0.40	0.81	0.87
	v)		75	0.79	1.00	1.00
4.	i)	Only patients with BLC ≥ 3 benefit from new therapy	0	0.05	0.05	0.05
	ii)		10	0.06	0.06	0.06
	iii)		25	0.09	0.12	0.15
	iv)		50	0.23	0.52	0.60
	v)		75	0.47	0.93	0.96
5.	i)	Only patients with BLC ≥ 4 benefit from new therapy	0	0.05	0.05	0.05
	ii)		10	0.05	0.05	0.06
	iii)		25	0.07	0.09	0.11
	iv)		50	0.14	0.28	0.35
	v)		75	0.29	0.69	0.76

*Overall test based on 5,000 generated samples

◊Procedures A and B based on 5,000 generated samples. The approximation of the permutation distribution was based on 1,000 permutations of the data and was used to calculate the p-value and find the empirical power for Procedures A and B

Figure 3.5: Empirical power as a function of treatment effect across BLC cut-off values



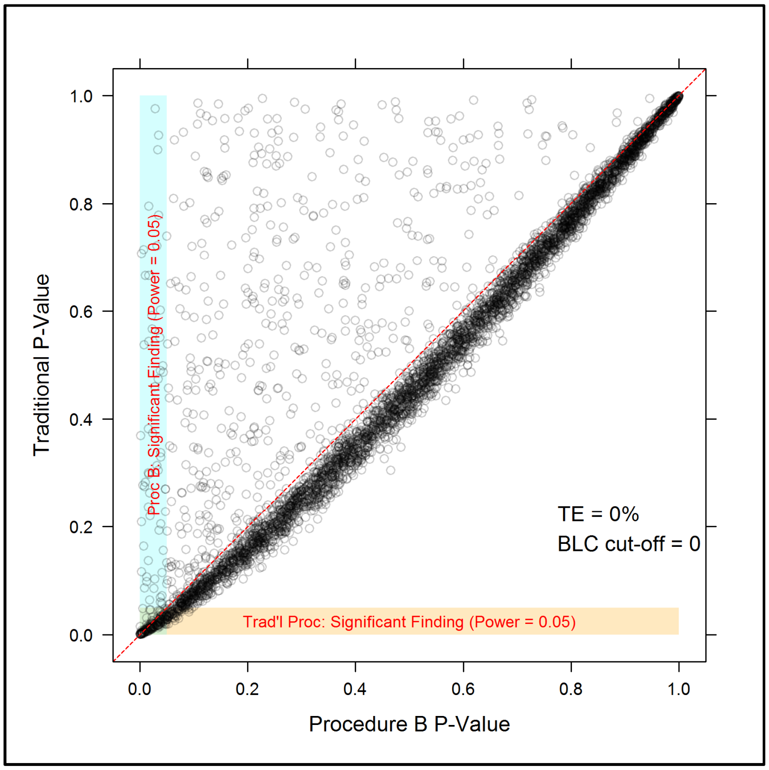
The adaptive procedures outperform the overall effect test in terms of empirical power across all of the scenarios for which the treatment effect is confined to a patient subset and in which the TE is larger than 10%. For the case where everyone benefits, the adaptive procedures perform nearly as well as the OET for the scenarios considered. In the case of no treatment effect, all the methods perform the same and have a 5% chance of detecting a treatment effect. This corresponds to the Type I error for the methods.

The adaptive procedures perform quite similarly. Procedure A slightly outperforms Procedure B in the case that the entire treatment group benefits and Procedure B performs slightly better than Procedure

A when benefit is confined to a patient subset. Compared with the OET, the adaptive procedures have the greatest gains in power for treatment effects which are confined to smaller subsets.

A comparison of the empirical power across the scenarios provides evidence that the adaptive procedures are, in general, superior at detecting a treatment effect for the simulated population when the treatment effect is confined to a patient subset. Also of interest is the performance of the adaptive procedures compared to the OET for specific samples. To investigate this, we first consider the p-values from the OET as compared to those from Procedure B for the 5,000 generated samples from Scenario 1.i) (i.e.: No treatment effect) in Figure 3.6.

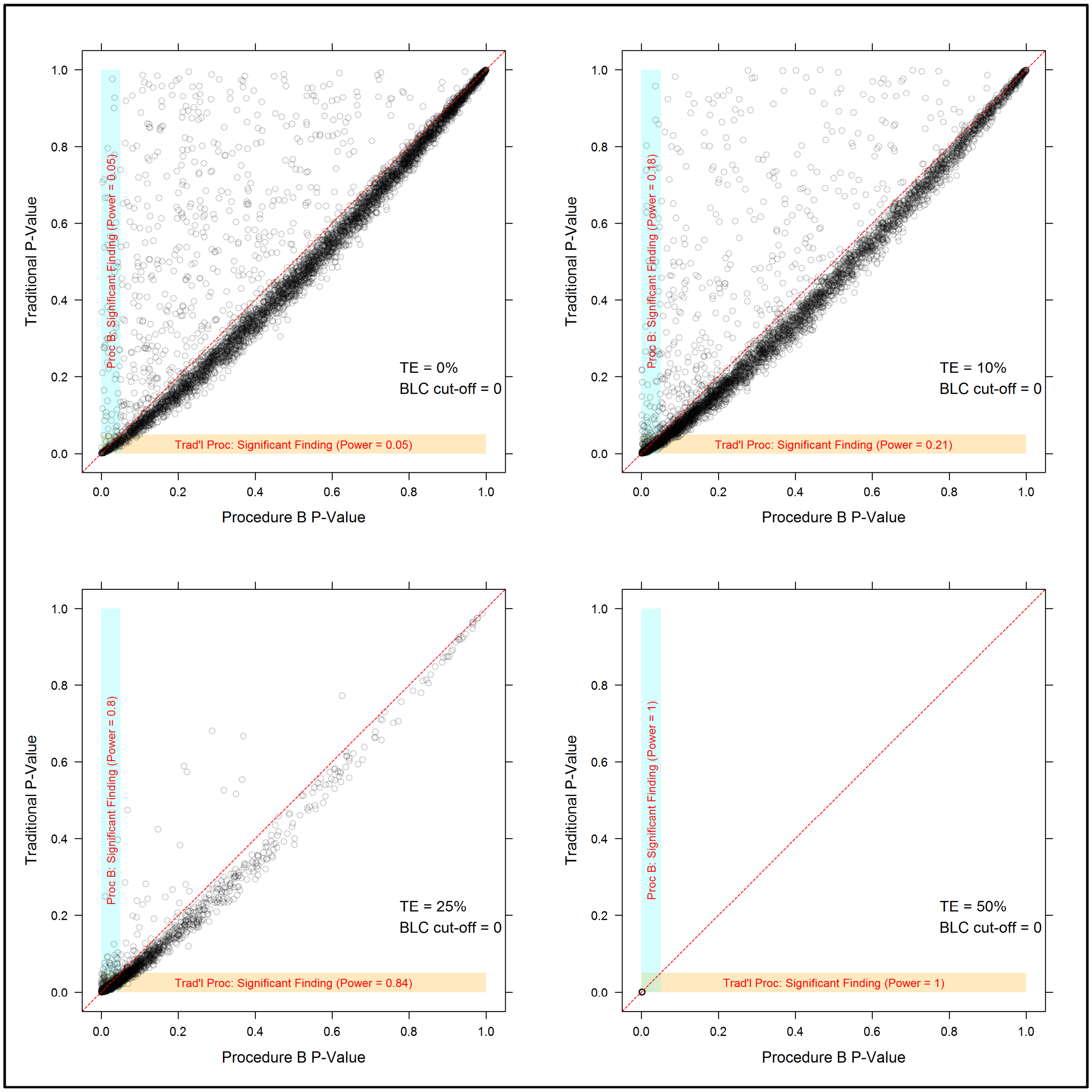
Figure 3.6: Comparison of p-values for Scenario 1.i) where TE = 0%



This figure is a scatter plot of the p-values calculated using the OET (on the y-axis) versus the p-values from Procedure B (on the x-axis). If a point lies in the unshaded region, then neither method has a significant treatment effect for the corresponding sample. A point which lies in the orange region has a significant p-value using the traditional method, whereas a point which lies in the blue region has a significant p-value using Procedure B. The labels on the orange and blue shaded regions denote the proportion of samples which lie in each of these regions. For this scenario, approximately 5% of the data lies in each of these shaded regions respectively, denoting the empirical Type I Error of these methods. The points in the lower left corner of the graph (where the shaded regions overlap) are considered to have statistically significant treatment effects using both of the methods. Lastly, the red dashed line denotes complete agreement between the two methods.

For ease of comparison, we display the scatter plots from Scenario 1.i) – iv) in Figure 3.7. We do not include the plot corresponding to Scenario 1.v) as it is the same as Scenario 1.iv).

