Magnetic Resonance Imaging Lesion Count as a Surrogate Endpoint in Relapsing-Remitting Multiple Sclerosis Clinical Trials

by

Lang Qin

B.Sc., Jinan University, 2009

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

Master of Science

in

The Faculty of Graduate Studies
(Statistics)

THE UNIVERSITY OF BRITISH COLUMBIA
(Vancouver)

August 2011

Abstract

The count of active lesions based on magnetic resonance imaging (MRI) is often used as a potential surrogate endpoint in phase 2 clinical trials for relapsing-remitting multiple sclerosis (RRMS) patients. However, this surrogacy relationship has not been completely validated. In this report, we study whether at the trial level, the MRI lesion count is a good surrogate endpoint for the relapse rate, the usual clinical endpoint for RRMS clinical trials.

Two different approaches to assess this surrogacy relationship are applied to the dataset used by Sormani et al. [1] (SBRCMB) which contains the summary results from 23 randomized, placebo-controlled clinical trials in RRMS. The SBRCMB approach uses simple linear regression with weighted least squares estimation, while our more comprehensive approach develops a detailed model for the endpoints and the treatment effects to take into account estimation errors and the correlated contrasts. Both approaches are based only on the summary results from each clinical trial.

The shortcomings of the SBRCMB approach are discussed and the results from the two approaches are compared. Both approaches show that the MRI lesion count is a good surrogate endpoint, while our more comprehensive approach shows a nearly perfect surrogacy relationship. When the estimated surrogacy relationship is used to predict the true treatment effect on the clinical endpoint for the trials in the SBRCBM dataset, the approaches give similar point predictions, but

the approximate 95% prediction intervals from the comprehensive approach are generally shorter. In practice, the estimated surrogacy relationship based on the comprehensive approach can give a precise prediction for the true treatment effect on the clinical endpoint if the treatment displays a large effect on the surrogate endpoint, but may otherwise lead to an inconclusive result.

Table of Contents

Abstract ii

Table of Contents iv

List of Tables vii

List of Figures viii

Acknowledgments ix

Dedication x

1 Introduction 1

1.1 What is a Surrogate Endpoint? 1

1.2 Surrogate Endpoints in Multiple Sclerosis 2

1.3 Outline of the Report 4

2 Literature Review: Validation of Surrogate Endpoints 5

2.1 Importance of Validating a Potential Surrogate Endpoint 5

2.2 Methods of Validating Surrogate Endpoints 8

2.2.1 The Prentice Operational Criteria for Validation 8

2.2.2 Validation in a Single Clinical Trial 9

2.2.3 Validation in Multiple Clinical Trials 12

2.3 Validation in Multiple Clinical Trials with Individual Data Unavailable 15

2.3.1 Review of Daniels and Hughes [2] 16
2.3.2 Review of Korn et al. [3] 19
2.3.3 Comparison of These Two Approaches 24

3 Lesion Counts as a Surrogate Endpoint in RRMS: the SBRCMB Approach 26
3.1 Introduction and the SBRCMB Dataset 26
3.2 The SBRCMB approach 28
3.3 Critique of the SBRCMB Approach 33
3.3.1 The Appropriateness of the Weights 34
3.3.2 Correlation of the Contrasts 39
3.3.3 Influence of Estimation Errors 40

4 Lesion counts as a Surrogate Endpoint in RRMS: A More Comprehensive Approach 44
4.1 Model for the Single-contrast Clinical Trials 45
4.1.1 Model for the True Treatment Effects 45
4.1.2 Model for the Observed Annualized Relapse Rate and MRI Lesion Count Per Patient Per Scan 47
4.1.3 Model for the Estimated Treatment Effects 48
4.2 Model for the Multiple-contrast Clinical Trials 50
4.3 Parameter Estimation 54
4.4 Comparison between the Comprehensive Approach and the SBRCMB Approach 59
4.5 Assessment of the Estimated Surrogacy Relationship in Practice . 72

5 Conclusions and Discussion 79

Bibliography 85

Appendices 88

A The SBRCMB Dataset 88

B Partial Derivatives of $E(Y_0^{true}|X_0 = x_0)$ 91

List of Tables

Table 3.1 Results of the Sensitivity Study 30
Table 3.2 Results of the Interaction Study 31

Table 4.1 Results of the Model Fit 57
Table 4.2 Comparison of the Approximate 95% Prediction Intervals for $\exp(Y_0^{true}(x_0))$ for the SBRCMB and Comprehensive Approaches 68
Table 4.3 Influence of the Sample Size N_0 and the Magnitude of the Estimated Treatment Effect on the Surrogate Endpoint on the 95% Prediction Intervals for the True Treatment Effect on the Clinical Endpoint for Trials with $K_0 = 6$ Scans per Patient. The Entries are the Point Predictions and Approximate 95% Prediction Intervals for $\exp(Y_0^{true}(x_0))$. 73

List of Figures

Figure 2.1 Scenarios of Perfect (a) and Imperfect (b,c,d) Surrogates . . . 7

Figure 3.1 Scatter Plot of Estimated Treatment Effects 27
Figure 3.2 Results of the Validation Study 33
Figure 3.3 Scatter Plot of $(c, 1/w)$ 38
Figure 3.4 Scatter Plot of $(c, 1/w')$ 38

Figure 4.1 Regression Prediction Lines: the SBRCMB Approach ($y = -0.02 + 0.55x$) and the Comprehensive Approach with $K_0 = 6$ and $N_{a0} = N_{c0} = 50$ ($y = 0.50x$). 63
Figure 4.2 Regression Prediction Lines: the SBRCMB Approach ($y = -0.02 + 0.55x$) and the Comprehensive Approach with $K_0 = 6$ and $N_{a0} = N_{c0} = \infty$ ($y = 0.08 + 0.62x$). 64
Figure 4.3 Point Predictions for the 40 Contrasts 66
Figure 4.4 Comparison of Point Predictions for the 40 Contrasts 66
Figure 4.5 Comparison of the Approximate 95% Prediction Intervals for $\exp(Y_0^{true}(x_0))$ for the SBRCMB and Comprehensive Approaches 71
Figure 4.6 Threshold Value of $\exp(X_0)$ versus Sample Size N_0 when a Beneficial Treatment Effect is Observed on the Surrogate Endpoint 77
Figure 4.7 Threshold Value of $\exp(X_0)$ versus Sample Size N_0 when a Negative Treatment Effect is Observed on the Surrogate Endpoint 77

Acknowledgments

I would like to express my sincerest gratitude to my supervisor, Professor John Petkau. I could never finish my thesis without his insightful guidance and constant encouragement. I am also grateful for his enlightening teaching in STAT550 and STAT551, from which I started to learn how to think as a statistician. I would like to thank Professor Lang Wu for being my second reader and providing valuable comments.

I would like to thank my best friend Guannan Li, to whom I can always express my sadness when I am depressed. I would like to thank my statistic colleague Yumi Kondo, who always discussed statistics with me and made my days at the department interesting and cheerful. I would also like to thank Jun Chen for being a very nice roommate who beard my irregular working schedule.

Finally, I would like to express my deepest gratitude to my beloved parents. Without their love, I could never complete my graduate study.

To my family.

Chapter 1

Introduction

1.1 What is a Surrogate Endpoint?

In clinical trials, a clinical endpoint generally refers to occurrence of a disease, a symptom, a sign or a laboratory abnormality that constitutes one of the target outcomes of the trial. It directly measures how a patient feels, functions or survives and thus, is used to determine whether the treatment being studied is beneficial. A surrogate endpoint is an outcome which can be used as a substitute for a clinical endpoint. When assessing the treatment effect, a surrogate endpoint can be used to generate reliable conclusions instead of using the corresponding clinical endpoint directly. Examples of potential surrogate endpoints include CD4 cell count for HIV-related disease progression in clinical trials of anti-HIV treatments, progression-free survival time for survival time in clinical trials of treatments for advanced ovarian cancer and serum cholesterol levels for survival in clinical trials of treatments for cardiovascular disease. More examples of potential surrogate endpoints can be found in Burzykowsky et al. [4].

Why are surrogate endpoints required? The principal reason is that in many clinical trials, it is difficult to use the desired clinical endpoints directly. The clinical endpoint may be rare, so a large number of patients would be required for

a trial with adequate power (e.g. short-term mortality in patients with suspected acute myocardial infarction). The clinical endpoint may need a very long follow-up time to be detected (e.g. survival of patients in early-stage cancers), but too many patients might then be lost to follow-up. The clinical endpoint may also be difficult or costly to measure. In contrast, surrogates endpoints are outcomes that occur more often or are easier to measure. The motivation for the use of a surrogate endpoint is therefore the possibility of a reduction in the number of required patients or in the required trial duration.

In order to effectively substitute for a formal clinical endpoint, a surrogate endpoint must have the potential to yield unambiguous information about differential treatment effects on a clinical endpoint. The formal definition of a surrogate endpoint is given by Prentice [5] as "***a response variable for which a test of the null hypothesis of no relationship to the treatment groups under comparison is also a valid test of the corresponding null hypothesis based on the clinical endpoint***". The Prentice definition means that if a treatment has an effect on a clinical endpoint, then the treatment also has an effect on the surrogate endpoint, and the converse is also true. Mathematically, if S and C denote the surrogate endpoint and the clinical endpoint respectively, and Z denotes the treatment. then the Prentice definition can be written as:

$$f(S|Z) = f(S) \Leftrightarrow f(C|Z) = f(C), \tag{1.1}$$

where $f(X)$ denotes the probability distribution of the random variable X and $f(X|Z)$ denotes the probability distribution of X conditional on the value of Z.

1.2 Surrogate Endpoints in Multiple Sclerosis

Multiple sclerosis (MS) is a chronic and often disabling disease of the central nervous system. MS affects the ability of nerve cells in the brain and spinal cord

to communicate with each other. Nerve cells communicate by sending electrical signals called action potentials down long fibers called axons, which are wrapped in an insulating substance called myelin. In MS, the body's own immune system attacks and damages the myelin. When myelin is lost, the axons can no longer effectively conduct signals. The name multiple sclerosis refers to scars particularly in the white matter of the brain and spinal cord, which is mainly composed of myelin.

MS results in symptoms including difficulties in moving and coordination, deterioration of sensory functions, problems in bowel and bladder functions, among many others. MS onset usually occurs in young adults, and it is more common in women. Although much is known about the mechanisms involved in the disease process, the cause remains unknown, and there is no known cure for the disease to date.

There are several types of MS characterized by disease progression in terms of severity of disability. Relapsing-remitting MS (RRMS), the most common type, is characterized by unpredictable relapses followed by periods of months to years of relative quiet (remission) with no new signs of disease activity.

Until now, the only accepted primary endpoints for pivotal clinical trials of new treatments for RRMS are clinical outcomes, including relapse rate and accumulation of permanent disability, usually measured by the Extended Disability Status Scale (EDSS). There is no fully validated surrogate endpoint for RRMS yet. In RRMS clinical trials, magnetic resonance imaging (MRI) scans of the brain are often utilized to help monitor patients' health and the progression of their disease. McFarland et al. [6] argue that changes in MS brain lesion patterns determined by MRI scans, which reflect the underlying disease pathology, may be the best candidate for a surrogate endpoint in RRMS.

1.3 Outline of the Report

The objective of our study is to address the question: Are changes in brain lesion patterns determined by MRI a good surrogate endpoint for the relapse rate, the clinical endpoint in RRMS clinical trial?

This chapter has provided some background information about surrogate endpoints and MS. Chapter 2 provides a general review of how to validate a potential surrogate endpoint. We first discuss the importance of validation and then review different approaches, in situations where data is from a single clinical trial and data is from multiple clinical trials respectively. In the situation of multiple clinical trials, we focus on the scenario where only summary statistics for each trial are available. We review the methods adopted in Daniels and Hughes [2] and Korn et al. [3] in detail.

Chapter 3 considers validation in the RRMS setting. The specific potential surrogate endpoint we will focus on is the MRI lesion count, and the corresponding clinical endpoint is the annualized relapse rate. Information is presented on the dataset of Sormani et al. [1] (hereafter referred to as SBRCMB) used to assess the surrogacy relationship. The methodology of SBRCMB is discussed in detail, as well as the potential drawbacks of their approach.

In Chapter 4, we develop a related but different model to assess the surrogacy relationship. We focus on dealing with the issue of measurement error existing in estimating the surrogate endpoint and the clinical endpoint, and the context where data is available from several clinical trials, including some having more than two arms. We compare the results from the SBRCMB model and from our model. We also evaluate the prediction ability of the estimated surrogacy relationship to determine whether the surrogate endpoint is useful in practice. Chapter 5 summarizes the overall findings and discusses problems that remain to be investigated.

Chapter 2

Literature Review: Validation of Surrogate Endpoints

2.1 Importance of Validating a Potential Surrogate Endpoint

It is essential to validate a potential surrogate endpoint before using it as the primary outcome in a clinical trial. A surrogate endpoint should be able to assess the treatment effect in a clinical trial and the result obtained from the surrogate endpoint should be consistent with that obtained from the corresponding clinical endpoint. Inconsistent results will lead to an incorrect conclusion about the treatment effect, and thus misuse of the treatment in future, which may cause ineffective or even harmful impact on patients. For example, in some clinical trials regarding cardiologic disorder, blood pressure is used as a surrogate endpoint for actual survival of a patient. However, some treatments that are useful in lowering a patient's blood pressure have been shown to have no effect in reducing the risk of death from myocardial infarction. More examples of misuse of potential surrogate endpoints can be found in Fleming and DeMets [7].

Most potential surrogate endpoints are prognostic biomarkers, which means there is a strong association between the biomarker and the clinical endpoint at the level of the individual patient. Such association reflects a potential biological relationship between the biomarker and the clinical endpoint. However, as many studies have shown, a strong association is not enough. Surrogate endpoints are about assessing treatment effects. This means, at the trial level, the treatment effect obtained from a surrogate endpoint must reliably predict the treatment effect obtained from the clinical endpoint. Examples of the misuse of prognostic biomarkers as surrogate endpoints can be found in Fleming and DeMets [7].

Focusing on the Prentice definition of a surrogate endpoint (1.1), we require that if a treatment has an effect on a surrogate endpoint, then it also has an effect on the clinical endpoint. However, we also require that if a treatment doesn't have an effect on the surrogate endpoint, then it doesn't have an effect on the clinical endpoint either. Biologically, this implies the surrogate endpoint is on the sole causal pathway of the disease process to the clinical endpoint.

Figure 2.1 illustrates the perfect scenario for a surrogate as well as some imperfect scenarios: D, S and C stand for the disease, the surrogate endpoint and the clinical endpoint in a clinical trial respectively, while Z stands for the treatment applied in this clinical trial. Panel (a) shows the situation of a perfect surrogate endpoint, in which S is on the sole causal pathway from D to C. So the entire effect of Z on S will extend to C, and Z cannot affect C without affecting S. Panels (b), (c) and (d) show some situations of imperfect surrogate endpoints. Note that, in all these 3 situations, S is associated with C since they both are influenced by the same disease process D. However, in panel (b), S is not on the causal pathway from D to C. In the case illustrated, Z could affect S but not C, so S is not a surrogate endpoint for C. In panel (c), there are two pathways from D to C, and S is on one of them. If Z affects C only through X on the second pathway, then S is not a surrogate endpoint for C; if Z can affect C through both S and X, then

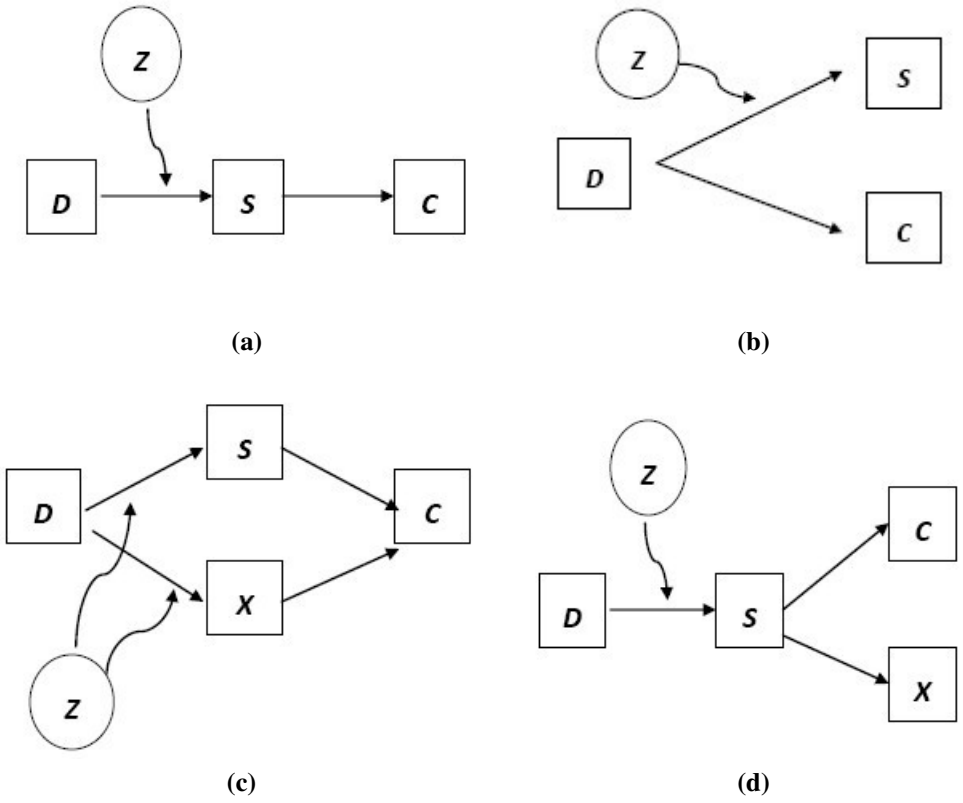

Figure 2.1: Scenarios of Perfect (a) and Imperfect (b,c,d) Surrogates

S is an imperfect surrogate endpoint for C. In such a case, an effect of Z on S could imply an effect of Z on C. However, since Z can bypass S and still influence C through X, it is possible that there is an effect on C but no effect on S. On the other hand, the effect of Z on S and the effect of Z on X may counteract each other, leading to no net effect of Z on C. In panel (d), it is possible that the effect of Z on S doesn't extend to C, but to X instead. In this case, if there is no treatment effect on S, then there is no treatment effect on C, but the converse is not always true.

2.2 Methods of Validating Surrogate Endpoints

2.2.1 The Prentice Operational Criteria for Validation

Prentice [5] proposed 4 operational criteria to validate a potential surrogate endpoint. Recalling his definition of a surrogate endpoint (1.1): $f(S|Z) = f(S) \Leftrightarrow f(C|Z) = f(C)$, and using the same notation, we can express the Prentice operational criteria as:

$$f(S|Z) \neq f(S) \tag{2.1}$$

$$f(C|Z) \neq f(C) \tag{2.2}$$

$$f(C|S) \neq f(C) \tag{2.3}$$

$$f(C|S,Z) = f(C|S) \tag{2.4}$$

Essentially, (2.1) requires that the treatment has an effect on the surrogate endpoint, (2.2) requires that the treatment has an effect on the clinical endpoint, (2.3) requires that different values of the surrogate endpoint result in different values of the clinical endpoint, which means the surrogate endpoint is a prognostic biomarker, and (2.4) requires that the surrogate endpoint should completely capture the dependence of the clinical endpoint on the treatment.

In practice, (2.1) and (2.2) are considered as necessary conditions for an outcome to be a surrogate endpoint, but not "actual" validation criteria. Note that (1.1) is equivalent to $f(S|Z) \neq f(S) \Leftrightarrow f(C|Z) \neq f(C)$, so (2.1) and (2.2) need to be satisfied or not simultaneously. Criteria (2.3) and (2.4) are the "actual" validation criteria. Usually, (2.3) is examined before (2.4), because a surrogate endpoint is expected to be a good prognostic biomarker. Criterion (2.4) is the essential part of the Prentice operational criteria. It means the treatment effect on the clinical endpoint can be entirely captured by the surrogate endpoint. A common way to examine (2.4) is to assume a regression model of form $C = \alpha + \beta Z + \gamma S + \varepsilon$ and

to check if the estimated regression coefficient for S is significantly different from 0 and that for Z is not. For this approach to be valid, one has to believe that the regression model describes the true relationship among C, S and Z.