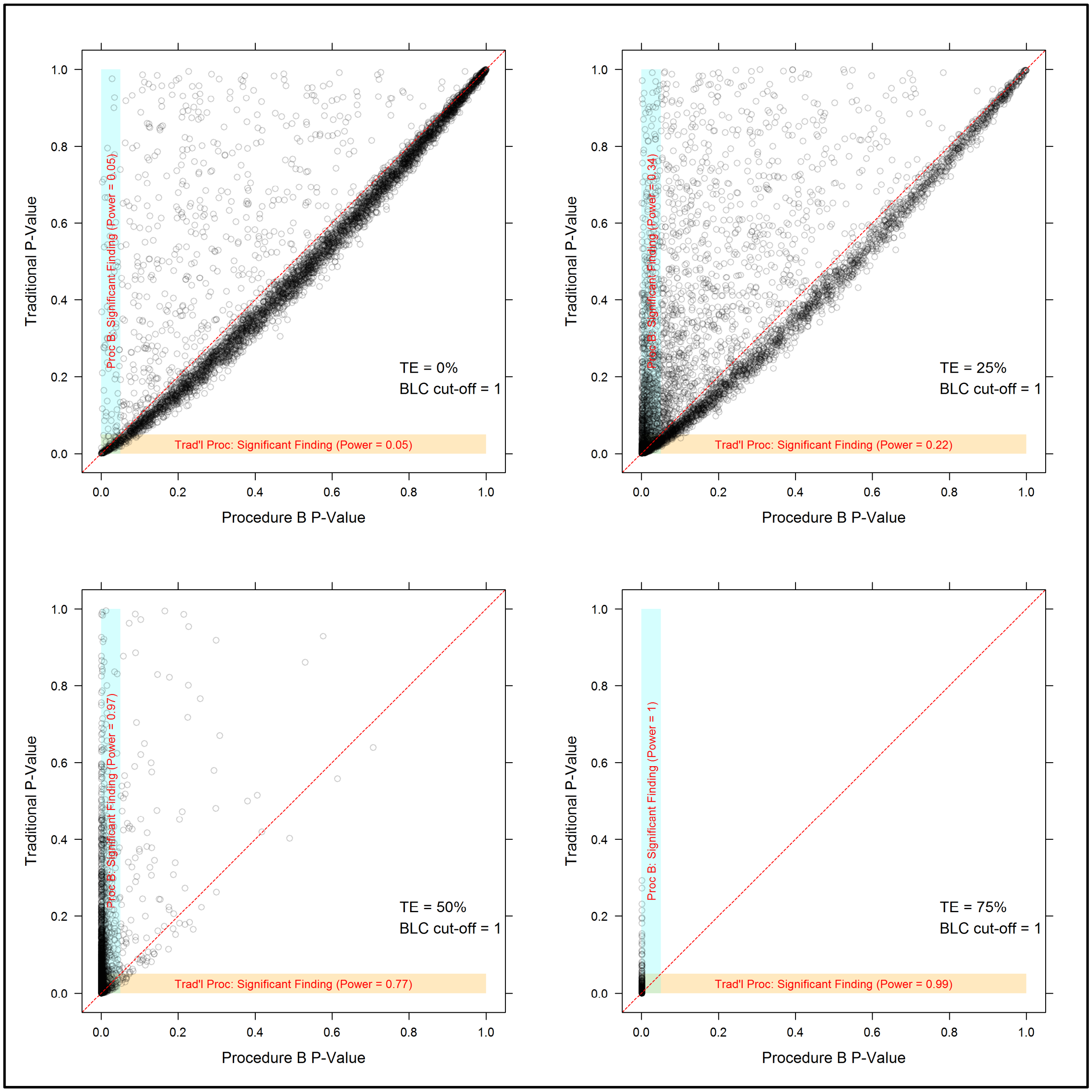
Figure 3.7: Comparison of p-values across Scenario 1 where everyone treated benefits



In the case of no treatment effect (TE=0%), much of the data lies close to the red diagonal, denoting nearly complete agreement between the methods. There are also a fair number of points above the diagonal and no points below the diagonal that are below the main cluster of data. As the treatment effect increases, we see a movement of points towards the lower-left corner of the graph, with slightly more points lying in the orange shaded region than in the blue shaded region for treatment effects of 10 and 25%. For a treatment effect of 50% all the samples shows significant treatment effects under both methods since the size of the study was powered for far smaller treatment effects.

Recall that Scenario 1 corresponds to the case in which everyone benefits from the treatment, and the remainder of the scenarios correspond to benefit being confined to a patient subset. The overall effect test was most highly powered for Scenario 1 and the relationship between the performances of the methods as noted for Figure 3.7 is not the same as what we find for the other scenarios, in which the adaptive procedures are more highly powered. Figures 3.8-3.11 provide a similar comparison of p-values for Scenarios 2-5.

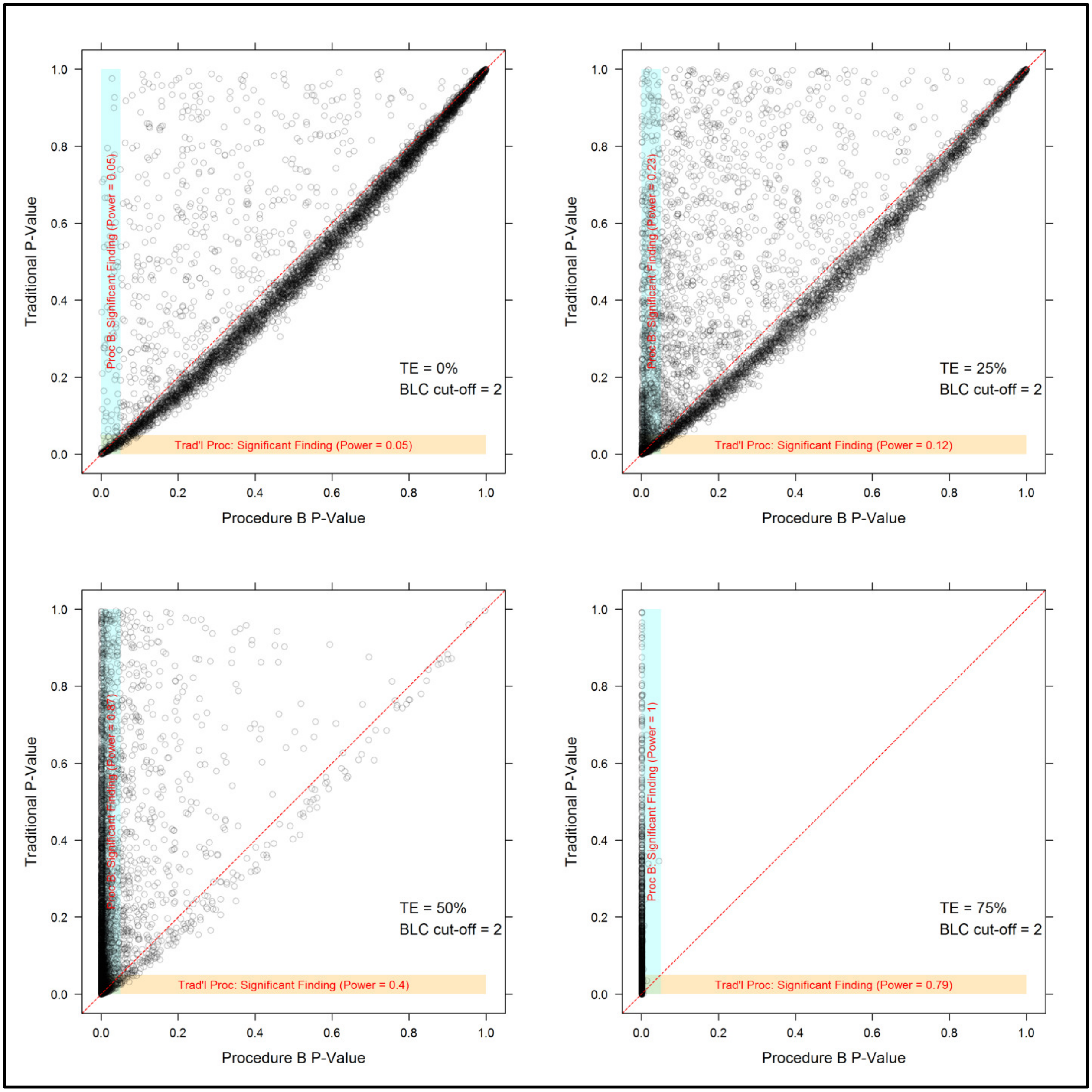
Figure 3.8: Comparison of p-values across Scenario 2 where the TE is confined to those with BLC ≥ 1



Scenario 2.ii) not depicted here (where TE=10%)

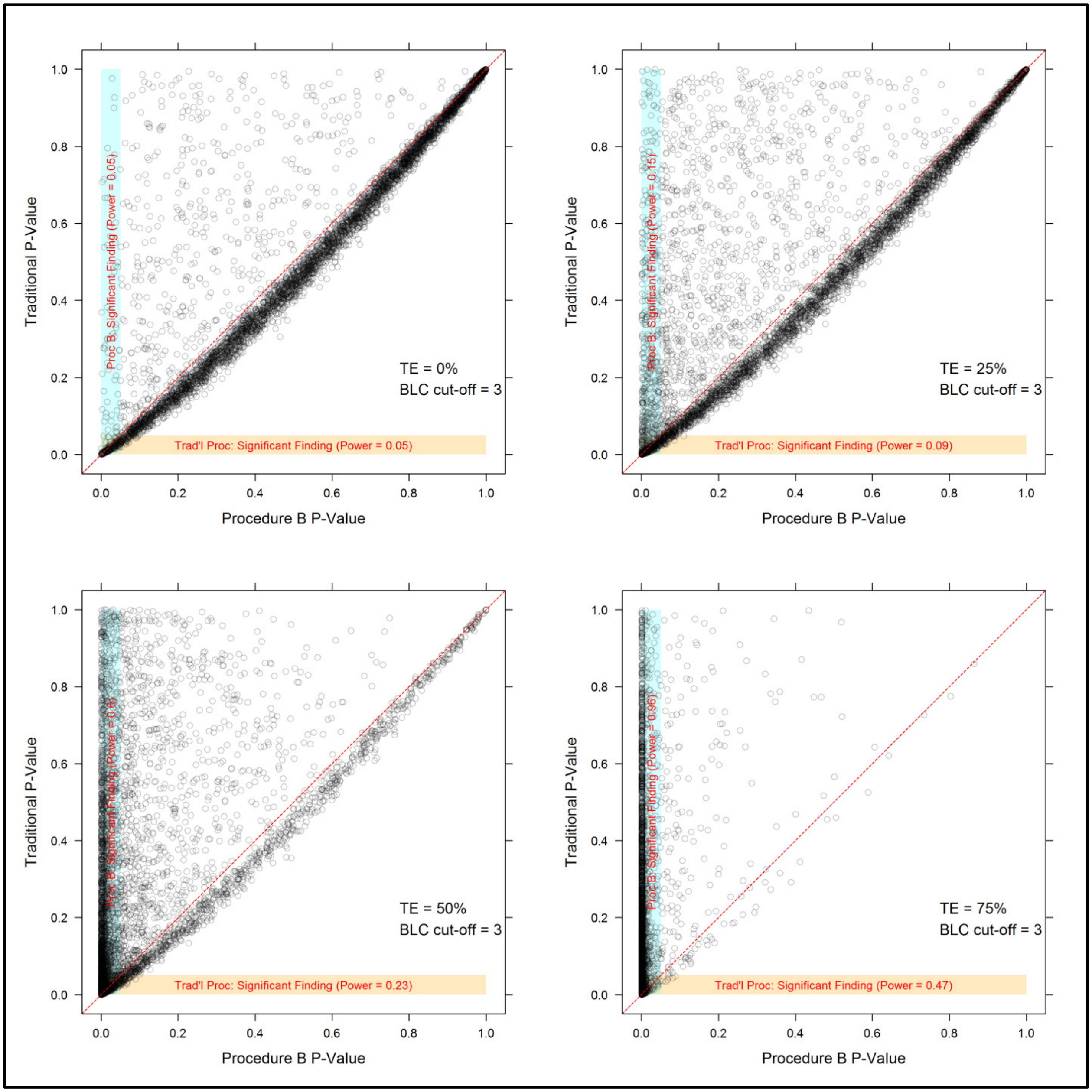
For Scenario 2 (Figure 3.8), as the treatment effect begins to increase (comparing the plots for TE=0% to TE=25%) we see that more points lie in the shaded regions and more points fall above the diagonal of the graph. Points which lie above the red dashed line are those with smaller p-values on Procedure B than on the traditional test. As the treatment effect becomes even larger, most samples have significant findings using both of the methods, but the adaptive method performs materially better than the traditional method when the treatment effect is 50%. In the extreme case of a very large treatment effect, Procedure B was able to detect this effect in all of the samples, while the OET detected the effect in nearly all of the samples.