Buyse and Molenberghs [8] show that (2.3) and (2.4) are necessary and sufficient conditions to establish (1.1) when the surrogate endpoint of interest is a binary outcome. When the surrogate endpoint is not binary, the criteria are only sufficient but not necessary; that is, if (2.3) and (2.4) are satisfied, then a treatment effect on the clinical endpoint ensures a treatment effect on the surrogate endpoint, but a treatment effect on the surrogate endpoint may not imply a treatment effect on the clinical endpoint. In terms of Figure 2.1, (2.3) and (2.4) exclude the situations (*b*) and (*c*), but not (*d*). (In (d), (2.3) holds because both S and C are influenced by D, and (2.4) holds because Z cannot affect C without affecting S.) Some counter examples are given in Buyse and Molenberghs [8] and Berger [9].

2.2.2 Validation in a Single Clinical Trial

To check the criterion (2.4), one needs to show that the statistical test for the treatment effect on the clinical endpoint to be nonsignificant after adjustment for the surrogate endpoint. However, this requirement raises a conceptual difficulty in validation since a nonsignificant result may simply be due to insufficient power of the statistical test. Hence, (2.4) is useful in rejecting a poor surrogate endpoint (the statistical test leads to a significant result), but is inadequate to validate a good surrogate endpoint. To overcome this difficulty, Freedman and Graubard [10] proposed a quantity called "proportion of the treatment effect explained by the surrogate" (PE) to measure the quality of a potential surrogate.

Let β and β_s be the parameters representing the treatment effect on the clinical endpoint C without and with adjustment for the surrogate endpoint S respectively.

Then PE is defined as:

$$PE = \frac{\beta - \beta_s}{\beta} = 1 - \frac{\beta_s}{\beta}. \tag{2.5}$$

It is expected that $\beta_s = 0$ when the surrogate is perfect; in this case, $PE = 1$. Naturally, PE being closer to 1 implies the surrogate endpoint explains more of the treatment effect on the clinical endpoint. In practice, β and β_s are replaced by their estimates, and the 2-sided 95% confidence interval for PE is constructed. Freedman and Graubard [10] suggested the lower limit of the interval should be greater than a critical value, say 0.5, for the surrogate endpoint to be considered useful.

For example, in a clinical trial, let α and β denote the treatment effect on the surrogate endpoint and the clinical endpoint respectively, and let S_j, C_j and Z_j denote the surrogate endpoint, the clinical endpoint and the treatment received for the jth patient. Here, Z_j is an indicator variable, which can be either 1 (the jth patient is in the active arm) or 0 (the jth patient is in the control arm). We often refer to the combination of an active arm and a control arm as a "contrast". So, α and β are the treatment effects obtained from the contrast in this clinical trial (i.e., by comparing the active arm and the control arm).

Assume the model:

$$\begin{aligned} S_j &= \mu_s + \alpha Z_j + \varepsilon_{sj}, \\ C_j &= \mu_c + \beta Z_j + \varepsilon_{cj}, \end{aligned} \tag{2.6}$$

where the error terms (ε_{si} and ε_{cj}) have a bivariate normal distribution with mean 0 and variance-covariance matrix

$$\Sigma = \begin{pmatrix} \sigma_{ss} & \sigma_{sc} \\ \sigma_{sc} & \sigma_{cc} \end{pmatrix}. \tag{2.7}$$

Then, one can obtain the conditional distribution of C_j given S_j, which is param-

eterized as:

$$C_j|S_j = \mu + \beta_s Z_j + \gamma S_j + \varepsilon_j, \tag{2.8}$$

where $\beta_s = \beta - \frac{\sigma_{sc}}{\sigma_{ss}}\alpha$. In this model, PE is given by:

$$PE = 1 - \frac{\beta_s}{\beta} = \frac{\sigma_{sc}}{\sigma_{ss}}\frac{\alpha}{\beta}. \tag{2.9}$$

Despite PE's description as the "proportion" of the treatment effect explained by the surrogate endpoint, it is not actually a "proportion". Molenberghs et al. [11] point out that the range of PE is not between 0 and 1 and discuss the interpretation problems of PE. For instance, the PE defined by (2.9) can take any value on the real line, because the range of $\frac{\alpha}{\beta}$ is unrestricted.

Buyse and Molenberghs [8] propose two quantities to replace PE in validating a potential surrogate endpoint. The first is the "adjusted association" ρ_A, a measure of the association between the surrogate endpoint and the clinical endpoint after adjustment for the treatment. In terms of model (2.6), ρ_A can be expressed as:

$$\rho_A = \frac{\sigma_{sc}}{\sqrt{\sigma_{ss}\sigma_{cc}}}. \tag{2.10}$$

The adjusted association ρ_A measures how good a surrogate endpoint performs at the level of the individual patient. In the above model, if $\rho_A = 1$, then the variance of ε_j in (2.8) is 0. So, C_j becomes a linear function of S_j, which means given the value of S_j, one can estimate the value of C_j without error. In this case, the surrogate endpoint and the clinical endpoint contain equivalent information about the treatment, hence one can determine the treatment effect on the clinical endpoint exactly from the treatment effect on the surrogate endpoint, and the Prentice definition (1.1) is satisfied [12].

The second quantity Buyse and Molenberghs [8] propose is the "relative effect" (RE), which is defined as the ratio of the treatment effect on the clinical endpoint to the treatment effect on the surrogate endpoint. In terms of model (2.6), RE is defined as:

$$RE = \frac{\beta}{\alpha}. \tag{2.11}$$

The relative effect RE is useful in predicting the treatment effect on the clinical endpoint from that on the surrogate endpoint. In practice, α and β are replaced by their estimates and a confidence interval for RE is constructed. A narrow confidence interval results in a good prediction of the treatment effect on the clinical endpoint. For example, based on the data from the current trial, one can obtain $\hat{RE} = \frac{\hat{\beta}}{\hat{\alpha}}$. For a future trial, one can estimate its treatment effect on the surrogate endpoint as $\hat{\alpha}_0$. Then, the treatment effect on the clinical endpoint from that future trial can be estimated as $\hat{\beta}_0 = \hat{\alpha}_0 \cdot \hat{RE}$. However, to make use of RE for such predictions, it is necessary to assume that the relationship (2.11) also holds in the future trial. This assumption may not be correct and cannot be checked in a single clinical trial.

2.2.3 Validation in Multiple Clinical Trials

When multiple clinical trials study the efficacy of the same treatment or treatments with a similar mechanism on the same disease, the validation procedure can use the information from these multiple trials. In this section, we review the methods used when the individual patient level data is available from each trial. In the next section, we discuss the methods used when only summary information from each trial is available.

Buyse et al. [13] consider the situation where individual patient level data is available and the surrogate endpoint and the clinical endpoint are both continu-

ously, normally distributed. Let S_{ij}, C_{ij} and Z_{ij} denote the surrogate endpoint, the clinical endpoint and the treatment received for the jth patient from the ith trial. Assume the model:

$$\begin{aligned} S_{ij} &= \mu_s + \mu_{si} + \alpha Z_{ij} + \alpha_i Z_{ij} + \varepsilon_{sij}, \\ C_{ij} &= \mu_c + \mu_{ci} + \beta Z_{ij} + \beta_i Z_{ij} + \varepsilon_{cij}, \end{aligned} \tag{2.12}$$

where μ_s and μ_c are fixed intercepts, α and β are the fixed effects of treatment on the surrogate endpoint and the clinical endpoint, μ_{si} and μ_{ci} are random intercepts and α_i and β_i are the random effects of treatment on the endpoints in trial i. The error terms ε_{sij} and ε_{cij} are assumed to follow the joint normal distribution with mean 0 and variance-covariance matrix given by (2.7), and the random effects $(\mu_{si}, \mu_{ci}, \alpha_i, \beta_i)^T$ are assumed to follow a joint normal distribution with mean 0 and variance-covariance matrix D given by:

$$D = \begin{pmatrix} d_{ss} & d_{sc} & d_{s\alpha} & d_{s\beta} \\ d_{sc} & d_{cc} & d_{c\alpha} & d_{c\beta} \\ d_{s\alpha} & d_{c\alpha} & d_{\alpha\alpha} & d_{\alpha\beta} \\ d_{s\beta} & d_{c\beta} & d_{\alpha\beta} & d_{\beta\beta} \end{pmatrix}. \tag{2.13}$$

Buyse et al. [13] suggest to evaluate the surrogate endpoint at two different levels. One is at the trial level, the other is at the individual patient level. At the trial level, the surrogacy relationship is assessed by the conditional variance of $\beta + \beta_i$ given μ_{si} and α_i. From (2.12) and (2.13), the conditional variance is given by:

$$Var(\beta + \beta_i | \mu_{si}, \alpha_i) = d_{\beta\beta} - \begin{pmatrix} d_{s\beta} \\ d_{\alpha\beta} \end{pmatrix}^T \begin{pmatrix} d_{ss} & d_{s\alpha} \\ d_{s\alpha} & d_{\alpha\alpha} \end{pmatrix}^{-1} \begin{pmatrix} d_{s\beta} \\ d_{\alpha\beta} \end{pmatrix}. \tag{2.14}$$

This conditional variance describes how precisely one can predict the treatment effect on the clinical outcome given the treatment effect on the surrogate outcome in a certain trial. Equivalently, a proportion type measure of "trial level" surrogacy is defined as:

$$R^2_{trial} = \frac{\begin{pmatrix} d_{s\beta} \\ d_{\alpha\beta} \end{pmatrix}^T \begin{pmatrix} d_{ss} & d_{s\alpha} \\ d_{s\alpha} & d_{\alpha\alpha} \end{pmatrix}^{-1} \begin{pmatrix} d_{s\beta} \\ d_{\alpha\beta} \end{pmatrix}}{d_{\beta\beta}}. \tag{2.15}$$

Moreover, one can quantify the relationship between the treatment effects on the surrogate endpoint and on the clinical endpoint by using the conditional expectation of $\beta + \beta_i$ given μ_{si} and α_i, which is:

$$E(\beta + \beta_i | \mu_{si}, \alpha_i) = \beta + \begin{pmatrix} d_{s\beta} \\ d_{\alpha\beta} \end{pmatrix}^T \begin{pmatrix} d_{ss} & d_{s\alpha} \\ d_{s\alpha} & d_{\alpha\alpha} \end{pmatrix}^{-1} \begin{pmatrix} \mu_{si} \\ \alpha_{si} \end{pmatrix}. \tag{2.16}$$

The equation (2.16) characterizes how the treatment effect on the clinical endpoint changes with the treatment effect on the surrogate endpoint. Given a new trial, after estimating the treatment effect on the surrogate endpoint, $\hat{\mu}_{si}$ and $\hat{\alpha}_{si}$, one can predict the expected treatment effect on the clinical endpoint through (2.16). Note that if we only have one trial, then we are not able to characterize this relationship.