Figure 3.9: Comparison of p-values across Scenario 3 where the TE is confined to those with BLC ≥ 2



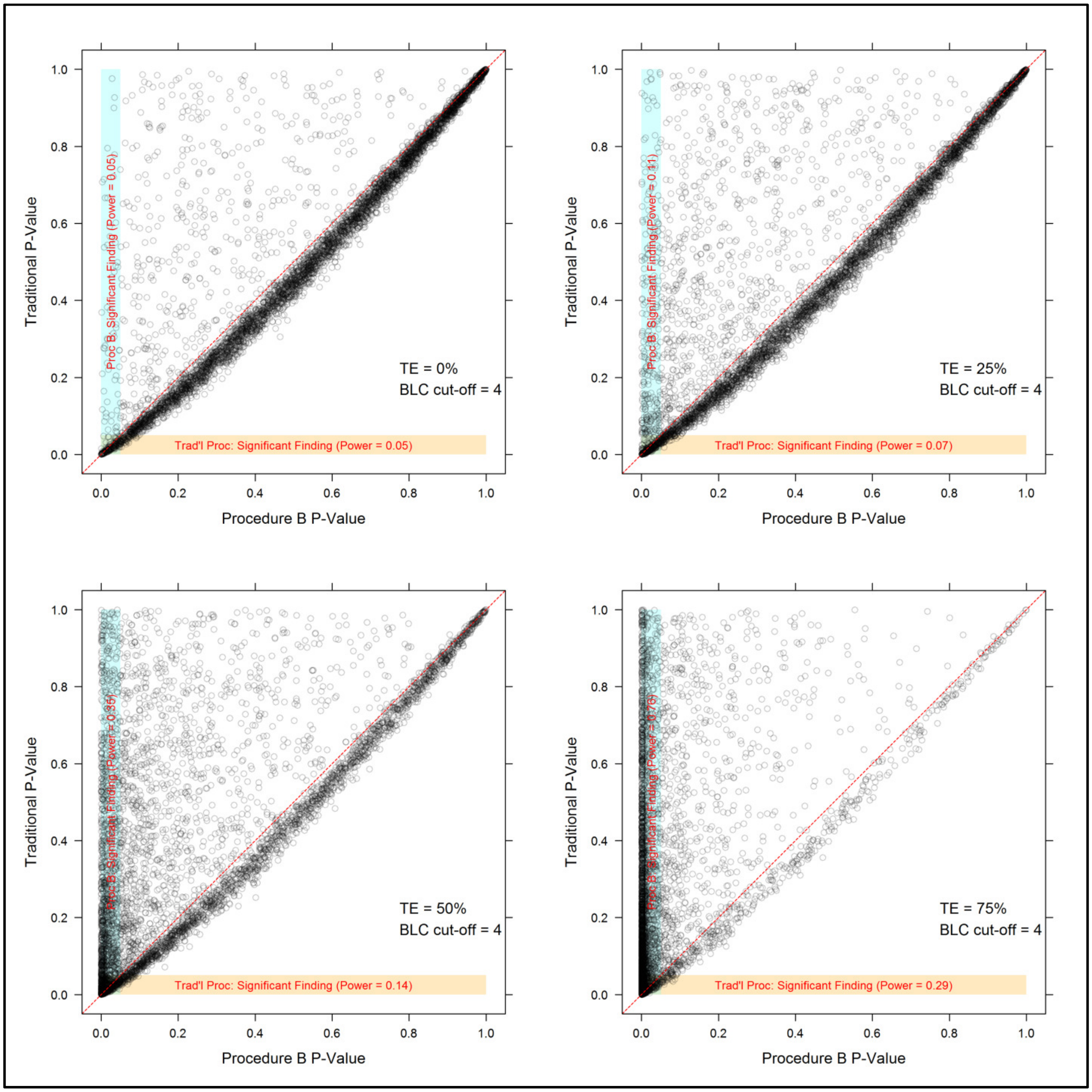
Scenario 3.ii) not depicted here (where TE=10%)

Figure 3.10: Comparison of p-values across Scenario 4 where the TE is confined to those with BLC ≥ 3



Scenario 4.ii) not depicted here (where TE=10%)

Figure 3.11: Comparison of p-values across Scenario 5 where the TE is confined to those with BLC ≥ 4



Scenario 5.ii) not depicted here (where TE=10%)

In Figures 3.9-3.11 we see the same general trends that were found for Scenario 2 (Figure 3.8) as the treatment effect increases. As well, we can make additional comparisons across figures while holding the treatment effect constant and increasing the cut-off denoting commencement of treatment benefit.

As the BLC cut-off increases (i.e.: as the size of the patient subset decreases), we see that the samples move from the shaded regions of significance towards the middle non-significant region.

Another way to quantify the gain of using the adaptive procedure is to consider the following question: Given that a treatment effect exists which would not be deemed significant by the OET, what is the probability that the adaptive method will find a significant difference? This conditional probability quantifies the gain in performance by using the adaptive procedure in the case that a treatment effect exists. Of course, this gain differs by treatment scenario. This probability is shown in the last column of Table 3.6 for each of the investigated scenarios.

Table 3.6: Conditional probability of finding a significant result using Procedure B

Scenario		Model	Treatment Effect (%)	P(Overall Significant)	P(B Significant| Overall Significant)	P(B Significant| Overall Insignificant)
1)	ii.	Everyone treated benefits from new therapy	10%	21%	82%	2%
	iii.		25%	84%	95%	3%
	iv.		50%	100%	100%	NA
	v.		75%	100%	100%	NA
2)	i.	Only patients with BLC ≥ 1 benefit from new therapy	10%	8%	77%	4%
	ii.		25%	22%	89%	19%
	iii.		50%	77%	100%	90%
	iv.		75%	99%	100%	100%
3)	i.	Only patients with BLC ≥ 2 benefit from new therapy	10%	6%	74%	3%
	ii.		25%	12%	85%	15%
	iii.		50%	40%	99%	79%
	iv.		75%	79%	100%	100%
4)	i.	Only patients with BLC ≥ 3 benefit from new therapy	10%	6%	73%	2%
	ii.		25%	9%	80%	8%
	iii.		50%	23%	95%	50%
	iv.		75%	47%	100%	92%
5)	i.	Only patients with BLC ≥ 4 benefit from new therapy	10%	5%	73%	2%
	ii.		25%	7%	77%	6%
	iii.		50%	14%	89%	26%
	iv.		75%	29%	98%	67%

All of the probabilities shown in the preceding table can be interpreted clinically. Let's first consider Scenario 2.ii), in which there was a relatively small treatment effect in a patient subset. Suppose that there were several trials conducted in a population like the one we considered in the simulation. Then of these trials, 22% would have found a significant result had the traditional test been used to detect a treatment effect. Now, of the trials which were considered significant using the traditional test, 89% of these, or roughly 9 trials out of 10, would have also been considered significant had Procedure B been used instead of the traditional procedure to detect the treatment effect. Thus, for Scenario 2.ii), the adaptive procedure is not detecting the treatment effect that would have otherwise been found using the traditional procedure about 10% of the time.

Now, consider the remaining 88% of trials which were considered non-significant according to the traditional procedure. Of these trials, for 19% of them, or for about 2 out of every 10, the adaptive technique would have successfully been able to detect the treatment effect.

Across Scenarios 2-5, we see that the probability that Procedure B will find a significant result given that one has been found using the traditional method is generally high and stays above 72%. Thus, by using the adaptive method there is only a small chance of missing the detection of a treatment effect, given that one exists. On the other hand, the gain in detection (corresponding to the probability that Procedure B will find a significant effect when the traditional method does not) increases and becomes very appealing for relatively large treatment effects.

Considering Scenario 1, there is still only a small chance (i.e.: between 0-18% across the sub-scenarios) of not detecting a treatment effect which would have otherwise been detected by the traditional procedure. However, the gain in detection is only very small. Thus, in the case that the overall group indeed benefits, the overall effect test remains superior.

Researchers are required to decide *a priori* which method to use. When making this decision, the trade-off between the loss and gain in the detection of a treatment effect will need to be weighed. Across the scenarios, however, the gain in detection is relatively high and the loss is relatively low, making Procedure B an attractive alternative to the OET when the treatment effect might be expected to be larger in a sensitive patient subset.

4 Application to a Real Dataset

In this chapter the developed methodology is applied to data from a previously completed clinical trial.