At the individual patient level, the surrogacy relationship is evaluated using the adjusted association ρ_A used in the single trial situation. The conditional variance of C_{ij} given S_{ij} and the random effects is $\sigma_{cc} - \sigma^2_{cs}\sigma^{-1}_{ss}$. Thus, a proportion type measure of "individual level" surrogacy is defined as:

$$R^2_{ind} = \rho^2_A = \frac{\sigma^2_{cs}}{\sigma_{ss}\sigma_{cc}}. \tag{2.17}$$

A surrogate endpoint is considered to be perfect when both R^2_{trial} and R^2_{ind} are equal to 1. Large values of R^2_{trial} implies precise prediction of the treatment effect on the clinical endpoint, while large values of R^2_{ind} implies strong association be-

tween the surrogate endpoint and the clinical endpoint, which is useful in patient management. It is possible that R^2_{trial} is large and R^2_{ind} is small, or vice versa.

2.3 Validation in Multiple Clinical Trials with Individual Data Unavailable

In some contexts, only summary data of each trial, not the individual patient data, is available. For example, only results about the estimated treatment effect on the endpoints and the corresponding estimated standard errors may be available, not the outcomes of each patient. Then, the surrogacy relationship can only be evaluated at the trial level. Since we don't know the outcomes of each patient, we cannot evaluate the strength of the association between the surrogate endpoint and the clinical endpoint at the individual patient level (e.g., calculate R_{ind} in (2.17)). However, we are still able to assess the relationship between the treatment effect on the clinical endpoint and on the surrogate endpoint.

When only summary results from each trial are available, caution must be taken in the validation procedure because these summary results are only "estimates", which are different from the "true" quantities. For example, an estimated treatment effect on the endpoint from one trial is different from the true treatment effect on the endpoint from this trial. The true treatment effect is the effect obtained when the clinical trial includes an infinite number of patients. In practice, due to the limited number of patients, there always exist non-negligible estimation errors between the estimated and the true effects. How to appropriately model these estimation errors is important in assessing surrogacy relationships at the trial level. In the following subsections, we will review papers by Daniels and Hughes [2] (DH, hereafter) and Korn et al. [3] (KAM, hereafter), in which models are constructed to evaluate surrogacy relationships in multiple clinical trials for the situation when individual patient level data is unavailable.

2.3.1 Review of Daniels and Hughes [2]

Suppose N trials are used to analyze the performance of the surrogate endpoint of interest. In the ith trial, denote the true treatments effect on the surrogate endpoint and on the clinical endpoint as X_i^{true} and Y_i^{true} respectively. Correspondingly, let X_i and Y_i denote their estimates, i.e. the summary results obtained from the ith trial. Generally, unless the the number of patients in the ith trial is very large, X_i and Y_i are different from X_i^{true} and Y_i^{true}.

Given the ith trial, X_i is assumed to be normally distributed with mean X_i^{true} and variance δ_i^2 and Y_i is assumed to be normally distributed with mean Y_i^{true} and variance σ_i^2. Furthermore, the correlation between X_i and Y_i is assumed to be ρ_i. Here, δ_i^2 and σ_i^2 represent the effect of estimation error in the ith trial, and ρ_i represents the correlation between the estimation errors on X_i^{true} and Y_i^{true}. In mathematical form:

$$\begin{pmatrix} Y_i \\ X_i \end{pmatrix} \Bigg| \begin{pmatrix} Y_i^{true} \\ X_i^{true} \end{pmatrix} \sim N_2 \left(\begin{pmatrix} Y_i^{true} \\ X_i^{true} \end{pmatrix}, \begin{pmatrix} \sigma_i^2 & \rho_i\sigma_i\delta_i \\ \rho_i\sigma_i\delta_i & \delta_i^2 \end{pmatrix} \right). \tag{2.18}$$

The surrogacy relationship of interest is the relationship between the true treatment effects X_i and Y_i. DH assume the following structure:

$$Y_i^{true} | X_i^{true} \sim N(\alpha + \beta X_i^{true}, \tau^2). \tag{2.19}$$

Here, β measures the association between the true treatment effects on the clinical and the surrogate endpoint. If $\beta = 0$, then there is actually no such surrogacy relationship. When $\beta \neq 0$, a perfect surrogacy relationship also requires that $\alpha = 0$ so that the treatment having no effect on the surrogate endpoint suggests no effect on the clinical endpoint. Having $\alpha \neq 0$ implies that there is a treatment effect on the clinical endpoint unexplained by the surrogate endpoint. The variance τ^2

represents the uncertainty of using X_i^{true} to predict Y_i^{true}. If $\tau^2 = 0$, then Y_i^{true} will be perfectly determined when X_i^{true} is given.

At this stage, DH assume the X_i^{true}s are fixed quantities. The reason why they choose X_i^{true}s as fixed rather than random is to avoid having to propose specific distributions for the X_i^{true}s, which they think may not be appropriate. (Though later, they put very flat prior distributions on X_i^{true}s when estimating the model parameters in the Bayesian framework.) Then combining (2.18) and (2.19), we obtain the bivariate normal model for Y_i and X_i:

$$\begin{pmatrix} Y_i \\ X_i \end{pmatrix} \sim N_2 \left(\begin{pmatrix} \alpha + \beta X_i^{true} \\ X_i^{true} \end{pmatrix}, \begin{pmatrix} \sigma_i^2 + \tau^2 & \rho_i \sigma_i \delta_i \\ \rho_i \sigma_i \delta_i & \delta_i^2 \end{pmatrix} \right). \tag{2.20}$$

In some clinical trials, there may be more than one active arm, in addition to the control arm. A common situation is that different patients receive different levels of dosage of a treatment. For example, if a treatment is applied at 2 dosage levels, then this clinical trial consists of 3 arms. Patients on the first arm receive treatment with dosage level one, patients on the second arm receive treatment with dosage level two, and patients on the third arm receive control. Since the combination of any active arm and a control arm yields a contrast, this clinical trial consists of 2 contrasts.

From a clinical trial with multiple contrasts, we obtain multiple estimated treatment effects on both endpoints. Suppose there are 3 arms in the ith trial, which can generate 2 contrasts. Let Y_{i1} and X_{i1} be the estimated treatment effects on the clinical and surrogate endpoints from the first contrast, and Y_{i2} and X_{i2} be those from the second contrast. Correspondingly, let $Y_{i1}^{true}, X_{i1}^{true}, Y_{i2}^{true}$ and X_{i2}^{true}

be the true treatment effects. Then model (2.18) can be generalized to:

$$\begin{pmatrix} Y_{i1} \\ X_{i1} \\ Y_{i2} \\ X_{i2} \end{pmatrix} \Bigg| \begin{pmatrix} Y_{i1}^{true} \\ X_{i1}^{true} \\ Y_{i2}^{true} \\ X_{i2}^{true} \end{pmatrix} \sim N4 \left(\begin{pmatrix} Y_{i1}^{true} \\ X_{i1}^{true} \\ Y_{i2}^{true} \\ X_{i2}^{true} \end{pmatrix}, \begin{pmatrix} \sigma_{i1}^2 & \rho_{i11}\sigma_{i1}\delta_{i1} & \rho_{iy}\sigma_{i1}\sigma_{i2} & \rho_{i12}\sigma_{i1}\delta_{i2} \\ \rho_{i11}\sigma_{i1}\delta_{i1} & \delta_{i1}^2 & \rho_{i21}\delta_{i1}\sigma_{i2} & \rho_{ix}\delta_{i1}\delta_{i2} \\ \rho_{iy}\sigma_{i1}\sigma_{i2} & \rho_{i21}\delta_{i1}\sigma_{i2} & \sigma_{i2}^2 & \rho_{i22}\sigma_{i2}\delta_{i2} \\ \rho_{i12}\sigma_{i1}\delta_{i2} & \rho_{ix}\delta_{i1}\delta_{i2} & \rho_{i22}\sigma_{i2}\delta_{i2} & \delta_{i2}^2 \end{pmatrix} \right). \tag{2.21}$$

The off-diagonal blocks of covariance terms in (2.21) are allowed to be non-zero, reflecting the possibility of correlations among the two pairs of estimated treatment effects arising because they all involve comparisons to the same control arm. Also, assuming X_{i1}^{true} and X_{i2}^{true} are fixed, (2.19) is generalized (M. J. Daniels, personal communication) as:

$$\begin{pmatrix} Y_{i1}^{true} \\ Y_{i2}^{true} \end{pmatrix} \Bigg| \begin{pmatrix} X_{i1}^{true} \\ X_{i2}^{true} \end{pmatrix} \sim N_2 \left(\begin{pmatrix} \alpha + \beta X_{i1}^{true} \\ \alpha + \beta X_{i2}^{true} \end{pmatrix}, \begin{pmatrix} \tau^2 & 0 \\ 0 & \tau^2 \end{pmatrix} \right). \tag{2.22}$$

From (2.22), we can see that the marginal distributions of Y_{i1}^{true} and Y_{i2}^{true} have the same form. This is because all the treatments included in the analysis have similar mechanism of action; whether two contrasts are from one trial or different trials, they should reflect the same surrogacy relationship. DH assume the covariance between Y_{i1}^{true} and Y_{i2}^{true} is 0 in (2.22). In principle, this covariance could be non-zero and (2.22) can be replaced by substituting the 0 by an non-zero parameter.

Combining (2.21) and (2.22), we get:

$$\begin{pmatrix} Y_{i1} \\ X_{i1} \\ Y_{i2} \\ X_{i2} \end{pmatrix} \sim N_4 \left(\begin{pmatrix} \alpha + \beta X_{i1}^{true} \\ X_{i1}^{true} \\ \alpha + \beta X_{i2}^{true} \\ X_{i2}^{true} \end{pmatrix}, \begin{pmatrix} \sigma_{i1}^2 + \tau^2 & \rho_{i11}\sigma_{i1}\delta_{i1} & \rho_{iy}\sigma_{i1}\sigma_{i2} + \tau^2 & \rho_{i12}\sigma_{i1}\delta_{i2} \\ \rho_{i11}\sigma_{i1}\delta_{i1} & \delta_{i1}^2 & \rho_{i21}\delta_{i1}\sigma_{i2} & \rho_{ix}\delta_{i1}\delta_{i2} \\ \rho_{iy}\sigma_{i1}\sigma_{i2} + \tau^2 & \rho_{i21}\delta_{i1}\sigma_{i2} & \sigma_{i2}^2 + \tau^2 & \rho_{i22}\sigma_{i2}\delta_{i2} \\ \rho_{i12}\sigma_{i1}\delta_{i2} & \rho_{ix}\delta_{i1}\delta_{i2} & \rho_{i22}\sigma_{i2}\delta_{i2} & \delta_{i2}^2 \end{pmatrix} \right). \tag{2.23}$$

For a clinical trial with 3 or more contrasts, a similar extension can be applied.

DH assume a joint normal structure for the summary results (estimated treat-

ment effects) from each trial included in the study. To estimate the model parameters, they adopt a Bayesian approach. In the estimation procedure, all the within trial variances and correlations are assumed known and replaced by their estimates. The variance estimates for each trial are obtained from the summary results of that trial. For the correlation estimates, if the individual patient level data for one trial is available, the correlation estimates in this trial are calculated from the individual patient data. Otherwise, the correlation estimates for that trial are set to the average value of the correlation estimates from trials which individual patient level data are available. In the Bayesian procedure, priors are then placed on α, β, τ^2 and all the true treatment effects on the surrogate endpoint (i.e. X_i^{true} in single contrast trials and X_{ij}^{ture}s in multiple contrast trials).

To assess the surrogacy relationship, DH propose to examine if the 95% credible intervals for α, β and τ^2 exclude 0. Also, DH suggest to compute Bayes factors [14] to test if α, β and τ^2 are 0. If the tests reject the null hypothesis of $\beta = 0$ and don't reject the null hypotheses of $\alpha = 0$ and $\tau^2 = 0$, then the surrogacy relationship is considered to be validated.

2.3.2 Review of Korn et al. [3]

KAM discuss different models to assess the surrogacy relationship for two different types of clinical trials. One type of trial involves unordered treatment arms (i.e. there is no control arm in the trial), and the other type of trial involves ordered treatment arms (i.e. there is one control arm in the trial). Since the dataset we will use in the next chapter consists of only ordered trials, and also to make KAM's model comparable with DH's model, we only discuss their model for ordered trials.

In contrast to DH, KAM start their model at the arm level instead of at the contrast level. In the ith clinical trial, let C_{ij} and S_{ij} be the observed clinical endpoint

and the observed surrogate endpoint from the jth arm, where $j = 0, 1, 2, \ldots$ ($j = 0$ represents the control arm in the trial). Similarly, let C_{ij}^{true} and S_{ij}^{true} be the true clinical and surrogate endpoints. KAM's model begins by describing the estimation errors in estimating the endpoints. Correspondingly, let e_{ij} and f_{ij} denote the estimation errors in the surrogate endpoint and the clinical endpoint respectively. Then:

$$\begin{cases} S_{ij} = S_{ij}^{true} + e_{ij} \\ C_{ij} = C_{ij}^{true} + f_{ij}, \end{cases} , \tag{2.24}$$

Since the estimation errors happen in different arms with different patients, they are assumed to be independent. KAM further assume that $e_{ij} \overset{iid}{\sim} N(0, \sigma_{ij}^2)$ and $f_{ij} \overset{iid}{\sim} N(0, \delta_{ij}^2)$, and that e_{ij} and f_{ij} are independent.

As a next step, KAM model S_{ij}^{true} and C_{ij}^{true}. Let μ_i represent the expected level of the true surrogate endpoint on the control arm, and μ_S represent the expected difference on the true surrogate endpoint between the active and control arms. Let m_{ij} be the random effect representing the uncertainty in the true surrogate endpoint for each arm. KAM express S_{i0}^{true} and S_{ij}^{true} as:

$$S_{i0}^{true} = \mu_i + m_{i0} \quad \text{and} \quad S_{ij}^{true} = \mu_i + \mu_S + m_{ij}, \quad \text{for } j \neq 0. \tag{2.25}$$

KAM assume $m_{i0} \sim N(0, \lambda_0^2)$, $m_{ij} \sim N(0, \lambda^2)$ ($j \neq 0$), and all m_{ij}s are independent. Note that although m_{ij}s ($j \neq 0$) are the random effects for different active arms, they are assumed to have the same distribution. Similarly, all m_{i0}s are assumed to have the same distribution. Furthermore, m_{ij} is assumed to be independent of e_{ij} and f_{ij}, which means the estimation errors are not affected by the value of the true endpoints.

KAM assume there is a linear relationship between C_{ij}^{true} and S_{ij}^{true}, specifically:

$$C_{i0}^{true} = \alpha_i + \beta S_{i0}^{true} + g_{i0} \quad \text{and} \quad C_{ij}^{true} = \alpha_i + \mu_C + \beta S_{ij}^{true} + g_{ij}, \quad \text{for } j \neq 0, \tag{2.26}$$

where β represents the linear relationship between C_{ij}^{true} and S_{ij}^{true}, and α_i and $\alpha_i + \mu_C$ are the intercepts in the control arms and the active arms respectively. Here, μ_C represents the expected difference on the clinical endpoint between the active and control arms that cannot be explained by the influence of the true surrogate endpoint on the true clinical endpoint. The random effects g_{ij} account for the fact that C_{ij}^{true} and S_{ij}^{true} are not perfectly linearly related and are assumed to be independent and normally distributed with mean 0 and variance $\tau^2/2$. Note that all the g_{ij}s are assumed to have the same distributions though they are from different arms. Since g_{ij}s are not estimation errors, g_{ij}, e_{ij} and f_{ij} are assumed to be independent.

The treatment effect is estimated as the difference between the endpoints from the active arm and from the control arm. Let $X_{ij} = S_{ij} - S_{i0}$ and $Y_{ij} = C_{ij} - C_{i0}$ denote the estimated treatment effects on the surrogate and on the clinical endpoints respectively ($j \neq 0$). Corresponding, let $X_{ij}^{true} = S_{ij}^{true} - S_{i0}^{true}$ and $Y_{ij}^{true} = C_{ij}^{true} - C_{i0}^{true}$ denote the true treatment effects. From (2.24), (2.25) and (2.26), we have:

$$\begin{cases} X_{ij} = X_{ij}^{true} + (e_{ij} - e_{i0}) \\ Y_{ij} = Y_{ij}^{true} + (f_{ij} - f_{i0}) \end{cases} \quad \text{where} \quad \begin{cases} X_{ij}^{true} = \mu_S + (m_{ij} - m_{i0}) \\ Y_{ij}^{true} = \mu_C + \beta X_{ij}^{true} + (g_{ij} - g_{i0}) \end{cases} \tag{2.27}$$

From (2.27), we obtain:

$$\begin{aligned} &E(Y_{ij}^{true} | X_{ij}^{true}) = \mu_C + \beta X_{ij}^{true} \\ &Var(Y_{ij}^{true} | X_{ij}^{true}) = \tau^2, \end{aligned} \tag{2.28}$$

which describes the surrogacy relationship between the true treatment effects on the clinical endpoint and on the surrogate endpoint. We now see the interpreta-

tions of μ_C, β and τ^2 in the KAM model are the same as the interpretations of α, β and τ^2 in the DH model. As before, β measures the association between the true treatment effect on the clinical endpoint and on the surrogate endpoint.

For a trial with single contrast, from (2.27) we can obtain the joint distribution of the estimated treatment effects:

$$\begin{pmatrix} Y_{i1} \\ X_{i1} \end{pmatrix} \sim N_2 \left(\begin{pmatrix} \mu_C + \beta\mu_S \\ \mu_S \end{pmatrix}, \begin{pmatrix} \beta^2(\lambda_0^2+\lambda^2) + \tau^2 + (\sigma_{i1}^2 + \sigma_{i0}^2) & \beta(\lambda_0^2+\lambda^2) \\ \beta(\lambda_0^2+\lambda^2) & (\lambda_0^2+\lambda^2) + (\delta_{i1}^2 + \delta_{i0}^2) \end{pmatrix} \right). \tag{2.29}$$

For a trial with 2 contrasts, from (2.27), after similar calculation:

$$(Y_{i1}, X_{i1}, Y_{i2}, X_{i2})^T \sim N_4 \left(\begin{pmatrix} \mu \\ \mu \end{pmatrix}, \begin{pmatrix} \Sigma_{1i} & \Sigma_{3i} \\ \Sigma_{3i} & \Sigma_{2i} \end{pmatrix} \right), \tag{2.30}$$

where

$$\mu = \begin{pmatrix} \mu_C + \beta\mu_S \\ \mu_S \end{pmatrix}$$

and

$$\Sigma_{1i} = \begin{pmatrix} \beta^2(\lambda_0^2+\lambda^2) + \tau^2 + (\sigma_{i1}^2 + \sigma_{i0}^2) & \beta(\lambda_0^2+\lambda^2) \\ \beta(\lambda_0^2+\lambda^2) & (\lambda_0^2+\lambda^2) + (\delta_{i1}^2 + \delta_{i0}^2) \end{pmatrix},$$

$$\Sigma_{2i} = \begin{pmatrix} \beta^2(\lambda_0^2+\lambda^2) + \tau^2 + (\sigma_{i2}^2 + \sigma_{i0}^2) & \beta(\lambda_0^2+\lambda^2) \\ \beta(\lambda_0^2+\lambda^2) & (\lambda_0^2+\lambda^2) + (\delta_{i2}^2 + \delta_{i0}^2) \end{pmatrix},$$

$$\Sigma_{3i} = \begin{pmatrix} \beta^2\lambda_0^2 + \frac{\tau^2}{2} + \sigma_{i0}^2 & \beta\lambda_0^2 \\ \beta\lambda_0^2 & \lambda_0^2 + \sigma_{i0}^2 \end{pmatrix}.$$

For a trial with 3 or more contrasts, a similar extension can be applied.