4.1 Overview of the Dataset

For this illustrative example, we use data from the North American Trial of IFNβ-1b in subjects with SP MS which was completed in 2000 (The North American Study Group on Interferon beta-1b in Secondary Progressive MS, 2004). The primary outcome for this study was the number of days until confirmed progression of one point in the Expanded Disability Status Scale (EDSS), while a secondary outcome was the ARR. The analysis included data from 939 patients, of which roughly a third were enrolled in each of the placebo, the IFNβ-1b 250 μg arm, or IFNβ-1b 160 μg/m^2 body surface area arm.

The study found no significant treatment effect on EDSS progression based on a pooled analysis to assess all IFNβ-1b vs. placebo (p-value = 0.71). A significant treatment effect was found on the ARR for the pooled analysis (p-value = 0.02), and for an analysis comparing the 250 μg arm to placebo (p-value=0.01), but not for the 160 μg/m^2 arm against placebo (p-value=0.20).

As introduced in Section 3.2 of this thesis, we are interested in whether two potential biomarkers are predictive of treatment effect. This chapter investigates how the modifications of the adaptive procedures A and B introduced in Section 3.3.2 perform relative to the OET. Since our methodology has been developed for count outcomes, we focus the analysis on the ARR as the outcome of interest. We first consider baseline lesion activity as the potential predictor of treatment effect, followed by an investigation of MS duration as the predictive marker.

4.2 Lesion Counts as the Predictive Marker of the Treatment Effect on ARR

As mentioned in the preceding section, the original study found a significant effect on ARR using data from the pooled analysis (both treatment arms vs. placebo) and for the 250 μg arm alone vs. placebo. Analysis of variance (ANOVA) was used to assess the significance of the treatment effect. While we still pool information across treatment arms to compare it against placebo, we do not use ANOVA to assess the significance of the treatment effect. Instead, we use a test of deviance comparing two nested negative binomial regression models to assess the relationship, since our modification of the BATD was designed using this method of analysis. Differences between the analytical techniques used to assess significance of the OET and the ones used in the original analysis can account for some of the difference seen in the reported significance levels.

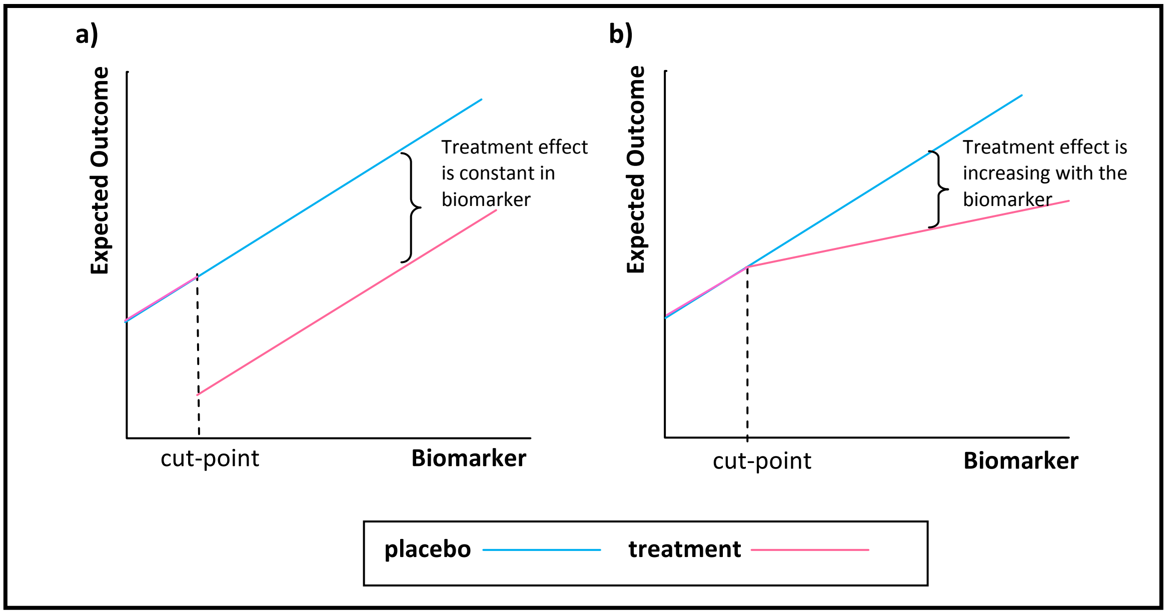
Frequent MRI measures, including the count of lesions at screening and baseline visits, were only taken on a sub-study population of 163 patients. For the investigation of lesion count as a predictive biomarker, we limit the analysis (including that of the OET) to this sub-study population. Additionally, eight of these patients did not have baseline or screening lesion counts recorded, so the analysis is further limited to 155 patients. The use of only a subset of the data accounts for some of the difference found in the significance levels reported in the original paper and that shown in this thesis.

Lastly, the predictive biomarker will be the average of the number of lesions at baseline and screening (referred to hereafter as ALC, for “average lesion count”), instead of just the number of lesions at baseline. For this population the screening visit took place before the baseline visit and using an average of these measures should provide a better (i.e. less variable) measure of baseline activity than one measure alone. Although the ALC is an average, the methodology developed for count biomarkers still applies as one can transform these averages into integers by multiplying them by two.

4.2.1 Descriptive Statistics

We first descriptively assess the relationship between ALC and the treatment effect. In particular, we would like to visually assess the notion of a predictive treatment effect of ALC on the ARR. In this setting, an ideal predictive biomarker influences the effect of treatment in one of the ways illustrated in Figure 4.1.

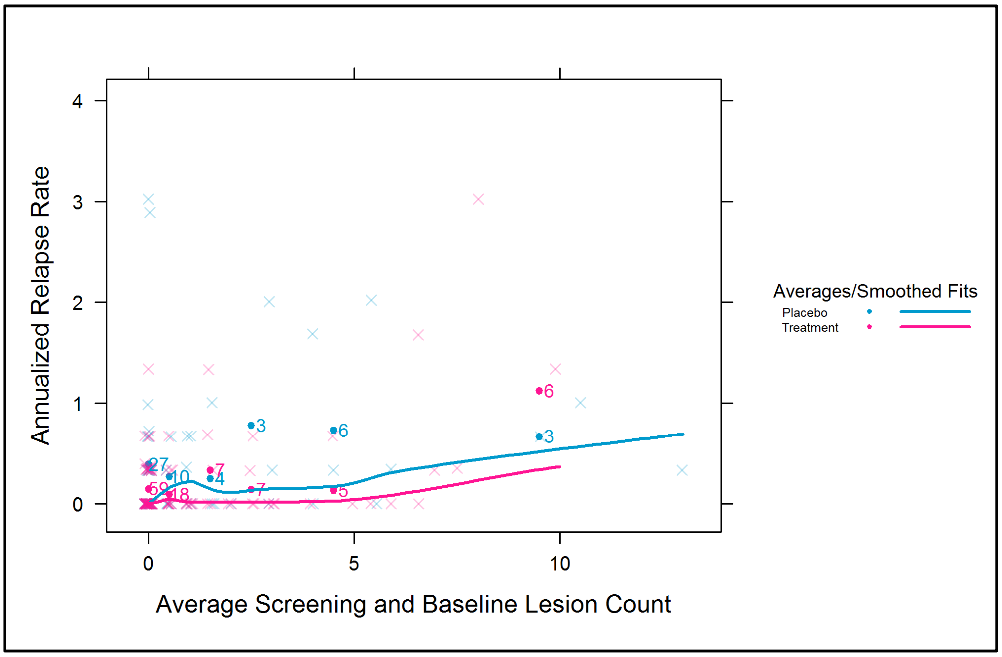
Figure 4.1: Idealized relationships between a predictive biomarker and the treatment effect



That is, at the biomarker cut-off value, the treatment effect begins (here, we assume a lowering of the outcome measure is desired). Different types of treatment effects can occur, of which two are shown in the above figure. Figure 4.1a) shows a treatment effect that is constant over the range of the biomarker after the cut-point is reached, while 4.1b) shows a treatment effect that increases with the biomarker. These are just two relationships, and other relationships are possible as well.

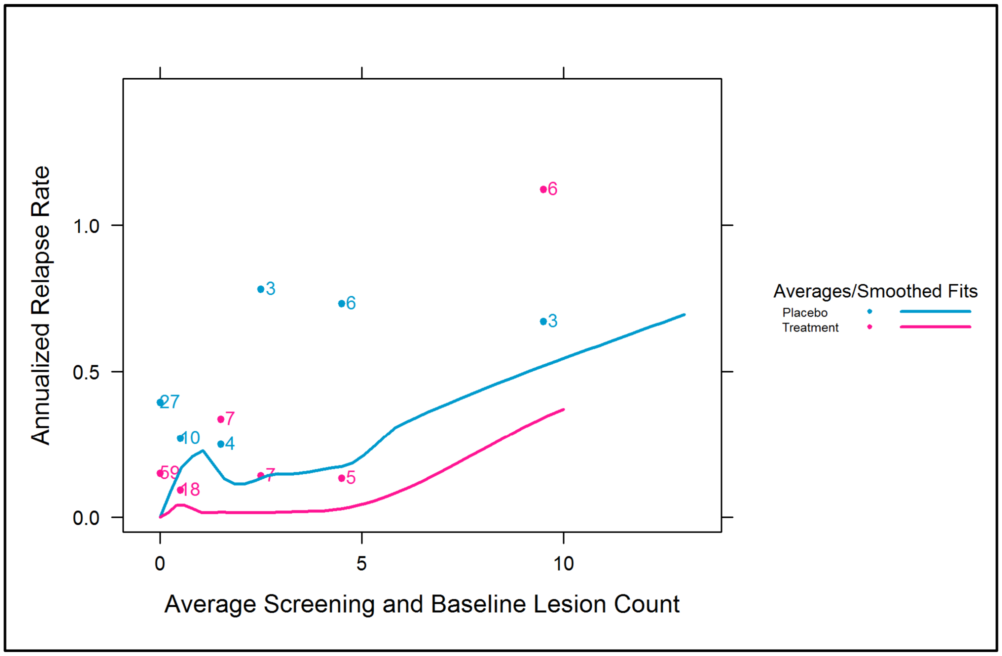
The actual relationship between the ALC and the treatment effect is illustrated in Figure 4.2, while Figure 4.3 shows the same plot without the raw data to allow for closer assessment of the relationships in the data.