When fitting their model, KAM use the maximum likelihood estimators obtained from the joint normal distributions (2.29) and (2.30). The estimation error

terms σ_{ij}^2 and δ_{ij}^2 are assumed known and replaced by their estimates when fitting the model.

To assess the surrogacy relationship, in addition to evaluating the estimates and the confidence intervals for μ_C, β and τ^2, KAM suggest that one can use a R^2-type measure. From (2.27) and (2.29), we know $Var(Y_{ij}^{true}) = \beta^2(\lambda_0^2 + \lambda^2) + \tau^2$, and $Var(Y_{ij}^{true}|X_{ij}^{true}) = \tau^2$. So, the R^2-type measure is defined as:

$$R_{trial}^2 = \frac{\beta^2(\lambda_0^2 + \lambda^2)}{\beta^2(\lambda_0^2 + \lambda^2) + \tau^2}. \tag{2.31}$$

This quantity is analogous to R_{trial}^2 in (2.15). Large values of R_{trial}^2 indicate a good surrogacy relationship.

Furthermore, to evaluate how a surrogate endpoint performs in practice, KAM suggest to estimate the parameter $E(Y_{ij}^{true}|X_{ij})$, which is useful in predicting the true treatment effect on the clinical endpoint given the estimated treatment effect on the surrogate endpoint. This parameter is analogous to $E(\beta + \beta_i|\mu_{si}, \alpha_i)$ in (2.16). However, KAM suggest to condition Y_{ij}^{true} on X_{ij}, rather than on X_{ij}^{true}. From (2.27), the parameter of interest is:

$$\Delta = E(Y_{ij}^{true}|X_{ij}) = (\mu_C + \beta\mu_S) + \Big(\frac{\beta(\lambda_0^2 + \lambda^2)}{(\lambda_0^2 + \lambda^2) + (\delta_{ij}^2 + \delta_{i0}^2)}\Big)(X_{ij} - \mu_S). \tag{2.32}$$

To estimate Δ, KAM plug in the estimates for $(\beta, \mu_S, \mu_C, \lambda_0^2, \lambda^2)$ and the observed value of X_{ij} from a new trial and replace δ_{ij}^2 and δ_{i0}^2 by their estimates from that trial.

2.3.3 Comparison of These Two Approaches

The first difference between DH and KAM is that their models start from different levels: DH start directly from the treatment effects (contrast level, since treatment effects are obtained from contrasts), where they build the model for (Y_i, X_i) given (Y_i^{true}, X_i^{true}) and for Y_i^{true} given X_i^{true}. In contrast, KAM start from the endpoints (arm level, since the endpoint values are obtained from the arms), where they first specify the joint distribution for $(S_{ij}, C_{ij}, S_{ij}^{true}, C_{ij}^{true})$, and take the difference to obtain the joint distribution for Y_{ij} and X_{ij}. Building the model from the arm level requires a more detailed specification. However, in (2.26), KAM assume the same coefficient β for control arms and active arms, which implies the relationships between the true surrogate endpoint and the true clinical endpoint are the same regardless of the arm. This assumption may not be very realistic in some situations, where a treatment may substantially influence the association between two endpoints and thus it may be more reasonable to assume different βs for control and active arms. In contrast, DH don't make assumptions about the relationship between the endpoints but model the surrogacy relationship directly in (2.19). We think the DH approach is more reasonable from this perspective.

Both papers deal with the estimation errors in the same way in the sense that the estimation errors are assumed to be independent of the true treatment effects. In (2.18), DH assume σ_i^2 and δ_i^2, the within trial estimation errors, do not depend on Y_i^{true} and X_i^{true}. This means the estimation errors on the treatment effects are not affected by the true treatment effects. Similarly, in (2.25), KAM assume m_{ij} are independent of e_{ij} and f_{ij}, which means the estimation errors on the endpoints are not affected by the true endpoints. This assumption implies that $(m_{ij} - m_{i0})$ are independent from $(e_{ij} - e_{i0})$ and $(f_{ij} - f_{i0})$, which also means the estimation errors on the treatment effects are not affected by the true treatment effects. However, it is possible that a large true treatment effect is associated with a large estimation error, while a small treatment effect is associated with a small estimation error. Thus, this independence assumption may not hold in some clinical trials.

To compare how these models differ in characterizing the treatment effects, we can compare the joint distributions for the treatment effects. For example, we can compare (2.20) with (2.29). Alternatively, from (2.27), we obtain:

$$\begin{pmatrix} Y_{i1} \\ X_{i1} \end{pmatrix} \Bigg|_{(Y_{i1}^{true}, X_{i1}^{true})} \sim N_2 \left(\begin{pmatrix} Y_{i1}^{true} \\ X_{i1}^{true} \end{pmatrix}, \begin{pmatrix} \sigma_{i1}^2 + \sigma_{i0}^2 & 0 \\ 0 & \delta_{i1}^2 + \delta_{i0}^2 \end{pmatrix} \right), \tag{2.33}$$

and

$$Y_{i1}^{true} | X_{i1}^{true} \sim N(\mu_C + \beta X_{i1}^{true}, \tau^2). \tag{2.34}$$

Comparing (2.33) and (2.34) with (2.18) and (2.19), it is evident that the DH model and the KAM model are essentially the same. One difference is that X_{i1}^{true} follows a normal distribution with mean μ_S and variance $\lambda_0^2 + \lambda^2$ in the KAM model, while DH treat X_i^{true} as fixed when specifying their model but then give it a prior distribution when carrying out the estimation. The prior is chosen to be normal with mean 0 and a very large variance, meaning it is "non-informative". Besides this, the conditional covariance in (2.33) is 0, while the conditional covariance in (2.18) is allowed to be non-zero. This is because KAM assume the within trial estimation errors e_{ij} and f_{ij} are independent, but DH allow a correlation ρ_i. It is likely that the two estimation errors are correlated in general. However, without individual patient level data, it is difficult to estimate this correlation.

In the following chapters, we will discuss validation of the surrogate endpoint in the MS context. Our dataset consists of multiple clinical trials and only summary results from these trials are available. We will discuss two approaches to validate the surrogate endpoint of interest: the SBRCMB approach and a more comprehensive approach. The comprehensive approach is similar in spirit to the DH and KAM models.

Chapter 3

Lesion Counts as a Surrogate Endpoint in RRMS: the SBRCMB Approach

3.1 Introduction and the SBRCMB Dataset

Recently, MRI measures of brain lesion counts on RRMS patients are widely used in clinical trials as a potential surrogate endpoint. One important clinical endpoint in RRMS clinical trials is the annualized relapse rate. A relapse is defined as appearance of new symptom or worsening of an existing symptom, attributable to MS, accompanied by an appropriate new neurologic abnormality. However, the surrogacy relationship between such MRI measures and this clinical endpoint has remained incompletely validated. Petkau et al. [15] show that the correlation between MRI lesion counts and the annualized relapse rate at the individual level is weak. The low degree of correlation at the individual level indicates that MRI measures would be unreliable predictors of the annualized relapse rate for an individual patient. However, this result does not exclude the possibility that the treatment effects on MRI measures and on the annualized relapse rate are highly associated, which means that MRI measures may still be useful for assessing the

treatment effect at the trial level.

To evaluate whether MRI measures are useful in assessing treatment effects, SBRCMB collected summary information from multiple MS clinical trials. The SBRCMB dataset includes 23 randomized, double-blind, placebo-controlled trials. The treatments in the trials are believed to have similar mechanism of action. There are 2 trials including both secondary progressive multiple sclerosis patients and RRMS patients. The remaining 19 trials include only RRMS patients. Among the 23 trials, there are 9 trials of 2 arms, 14 trials of 3 arms, 1 trial of 4 arms and 1 trial of 5 arms. Each trial has only 1 control arm but 1 to 4 active arms. In total, there are 63 arms, 40 contrasts and 6591 patients. The detailed SBRCMB dataset is included in Appendix *A*.