Figure 4.2: Relationship between ALC and ARR



Note: The LOESS curve uses a span of $1/2$ of the data to fit first degree local polynomials to these data subsets

Figure 4.3: Relationship between ALC and ARR at close range



Note: The LOESS curve uses a span of $1/2$ of the data to fit first degree local polynomials to these data subsets

In these figures, two summary measures of the raw data are shown. Firstly, the average ARR for patients grouped by ALC are shown alongside with the number of patients in each grouping. These averages are plotted at the midpoint of the range considered. The intervals and sizes of the patient groupings are as follows:

- Group 1: ALC of 0, for a total of 86 patients
- Group 2: $0 < \text{ALC} \leq 1$, for a total of 28 patients
- Group 3: $1 < \text{ALC} \leq 2$, for a total of 11 patients
- Group 4: $2 < \text{ALC} \leq 3$, for a total of 10 patients
- Group 5: $3 < \text{ALC} \leq 6$, for a total of 11 patients
- Group 6: $\text{ALC} > 6$, for a total of 9 patients

The second descriptor is a fitted LOESS curve to the data according to treatment group.

From this plot, the treatment effect is evident across all levels of the biomarker, as we see a reduction in the ARR for the treated patients. There appears to be a positive relationship between the ALC and relapse activity. However, there does not appear to be a predictive biomarker relationship between the ALC and the ARR as there is no apparent trend in the magnitude of the treatment effect as measured by the distance between the LOESS curves or a discernable sudden increase in the magnitude as the biomarker increases in value. If the biomarker behaved as hypothesized, we would expect that the size of the gap between the smoothed curves would increase in the biomarker, but this does not appear to be the case.

4.2.2 Analysis

We used the developed methodology to compare the p-values obtained from the OET to those obtained from the adaptive tests. An offset term was included in the negative binomial regression model to account for the differing number of days that patients were followed. A summary of the output from the test for OET is shown here.

Table 4.1: Summary of output for the overall effect test adjusted for ALC

Parameter	Estimate	Standard Error	Wald test p-value
Intercept	-0.76	0.21	<0.001
Log(ALC + 1/6)	0.26	0.09	<0.01
Treatment	-0.67	0.27	0.01
Dispersion	0.71	0.20	NA

The test of the treatment effect was based on the deviance test. The deviance table for this test is shown here.

Table 4.2: Analysis of deviance table adjusted for ALC

Model	Deviance	Difference in deviance	p-value
Intercept only	147.46	NA	NA
Intercept + Log(ALC + 1/6)	137.92	9.54	<0.01
Intercept + Log(ALC + 1/6) + Treatment	131.87	6.05	0.01

Based on this analysis, the treatment effect is significant for the overall group. The BATD is most useful when the treatment effect is larger for a sensitive patient subset when compared to the overall group. Thus, this biomarker-response relationship is not an ideal one for the use of the BATD. However, it is still a worthwhile exercise to consider the p-values that would be provided in the case that an adaptive method was specified as part of the pre-specified analysis plan for the trial. Table 4.3 shows the p-values from the fixed and adaptive methods if these methods were applied to assess the significance of treatment in this study.

Table 4.3: P-values to assess the significance of treatment using ALC as the biomarker

Procedure	P-value
Overall effect test	0.014* or 0.028×
Procedure AŦ	Same as overall, since Stage A p-value < 0.04
Procedure BŦ	0.022

Notes:
Ŧ Permutation distribution used in adaptive procedures is based on 10,000 permutations of the dataset
*Uses the theoretical chi-squared distribution to evaluate the significance of the treatment effect
×Uses the permutation distribution to assess the significance of the treatment effect

The p-values from the OET and the adaptive procedures are quite similar in magnitude. For the OET, a p-value using the same permutation distribution as was used to evaluate the significance of the adaptive procedures is also included here for comparison purposes. This p-value is even closer to that given by Procedure B than the one based on the assumption that the overall effect test statistic follows a chi-squared distribution. No matter whether the adaptive or fixed method was used, the conclusion that the treatment has a significant treatment effect on the overall group would remain unchanged. As well, the point estimate for the cut-off is given by the cut-point which has the highest likelihood ratio test statistic. For this dataset, we considered cut-points ranging from 0 (i.e.: the overall effect test) to 4. No cut-points larger than 4 were considered because less than 10% of patients had ALC > 4. The corresponding likelihood ratio test statistics are shown in Table 4.4.

Table 4.4: Likelihood ratio test statistics across patients sub-grouped by ALC

Patient Subgroup	Number of Patients	Likelihood Ratio Test Statistic
ALC ≥ 0	155	6.62
ALC ≥ 0.5	69	2.14
ALC ≥ 1	52	1.78
ALC ≥ 1.5	41	0.69
ALC ≥ 2	33	1.40
ALC ≥ 2.5	30	1.18
ALC ≥ 3	26	0.92
ALC ≥ 4	20	0.04

From the preceding table, it is clear that the likelihood ratio test statistic based on using all the patients in the dataset is most significant, and so there is no need to further investigate the notion of ALC as a predictive biomarker for the treatment under investigation.

4.2.3 Conclusions

The OET for treatment effect and the adaptive methods both had significant results. Furthermore, the point estimate for the threshold denoting where the treatment effect begins was at an ALC level of 0, indicating that the best cut-point includes all patients. Thus, ALC does not appear to be a predictive biomarker. Fortunately, if investigators chose to use one of the adaptive procedures instead of the traditional test to perform their analysis, a significant treatment effect would still have been found.

4.3 MS Duration as the Predictive Marker

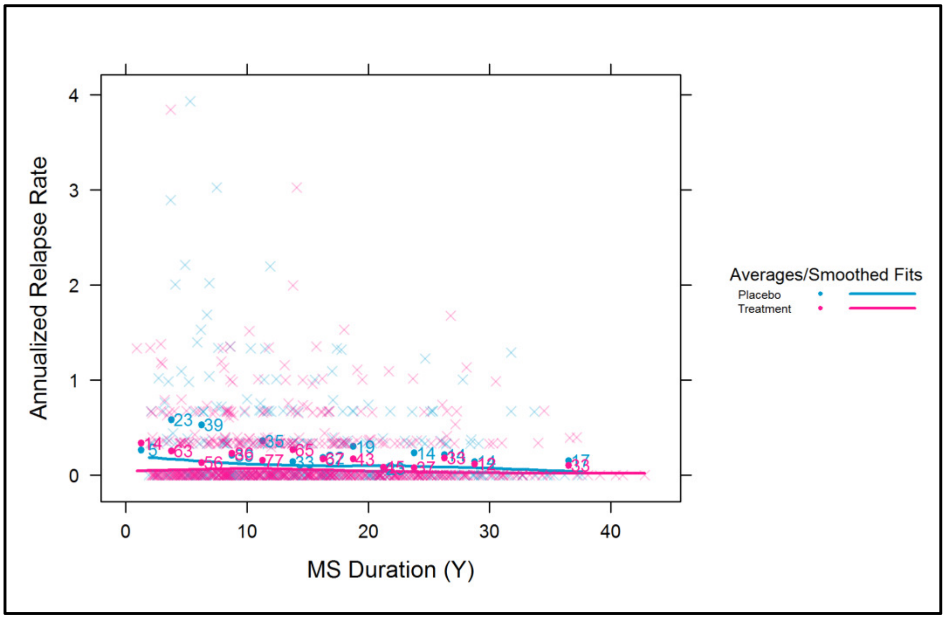
We described in Section 3.2.2 the hypothesis that MS duration is a predictive biomarker for the treatment effect on annualized relapse rate. Here, we hypothesize that the treatment effect increases as MS duration decreases. In order to use the adaptive method for scenarios with this property, we will consider patient subsets starting with the overall group, followed by those with biomarkers below a threshold, as opposed to above a threshold value.

Unlike ALC, which was only measured for a sub-study population, MS duration is known for all patients but one in the example dataset, so the procedures use data from 938 patients to assess the significance of the treatment effect. Again, we collapse the data across the treatment arms to assess the relationship of both treatment arms vs. placebo.

4.3.1 Descriptive Statistics

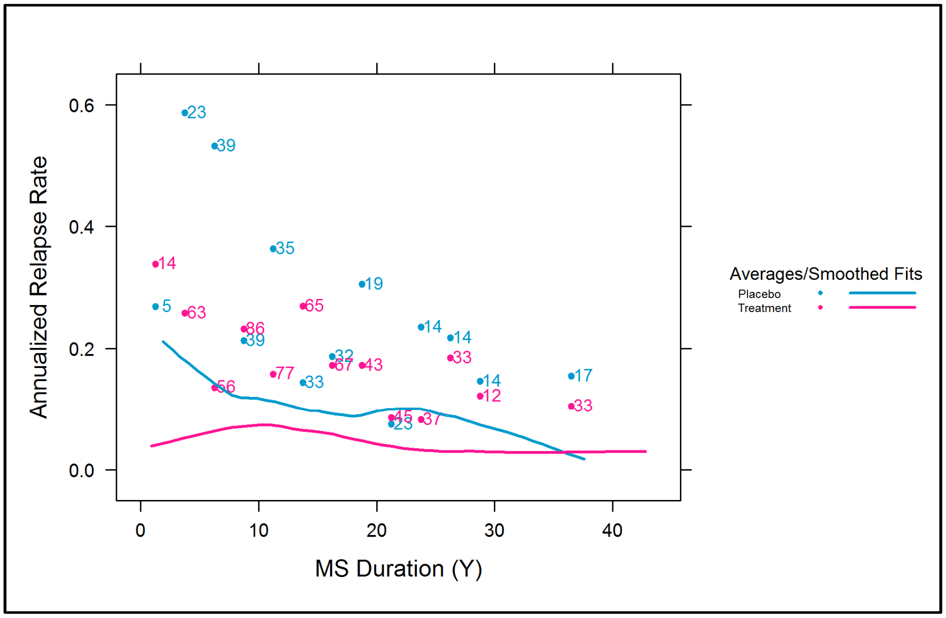
Figure 4.4 illustrates the relationship between MS duration and the ARR for patients from the placebo and treatment arms, while Figure 4.5 shows the same plot without the raw data to allow a more close assessment of the general relationships among the variables of interest.