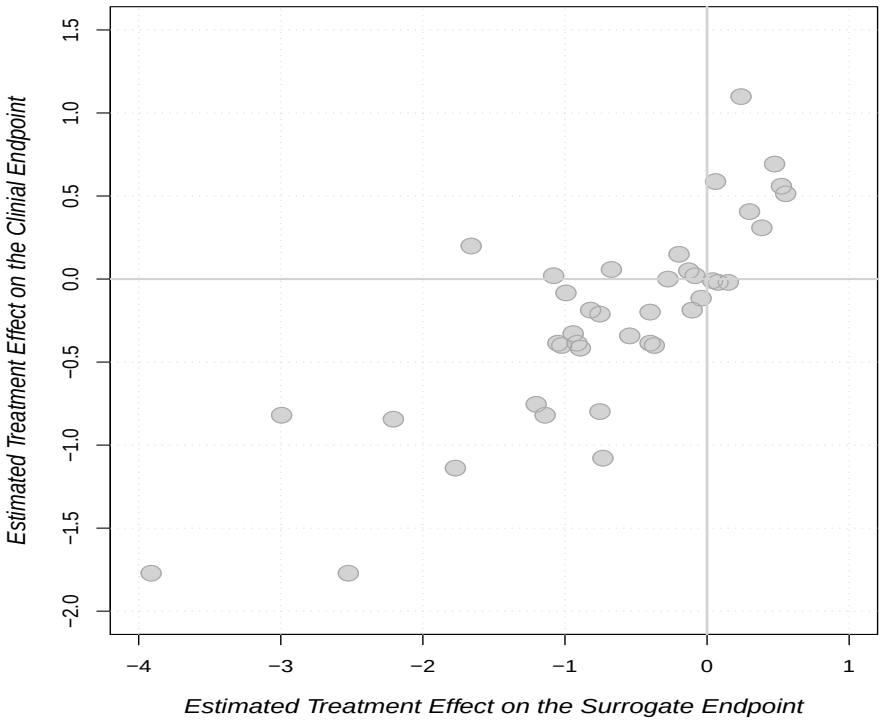

Figure 3.1: Scatter Plot of Estimated Treatment Effects

The SBRCMB dataset contains no individual patient level data, only the summary results from each clinical trial. The observed clinical endpoint for an arm is defined as the observed annualized relapse rate for this arm (it is assumed that all the patients in the same trial have the same follow-up time) and the observed surrogate endpoint for an arm is defined as the observed MRI lesion count per patient per scan from this arm (all the patients in the same trial are assumed to have the same scan times). The estimated treatment effect on the clinical endpoint is then defined as the log ratio between the observed clinical endpoints in the active and control arms. Similarly, the estimated treatment effect on the surrogate endpoint is then defined as the log ratio between the observed surrogate endpoints in the active and control arms. Since one contrast is formed by comparing one active arm and one control arm, we can obtain one estimated treatment effect on the clinical endpoint and one estimated treatment effect on the surrogate endpoint from each contrast. In total, we have 40 pairs of estimated treatment effects. Figure 3.1 shows the scatter plot of these pairs of estimated treatment effects. Note that, the observed endpoints are not equal to the true endpoints (unless the arm includes infinite number of patients), and thus the estimated treatment effects are not equal to the true treatment effects. The task is to assess the surrogacy relationship between the true treatment effects, which are not observable, based on the estimated treatment effects.

3.2 The SBRCMB approach

SBRCMB adopt a simple linear regression model and use weighted least squares (WLS) to assess the surrogacy relationship. The explanatory variable is the estimated treatment effect on the surrogate endpoint and the response variable is the estimated treatment effect on the clinical endpoint. In order to account for the influence of differences in trial size and trial duration for the contrasts, different weights are given to different contrasts. Specifically, let w_i denote the weight

given to the ith contrast, where $i = 1, 2, 3..., 40$. Then:

$$w_i = N_{\text{complete}_i} \cdot \sqrt{\frac{\text{follow-up (months)}_i}{12}}, \tag{3.1}$$

where follow-up (months)$_i$ is the duration of the MRI follow-up in months of the patients in the ith contrast, and N_{complete_i} is a number which SBRCMB choose to represent the total number of patients in this contrast. For a contrast from a trial with only 2 arms, N_{complete_i} is equal to the total number of patients in these two arms. For a contrast from a trial with more than 2 arms, N_{complete_i} is obtained by equally dividing the number of placebo patients between the treatment arms. For example, for a trial with 20 patients on each of the 3 arms, 2 contrasts are created with $N_{\text{complete}_i} = 20 + \frac{20}{2} = 30$ for both contrasts.

Let Y_i and X_i represent the estimated treatment effect on the clinical endpoint and surrogate endpoint from the ith contrast. SBRCMB assume the following regression model to describe the surrogacy relationship:

$$E(Y_i) = \alpha + \beta X_i, \tag{3.2}$$

and estimate the regression coefficients based on WLS; that is:

$$\min \sum w_i (Y_i - \alpha - \beta X_i)^2. \tag{3.3}$$

SBRCMB also carry out a sensitivity study, an interaction study and a validation study. The sensitivity study aims to check whether the regression coefficients are sensitive to the choice of the weights, or to the choice of the contrasts included in the analysis. To check the sensitivity with respect to the choice of the weights, SBRCMB refit the regression line with 2 other weights w_i' and w_i'', where w_i' gives more weight to the duration of the contrast:

$$w_i' = N_{\text{complete}_i} \cdot \frac{\text{follow-up (months)}_i}{12}, \tag{3.4}$$

and w_i'' is a constant weight (i.e. $w_i'' \equiv 1$). To check the sensitivity with respect to the choice of the contrasts, SBRCMB divide the whole dataset into different subsets with different features, and fits regression lines based on those subsets separately, all using the weights in (3.1). The first subset is a "highest contrasts" subset, which includes only data from "the active arm with the highest dose level versus control arm" contrast. The second subset is a "RRMS contrasts" subset, which includes data only from trials with only RRMS patients. The third subset is a "large effect contrasts" subset, which includes only data from the contrasts with estimated treatment effect on the clinical endpoint greater than 20%. Table 3.1 shows the results we reproduced for the sensitivity study; these are almost the same as those from SBRCMB.

Table 3.1: Results of the Sensitivity Study

Analysis	No. of trials	No. of contrasts	$\hat{\alpha}^*$	$\hat{\beta}^*$	R^2
w_i	23	40	-0.02 (0.05)	0.55 (0.04)	0.80
w_i'	23	40	-0.02 (0.05)	0.58 (0.04)	0.84
$w_i'' \equiv 1$	23	40	0.12 (0.07)	0.50 (0.06)	0.65
highest	23	23	-0.06 (0.08)	0.53 (0.06)	0.77
RRMS	21	36	-0.03 (0.05)	0.56 (0.05)	0.80
large effect	18	25	-0.01 (0.10)	0.58 (0.07)	0.75

* *estimate (estimated standard error)*

The values in the R^2 column are the weighted coefficients of determination:

$$R^2 = \frac{\sum w_i(\hat{y}_i - \bar{y})^2}{\sum w_i(y_i - \bar{y})^2}, \tag{3.5}$$

where $\hat{y}_i$ is the fitted value and w_i can be replaced by w_i' when (3.4) is used.

In the sensitivity study, none of the $\hat{\alpha}$s are significantly different from 0 but all the $\hat{\beta}$s are. Furthermore, SBRCMB claim that all the estimates of βs are close (all between 0.50 and 0.58) and all the R^2s are close (between 0.65 and 0.84). They interpret these findings as indicating that the fitted regression line is not sensitive

to the choice of weights or to the choice of contrasts involved.

The SBRCMB interaction study aims to check whether the regression coefficients depend on the characteristics of the trials. For example, Let I_i be an indicator variable, which takes the value 1 if the ith contrast is from a trial conducted after year 2000 and 0 otherwise. Then SBRCMB fit the following regression model with weight w_i:

$$E(Y_i) = \alpha + \beta_1 X_i + \beta_2 I_i + \beta_3 I_i \cdot X_i. \tag{3.6}$$

Through assessing β_2 and β_3, one can see whether there is a difference in the regression coefficients between the contrasts before year 2000 and after year 2000.

In addition to this "time period" factor, SBRCMB also examine the factors "drug class" (whether a contrast is from a trial whose treatment is an interferon) and "annualized relapse rate" (whether the observed annualized relapse rate in the placebo arm of a contrast is larger than 1). The reproduced results of the interaction study are shown in Table 3.2. The "P-value" column shows the P-values of testing if the coefficient of the interaction term is 0 (e.g. test if $\beta_3 = 0$ in (3.6)).

Table 3.2: Results of the Interaction Study

indicator variable	class	No. of contrasts	P-value
time period	> 2000	15	0.30
	< 2000	25	
drug class	with interferon	12	0.20
	not interferon	28	
annualized relapse rate	> 1	9	0.36
	< 1	31	

In the interaction study, as all these P-values are greater than 0.05, SBRCMB claim that there is no indication of differences in the slope of the fitted line for

contrasts with different characteristics, though SBRCMB also note that the power of this test is quite low due to the limited sample size.

Finally, SBRCMB carry out a validation study, where 4 new clinical trials are introduced, which result in 4 new contrasts (each of these trials has only 2 arms). Their estimated treatment effects on the clinical endpoint are compared with the predict counterpart obtained from the regression model with weight w_i. The reproduced results of the validation study are shown in Figure 3.2, where the hollow points represent the 40 actual contrasts used in the regression model, the solid line is the estimated regression line with weight w_i, the solid points represent the 4 new contrasts, and the bars are the 95% prediction intervals for the estimated treatment effects on the clinical endpoint for the 4 new contrasts. The prediction intervals are calculated by the standard regression approach: the X_is are assumed to be fixed, and the weights w_is are assumed to be proportional to the inverse of the variance of the Y_is.

It can be seen that all the solid points lie within the prediction intervals (except for the 2nd one from the left, which is at the very edge of the prediction interval). SBRCMB claim that the estimated regression model is able to give satisfactory predictions. However, these 4 new trials use active control arms rather than placebo-controlled arms. So, these 4 trials have different designs from the 23 trials in SBRCMB's dataset, and may not tell us whether the estimated regression equation can produce satisfactory predictions.

Based on all of these results, SBRCMB conclude that in RRMS, the treatment effect on MRI lesion count can be used to predict the treatment effect on the annualized relapse rate. They state that these results support for the use of MRI lesion count as a surrogate endpoint in RRMS clinical trials with treatments of analogous mechanism.