Figure 4.4: Relationship between MS duration and ARR



Note: The LOESS curve uses a span of $1/2$ of the data to fit first degree local polynomials to these data subsets

Figure 4.5: Relationship between MS duration and ARR at close range



Note: The LOESS curve uses a span of $1/2$ of the data to fit first degree local polynomials to these data subsets

Similar to Figures 4.2 and 4.3, two summary measures are shown on the figures, including the fitted LOESS curves to the raw data and the average ARR from patient groups based on small intervals of data which are plotted at the midpoints of the intervals. From the plot, we see that ARR appears to decrease as a function of MS duration for both the placebo and treatment patients. This negative relationship agrees with what has been stated in a natural history study (Tremlett, Zhao, Joseph, & Devonshire, 2008). As well, the size of the treatment effect can be assessed descriptively based on the distance between the smoothed curves. There does appear to be some evidence supporting the hypothesis of a predictive biomarker relationship between MS duration and the treatment effect on ARR. We see that the largest treatment effect (assessed by the distance between the smoothed fits) is for patients that have had MS for less than ten years. As well, the treatment effect decreases as duration increases.

4.3.2 Analysis

Table 4.5 contains a summary of the model output for the OET.

Table 4.5: Summary of output for the overall effect test adjusted for MS duration

Parameter	Estimate	Standard Error	Wald test p-value
Intercept	-0.81	0.14	<0.001
MS Duration	-0.04	0.01	<0.001
Treatment	-0.41	0.12	<0.001
Dispersion	0.71	0.10	NA

Based on the parameter estimates from the table, both MS duration and treatment have a significant effect on the average rate of relapse, in directions which agree with what is seen in the plot of the data.

Table 4.6 contains the analysis of deviance table used to assess the significance of the treatment effect.

Table 4.6: Analysis of deviance table adjusted for MS duration

Model	Deviance	Difference in deviance	p-value
Intercept only	783.90	NA	NA
Intercept + MS Duration	760.60	23.29	<0.001
Intercept + MS Duration+ Treatment	748.97	11.64	<0.001

Not surprisingly, treatment is highly significant for the overall group, in agreement with the previous section’s analysis. Again, we perform the adaptive procedures to determine the p-value that would have resulted, had these procedures been used instead of the traditional analysis. We also find the p-value using a permutation test for the traditional test, to see how it agrees with the p-value based on the theoretical chi-squared distribution. These results are found in Table 4.7.

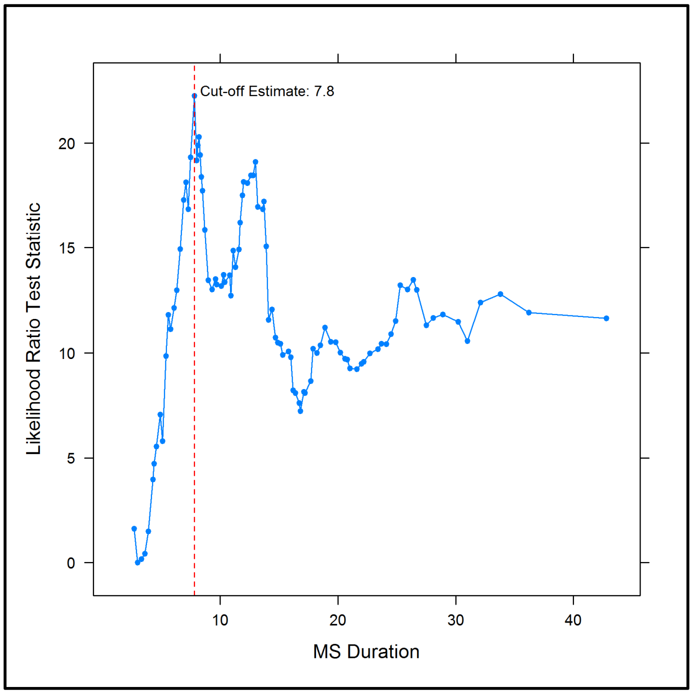
Table 4.7: Significance of the treatment effect using MS duration as the biomarker

Method	P-value
Overall	0.00065* or 0.00070[x]
Procedure A[Ŧ]	Same as overall since Stage A p-value < 0.04
Procedure B[Ŧ]	0.00070

Notes:
[Ŧ] Permutation distribution used in adaptive procedures is based on 10,000 permutations of the dataset
*Uses the theoretical chi-squared distribution to evaluate the significance of the treatment effect
[x]Uses the permutation distribution to assess the significance of the treatment effect

The p-values from the traditional and adaptive procedures agree very well. We next consider the point estimate of the biomarker cut-off denoting when the treatment effect commences. We considered patient subsets at every percentile of the data, beginning at the third percentile, since using any less than 3% of patients from this dataset results in convergence issues when conducting the statistical tests for treatment effects. As such, Figure 4.6 plots the likelihood ratio test statistic as a function of MS duration for durations with corresponding percentiles between 3-100%.

Figure 4.6: Likelihood ratio test statistic as a function of MS duration cut-off



Each point on the plot corresponds to the value of the likelihood ratio test statistic for the subset of patients with MS durations less than or equal to the corresponding x-value for that point.

From this plot we see that the threshold estimate corresponds to an MS duration of 7.8 years. That is, if a cut-point model exists (in which no treatment effect exists for patients with biomarker values above the cut-off and treatment effect exists for patients with values at the cut-off or below) then the most likely cut-off occurs at a duration of 7.8 years. Specifically, SP MS patients that have had MS for less than 7.8 years will benefit more from the drug (at least in terms of ARR) than patients with longer durations. Looking back to Figure 4.5 which shows the relationship between the treatment effect and MS duration as a predictive biomarker, we see that 7.8 years roughly corresponds to the beginning of a widening gap between the smoothed curves for placebo and treatment as duration decreases.

The technique of bootstrapping was used to find a confidence interval for the point estimate of the cut-off. Using 1,000 bootstraps, the 95% CI for our point estimate of 7.8 years is (6.5, 32.3).

4.3.3 Conclusions

Based on this analysis, it appears that MS duration may indeed be a predictive biomarker for the effect of IFNβ on ARR. However, the study was powered to detect the treatment effect in the overall group so the advantage of the adaptive procedures is not so apparent. As an exercise, the next section considers how the traditional and adaptive procedures would have performed if the study was not powered as highly by considering only a subset of the actual 938 patients used for this initial analysis.

4.4 MS Duration as the Predictive Biomarker using Data Subsamples

We first consider 500 samples, each containing 400 patients which have been sampled with replacement from the original dataset. The purpose of this initial investigation is to provide some general illustration of how the adaptive procedures perform when the OET is significant, and when the OET shows no significant result. Then, we perform additional analyses on one of the samples that had an insignificant OET.

4.4.1 Results Based on 500 Data Subsamples

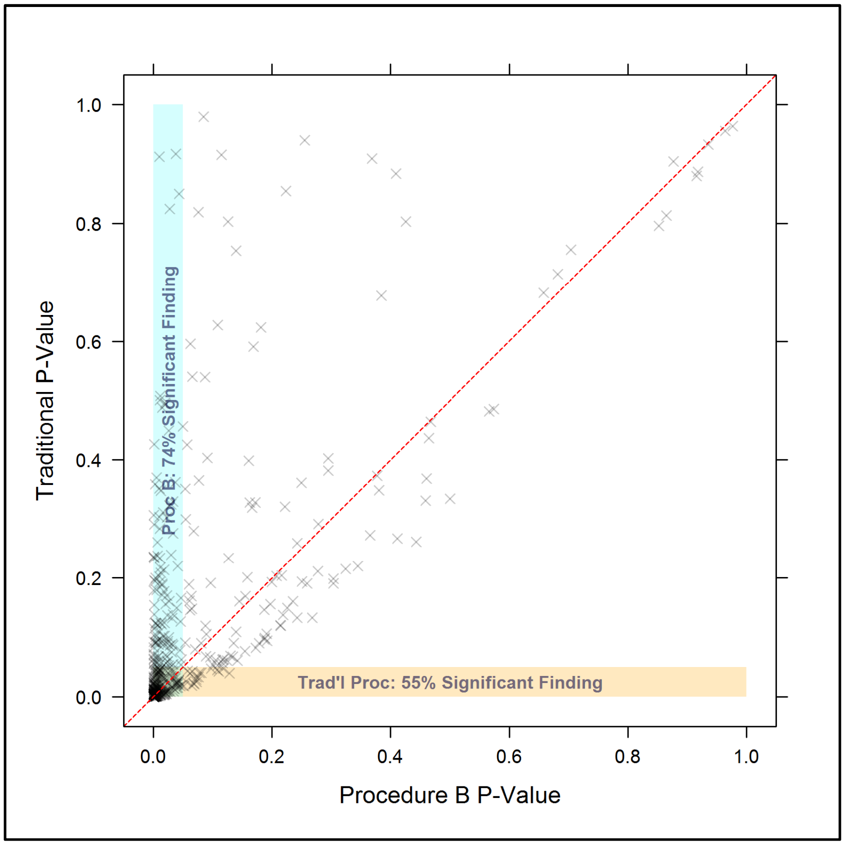
Five hundred samples of size 400 are drawn from patients that belong either to the placebo arm or the 160 μg/m^2 treatment arm. We did not include patients from the 250 μg arm in the samples in order to focus on the treatment effect between only the two arms (placebo vs. 160 μg/m^2 treatment arm). Table 4.8 is the contingency table which compares the significance results from the overall effect test and Procedure B for the subsamples.

Table 4.8: 2X2 table of the significance of the treatment effect determined by the overall test and Procedure B

		Procedure B P-value	
		Not Significant	Significant
Overall Effect	Not significant	111	113
Test P-value	Significant	19	257

About 51% of the random samples had a significant finding using both the overall effect test and Procedure B. Of the samples that had a significant finding on the overall test, more than 93% of these tests also had a significant finding on Procedure B. That is, if Procedure B is used instead of the traditional test for overall treatment effect, then there is only a 7% chance of missing a significant finding that would have been captured by the OET. As well, it is interesting to consider the chance that a treatment effect is missed when the OET is used. Of the samples which did not have a significant finding on the OET, just over half of these tests had a significant finding using the adaptive procedure. That is, using the adaptive procedure instead of the overall test would have meant finding a significant result 50% of the time when the traditional procedure would not have found any. For a more in-depth look at the relationship between the p-values given by the methods, a scatter plot of the p-values is shown in Figure 4.7.

Figure 4.7: Comparison of p-values for 500 subsamples of the data



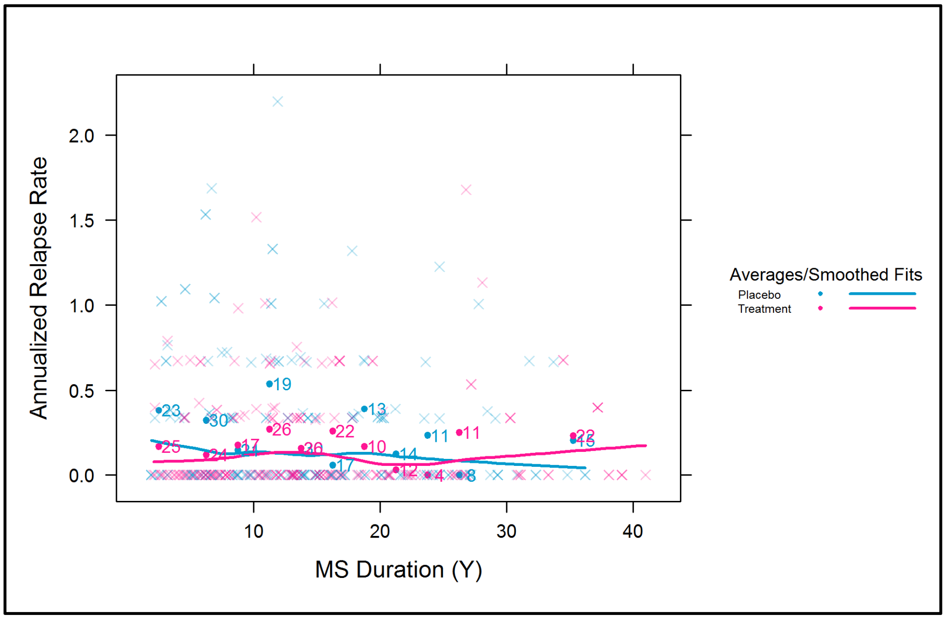
55% of the samples have significant results using the traditional procedure, while 74% of the samples found a significant treatment effect using Procedure B. The samples which are significant for each method have data points that fall within the orange and blue regions respectively. The red dashed line indicates complete agreement between the two procedures. We see that a fair proportion of the samples have data points close to this line. However, just over half of the samples have data points contained within the shaded lower-left corner of the graph – these samples have significant treatment effects using both of the procedures.

Based on this exercise we can conclude that a sample of size 400 drawn from a population with the distribution from this illustrative example will detect a treatment effect in the overall group about 55% of the time. For these samples, if the adaptive procedure was used instead of the OET, a significant result would have been found more than nine times out of ten. The benefit of the use of the adaptive procedures is most evident when considering those samples without a significant overall treatment effect. Of these samples, the adaptive procedures would have found a significant effect over 50% of the time, representing a sizable portion.

4.4.2 Descriptive Statistics for One Subsample

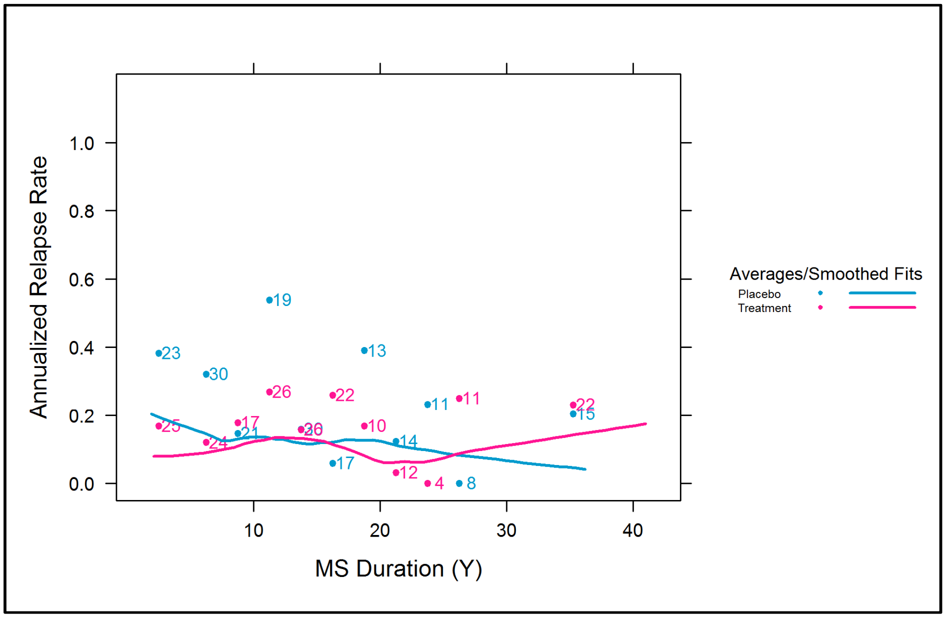
The preceding section provides evidence that the adaptive methodology was able to detect treatment effects that would be missed by the OET. In this section, we consider one such subsample and perform an analysis similar to what was done in Section 4.3. To select this subsample, we chose the first subsample generated that had an insignificant result on the OET. Once again, we begin the analysis by descriptively assessing the relationship between MS duration and the treatment effect on the ARR (Figure 4.8 and Figure 4.9 which does not include the raw data).

Figure 4.8: Relationship between MS duration and ARR in the subsample



Note: The LOESS curve uses a span of $1/2$ of the data to fit first degree local polynomials to these data subsets

Figure 4.9: Relationship between MS duration and ARR in the subsample at close range



Note: The LOESS curve uses a span of $1/2$ of the data to fit first degree local polynomials to these data subsets

The placebo group tends to show a decreasing ARR in the biomarker, as was also seen in the plot using all the data (Figure 4.5). The slight downward trend which was seen for the treatment group in Figure 4.5 is no longer apparent for this patient subsample. Based on the descriptive plot alone, it is difficult to assess the relationship between MS duration, the annualized relapse rate, and the magnitude of the treatment effect. Here, the analysis of the data will better quantify the relationships in the data.

4.4.3 Analysis for This Subsample

We first consider the OET, with a summary of the model output shown in Table 4.9.

Table 4.9: Summary of output for the overall effect test adjusted for MS duration for the subsample

Parameter	Estimate	Standard Error	Wald test p-value
Intercept	-1.16	0.17	<0.001
MS Duration	-0.02	0.01	0.09
Treatment	-0.26	0.16	0.10
Dispersion	1.1	0.30	NA

The regression parameter estimates for both MS Duration and Treatment are slightly smaller in magnitude than what we saw when all the data was used in the analysis. As well, the significance of the p-values are substantially lower, with a non-significant treatment effect as indicated in the corresponding p-value from the analysis of deviance table (Table 4.10).

Table 4.10: Analysis of deviance table adjusted for MS duration for the subsample

Model	Deviance	Difference in deviance	p-value
Intercept only	358.82	NA	NA
Intercept + MS Duration	355.46	3.46	0.06
Intercept + MS Duration+ Treatment	352.82	2.64	0.10

Since the overall effect test is insignificant, it is of interest to know how the p-values from the adaptive procedures will compare. Table 4.11 contrasts the p-values from the traditional and adaptive methods.

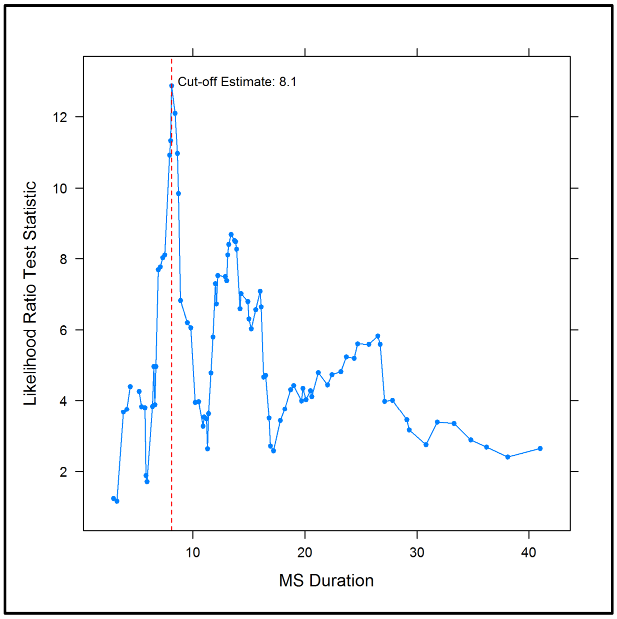
Table 4.11: Significance of the treatment effect using MS duration as the biomarker for the subsample

Method	P-value
Overall	0.10* or 0.09ˣ
Procedure AŦ	Stage 2 p-value: 0.004
Procedure BŦ	0.005

Notes:
Ŧ Permutation distribution used in adaptive procedures is based on 10,000 permutations of the dataset
*Uses the theoretical chi-squared distribution to evaluate the significance of the treatment effect
ˣUses the permutation distribution to assess the significance of the treatment effect

The difference in the p-values between the traditional and adaptive procedures is striking. While the overall test has an insignificant p-value near 0.10, the adaptive procedures were highly significant and suggest that a predictive biomarker relationship exists. As such, we estimate the biomarker cut-off by finding the threshold which maximizes the likelihood ratio test statistic (Figure 4.10).

Figure 4.10: Likelihood ratio test statistic as a function of MS duration cut-off in the subsample



Each point on the plot corresponds to the value of the likelihood ratio test statistic for the subset of patients with MS durations less than or equal to the corresponding x-value for that point.

Here, the estimate of the cut-off threshold corresponds to an MS duration of 8.1 years, which is close to the estimate using all the data. The corresponding 95% bootstrap CI is (3.4, 26.7).

4.4.4 Conclusions Based on the Analysis of Subsamples

Section 4.3 considers all of the data and found that both the OET and adaptive procedures produced significant findings. To investigate how the adaptive methods perform for studies that are less highly powered, Section 4.4 considers 500 subsamples of the data. In the case that the OET is significant, using the adaptive method instead results in a significant finding more than 90% of the time. However, when the OET is not significant, the adaptive procedure was able to detect a significant treatment effect more than half of the time, representing a large improvement in detection. Overall, the loss in detection associated with using the adaptive procedure appears to be small on average, whereas the gains appear to be quite large.

When we consider one subsample for which the OET was insignificant, the adaptive procedures detected a highly significant treatment effect, whereas the OET did not. This is suggestive of a predictive biomarker relationship between MS duration and the ARR such that lower durations correspond to larger benefit from treatment, in agreement with the findings of Section 4.3.

5 Generalizing the BATD to use Multiple Biomarkers

One could imagine that a sensitive patient subset may be characterized by more than one predictive biomarker. More precisely, many biomarkers may work in parallel resulting in different levels of sensitivity across patient subsets defined by various combinations of the levels of these markers.

In the genetics setting, where the set of potential markers is very large, the Adaptive Signature Design has been proposed for the described situation. The design is used to develop a sensitive patient classifier, which is defined by a combination of genetic markers, while testing for the treatment effect in the overall group (Freidlin, Jiang, & Simon, 2010). This design is suitable when the potential markers are based on genomic data.

In the case that only a few markers are to be considered and the directionality of the relationship among the markers and the treatment's effect on the response variable is hypothesized, an extension of the BATD is appropriate. For example, consider the two potential markers considered in Chapter 4: baseline lesion count and MS duration. To generalize the methodology for multiple markers, one could simply consider redefining the patient subsets for which the adaptive procedures' test statistics are based on by defining the subsets according to combinations of the levels of the two potential markers. That is, one patient subset for which the models are fit would include all patients who have had MS less than 5 years and have more than 2 lesions at baseline. By considering all possible combinations of levels of the markers one could define the test statistics using the same approach as is done based on one marker and also estimate the cut-off in the same manner as specified in the original methodology.

6 Conclusions

The Biomarker Adaptive Threshold Design can be generalized to accommodate a variety of biomarkers and responses. Through a simulation study, we have shown that the methodology can preserve statistical power across a variety of scenarios for which the treatment effect is confined to a sensitive patient subset. In the MS context, we have also shown that the adaptive design is able to detect a significant treatment effect on ARR, when considering MS duration as a predictive biomarker, which was not otherwise detected using a traditional approach to the analysis.

The design of a clinical trial is one of the most important aspects of any RCT as a well-planned and appropriate design is necessary in order for a study to have valid findings. One aspect of design is the sample size calculation. To perform this calculation, investigators define the smallest magnitude of the treatment effect that is of clinical interest. Traditionally, this treatment effect is thought to reflect the average benefit received by patients in the treatment arm as compared with patients in the placebo arm. If there is any evidence that treatment effect may vary with the level of a marker it is worth considering in advance how this relationship will affect the power of the test, especially if the treatment effect may be much larger in a small group. In this case, doing a sample size calculation which assumes the whole group will benefit will result in a realized power below the desired level.

The Biomarker Adaptive Threshold Design has important clinical implications because it allows researchers to investigate two research questions in sequence. Firstly, is there a significant treatment effect in the overall group? Secondly, if not, is there a significant treatment effect in a subgroup of patients according to the levels of a hypothesized potential biomarker? Use of the methodology elucidates the relationship between a potential marker, the response variable, and the treatment. If the methodology allows clinicians to pinpoint important markers indicative of treatment effect, than this finding can be used as a step towards providing personalized medicine to patients.

While the mindset behind the traditional clinical trial design involves the use of broad eligibility criteria, this often results in the over treatment of a patient population. The number needed to treat (NNT) is a concept which goes hand-in-hand with this. Treatments associated with relatively large NNT are those which require treating many patients in order for a few to benefit. A more efficient and arguably more ethical approach to treatment would put a greater emphasis on determining the characteristics of patients that are benefitting from treatment to distinguish them from those that do not benefit. Since data on a large-scale is collected during clinical trials, there is an opportunity to jointly investigate the magnitude and significance of the treatment effect and this secondary question regarding potential markers. The Biomarker-Adaptive Threshold Design is able to address both types of research questions while conserving statistical power across a variety of scenarios.

Bibliography

Beck, C. A., Metz, L. M., Svenson, L. W., & Patten, S. B. (2005). Regional variation of multiple sclerosis prevalence in Canada. *Multiple Sclerosis , 11* (5), 516-519.

Bielekova, B., & Martin, R. (2004). Development of biomarkers in multiple sclerosis. *Brain , 127* (7), 1463-1478.

Chang, M. (2008). *Adaptive Design Theory and Implementation Using SAS and R.* Boca Raton: Chapman & Hall/CRC.

Cook, T., & DeMets, D. L. (2010). Review of draft FDA adaptive design guidance. *Journal of Biopharmaceutical Statistics , 20* (6), 1132-1142.

Cox, D. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society , 34* (2), 187-220.

Ellenberg, S. S., Fleming, T. R., & DeMets, D. L. (2003). *Data Monitoring Committees in Clinical Trials: a Practical Perspective.* West Sussex: John Wiley & Sons Ltd.

Freidlin, B., Jiang, W., & Simon, R. (2010). The cross-validated adaptive signature design. *Clinical Cancer Research , 16* (2), 691-698.

Hosmer, D. W., & Lemeshow, S. (1999). *Applied survival analysis: regression modeling of time-to-event data.* New York: Wiley-Interscience.

Huitinga, I., Erkut, Z., van Beurden, D., & Swaab, D. F. (2004). Impaired hypothalamus-pituitary-adrenal axis activity and more severe multiple sclerosis with hypothalamic lesions. *Annals of Neurology , 55* (1), 37-45.

Jiang, W., Freidlin, B., & Simon, R. (2007). Biomarker-adaptive threshold design: A procedure for evaluating treatment with possible biomarker-defined subset effect. *Journal of the National Cancer Institute , 99* (13), 1036-43.

Kleinbaum, D. G., & Klein, M. (2005). *Survival Analysis: A Self Learning Text.* New York: Springer.

Lawless, J. F. (1987). Negative binomial and mixed poisson regression. *The Canadian Journal of Statistics, 15* (3), 209-225.

Mandrekar, S. J., & Sargent, D. J. (2009). Clinical trial designs for predictive biomarker validation: Theoretical considerations and practical challenges. *Journal of Clinical Oncology , 27* (24), 4027-4034.

Meyer, D., Zeileis, A., & Hornik, K. (2010). vcd: Visualizing Categorical Data. *R package version 1.2-8* .

National MS Society. (2010a, October). *About MS: National MS Society.* Retrieved March 31, 2011, from National MS Society: http://www.nationalmssociety.org/about-multiple-sclerosis/index.aspx

National MS Society. (2010b, December). *Treatments: National MS Society.* Retrieved March 31, 2011, from National MS Society: http://www.nationalmssociety.org/about-multiple-sclerosis/what-we-know-about-ms/treatments/index.aspx

Simon, R. (2010). Clinical trial designs for evaluating medical utility of prognostic and predictive biomarkers in oncology. *Personalized Medicine , 7* (1), 33-47.

Sormani, M. P., & Filippi, M. (2007). Statistical issues related to the use of MRI data in multiple sclerosis. *Journal of Neuroimaging , 17*, 56S-59S.

Sormani, M. P., Bonzano, L., Roccatagliata, L., Cutter, G. R., Mancardi, G. L., & Bruzzi, P. (2009). Magnetic resonance imaging as a potential surrogate for relapses in multiple sclerosis: A meta-analytic approach. *Annals of Neurology , 65* (3), 268-275.

Sormani, M. P., Bruzzi, P., Miller, D. H., Gasperini, C., Barkhof, F., & Filippi, M. (1999). Modelling MRI enhancing lesion counts in multiple sclerosis using a negative binomial model: implications for clinical trials. *Journal of the Neurological Sciences , 163* (1), 74-80.

The North American Study Group on Interferon beta-1b in Secondary Progressive MS. (2004). Interferon beta-1b in secondary progressive MS. Results from a 3-year controlled study. *Neurology , 63* (10), 1788-1795.

Tremlett, H., Zhao, Y., Joseph, J., & Devonshire, V. (2008). Relapses in multiple sclerosis are age- and time-dependent. *Journal of Neurology, Neurosurgery and Psychiatry , 79* (12), 1368-1375.

U.S. Department of Health and Human Services. Food and Drug Administration. (2010, February). *Guidance for Industry. Adaptive Design Clinical Trials for Drugs and Biologics.* Retrieved July 10, 2010, from U.S. Department of Health and Human Services. Food and Drug Administration: http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/ucm201790.pdf

Venables, W. N., & Ripley, B. D. (2002). *Modern Applied Statistics with S* (4th ed.). New York City: Springer.

Wang, Y. C., Meyerson, L., Tang, Y. Q., & Qian, N. (2009). Statistical methods for the analysis of relapse data in MS clinical trials. *Journal of the Neurological Sciences , 285* (1-2), 206-211.

Whitaker, J. N., McFarland, H. F., Rudge, P., & Reingold, S. C. (1995). Outcome assessment in multiple sclerosis clinical trials: a critical analysis. *Multiple Sclerosis , 1* (1), 37-47.

Yong, V. W., Chabot, S., Olaf, S., & Williams, G. (1998). Interferon beta in the treatment of multiple sclerosis. Mechanisms of action. *Neurology , 51* (3), 682-689.

Zhao, Y., Petkau, A. J., Traboulsee, A., Riddehough, A., & Li, D. (2010). Does MRI lesion activity regress in secondary progressive multiple sclerosis? *Multiple Sclerosis , 16* (4), 434-442.