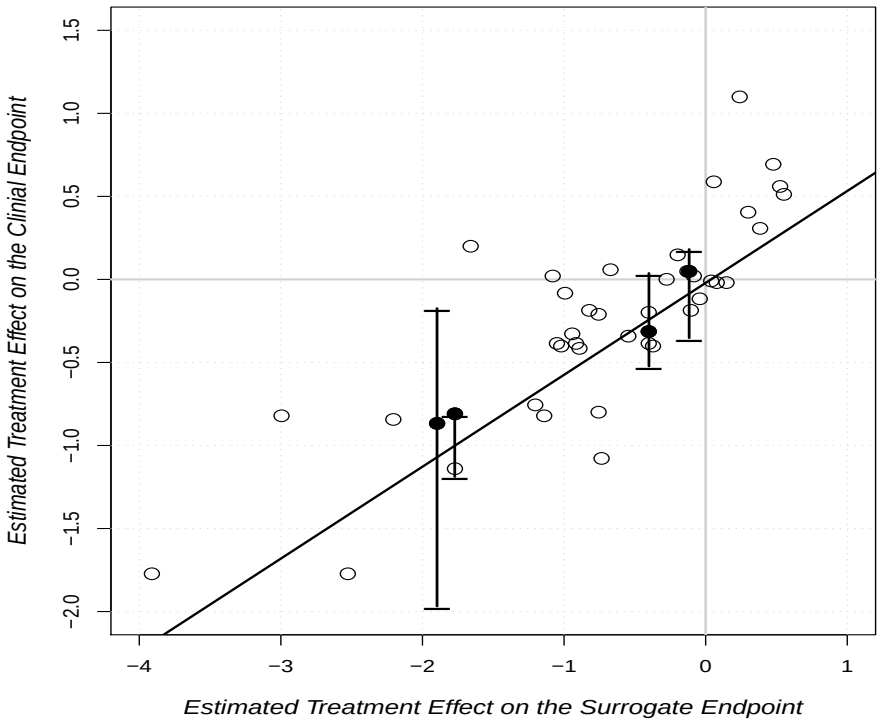

Figure 3.2: Results of the Validation Study

3.3 Critique of the SBRCMB Approach

In this section, we discuss shortcomings of the SBRCMB approach in assessing the surrogacy relationship. The fundamental issue is the WLS estimates may not be appropriate for the dataset. There are several reasons.

First, the explanatory variable X_i used in the SBRCMB model is defined as the log ratio between the **observed** MRI lesion counts per patient per scan in the active and the control arms, and the response variable Y_i used is defined as the log ratio between the **observed** annualized relapse rates in the active and the control arms. Since the observed endpoints are not equal to the true endpoints, X_i and Y_i are just estimates of the true treatment effects. If X_i^{true} and Y_i^{true} denote the corresponding true treatment effects, then the surrogacy relationship is the relationship between

X_i^{true} and Y_i^{true}, not that between X_i and Y_i. The SBRCMB approach doesn't take into account the influence of estimation errors in both X_i and Y_i, which may lead to a biased result.

Second, 14 of the 23 trials have more than 2 arms, which leads to correlated contrasts since the contrasts from the same trial share the same control arm. Therefore, even if we believe the estimation errors are negligible so that the relationship between Y_i and X_i should be an excellent approximation to the relationship between Y_i^{true} and X_i^{true}, the WLS approach is still not appropriate because some of the Y_is are correlated.

Third, the SBRCMB choices for the weights used in the WLS estimation are quite mysterious. SBRCMB simply state that the weights are chosen because they reflect the information conveyed by each trial. Suppose that there is no estimation error, and all the Y_is are independent so that it is reasonable to use the WLS approach. Then are these weights appropriate?

In the following subsections, we discuss each of these potential problems. We start with the appropriateness of the weights under the assumption that the WLS approach is reasonable. Then we discuss the correlation issue. Finally, we discuss the more fundamental issue of the influence of estimation errors in estimating the surrogacy relationship.

3.3.1 The Appropriateness of the Weights

In this section, we focus on the relationship between Y_i and X_i, and assume that all the Y_is are independent. Furthermore, we assume that all the X_is are fixed.

We assume the following regression model:

$$Y_i = \alpha + \beta x_i + \varepsilon_i, \tag{3.7}$$

where $E(\varepsilon_i) = 0$, $Var(\varepsilon_i) = \tau_i^2$ and all the ε_is are independent. Then theoretically, the weight w_i for Y_i should be proportional to the inverse of the variance of ε_i, i.e. $w_i \propto \tau_i^{-2}$. We use x_i instead of X_i here because x_is are assumed to be fixed.

In the following text, we omit the subscript i on every quantity to simplify notation. Let R_a and R_c be the observed annualized relapse rate in the active and the control arms respectively from a certain trial. Let R_a^{true} and R_c^{true} be the corresponding true annualized relapse rates. Then $Y = \log \frac{R_a}{R_c}$ and $Y^{true} = \log \frac{R_a^{true}}{R_c^{true}}$. Note that since R_a and R_c are from different arms with different patients, it is natural to consider them to be independent. Similarly, we consider R_a^{true} and R_c^{true} also to be independent.

Suppose that there are N_a and N_c patients in the active and control arm respectively, and assume that all $N_a + N_c$ patients have the same number of years of follow-up for the relapse data, namely T. Then, let F_j denote the total number of relapses of the jth patient in the active arm. We assume:

$$E(F_j|R_a^{true}) = TR_a^{true}, \quad Var(F_j|R_a^{true}) = \phi \cdot TR_a^{true}, \tag{3.8}$$

where ϕ is a dispersion parameter, describing how the variance of the number of relapses is related to its expectation. If $\phi = 1$, this corresponds to a Poisson assumption. We assume that ϕ is the same for all the patients in all the trials. Thus, ϕ has neither subscript j nor subscript i.

Then, $\frac{F_j}{T}$ is this patient's annualized relapse rate. From the above assumption, we have:

$$E(\frac{F_j}{T}|R_a^{true}) = R_a^{true}, \quad Var(\frac{F_j}{T}|R_a^{true}) = \phi \cdot \frac{R_a^{true}}{T}. \tag{3.9}$$

By definition, the observed annualized relapse rate in the active arm is:

$$R_a = \frac{F_1 + F_2 + ... + F_{N_a}}{TN_a}. \tag{3.10}$$

Then, by the delta method and the Central Limit Theorem, we obtain the following approximation to the conditional distribution of $\log R_a$:

$$\log R_a | R_a^{true} \approx N(\log R_a^{true}, \frac{\phi}{TN_a R_a^{true}}). \tag{3.11}$$

Similarly, for the control arm, we have:

$$\log R_c | R_c^{true} \approx N(\log R_c^{true}, \frac{\phi}{TN_c R_c^{true}}). \tag{3.12}$$

Unconditionally, we have:

$$\begin{aligned} Var(\log R_a) &= Var(E(\log R_a | R_a^{true})) + E(Var(\log R_a | R_a^{true})) \\ &\approx Var(\log R_a^{true}) + \tfrac{\phi}{TN_a} E(\tfrac{1}{R_a^{true}}), \end{aligned} \tag{3.13}$$

and similarly, for the control arm, we have:

$$\begin{aligned} Var(\log R_c) &= Var(E(\log R_c | R_c^{true})) + E(Var(\log R_c | R_c^{true})) \\ &\approx Var(\log R_c^{true}) + \tfrac{\phi}{TN_c} E(\tfrac{1}{R_c^{true}}). \end{aligned} \tag{3.14}$$

The independence assumption for R_a and R_c leads to:

$$\begin{aligned} Var(Y) = Var(\log \tfrac{R_a}{R_c}) = Var(\log R_a) + Var(\log R_c) \\ \approx Var(\log R_a^{true}) + Var(\log R_c^{true}) + \tfrac{\phi}{T}(\tfrac{1}{N_a} E(\tfrac{1}{R_a^{true}}) + \tfrac{1}{N_c} E(\tfrac{1}{R_c^{true}})). \end{aligned} \tag{3.15}$$

From the above formula, we can see that the variance of Y depends on the distribution of R_a^{true} and R_c^{true} as well as on the unknown parameter ϕ.

If we include the subscript i, (3.15) is actually $Var(Y_i) = Var(\log R_{ai}^{true}) + Var(\log R_{ci}^{true}) + \frac{\phi}{T_i}(\frac{1}{N_{ai}}E(\frac{1}{R_{ai}^{true}}) + \frac{1}{N_{ci}}E(\frac{1}{R_{ci}^{true}}))$, for $i = 1, 2, ..., 40$. Now we assume all the R_{ai}^{true}s are identically distributed. We think all the treatments included in the SBRCMB dataset have similar mechanism of action, so the distribution of the R_{ai}^{true} describes how the true clinical endpoint varies across contrasts. Similarly, we assume all the R_{ci}^{ture}s are identically distributed. As a result, the variances and the expectations in (3.15) are constant across trials.

One way to estimate $E(\frac{1}{R_a^{true}})$ and $E(\frac{1}{R_c^{true}})$ is to average all the R_as and all the R_cs across the contrasts and take their inverse. For the SBRCMB dataset, we obtain $\hat{E}(\frac{1}{R_a^{true}}) \approx 1.43$ and $\hat{E}(\frac{1}{R_c^{true}}) \approx 1.10$.

Let θ denote $Var(\log R_a^{true}) + Var(\log R_c^{true})$. Then the variance of Y can be written as:

$$\tau^2 = Var(Y) = \theta + \phi(\frac{1.43}{TN_a} + \frac{1.10}{TN_c}). \tag{3.16}$$

The values of T, N_a and N_c all depend on the contrast leading to Y. If we let $c = \frac{1.47}{TN_a} + \frac{1.12}{TN_c}$ and include the subscript i, we have:

$$\tau_i^2 = \theta + \phi c_i. \tag{3.17}$$

Based on (3.17), we can examine the appropriateness of the weights used in the SBRCMB approach. If $w_i = N_{\text{complete}_i} \cdot \sqrt{\frac{\text{follow-up (months)}_i}{12}}$ is appropriate, then w_i should be proportional to the inverse of the variance of the estimated clinical outcome; that is:

$$w_i = \frac{a}{\tau_i^2} = \frac{a}{\theta + \phi c_i} \Rightarrow \frac{1}{w_i} = \frac{\theta}{a} + \frac{\phi}{a} c_i, \tag{3.18}$$

where a is an arbitrary proportionality constant. The above result implies that if we draw the scatter plot of $(c_i, \frac{1}{w_i})$, the points should gather around a straight line.

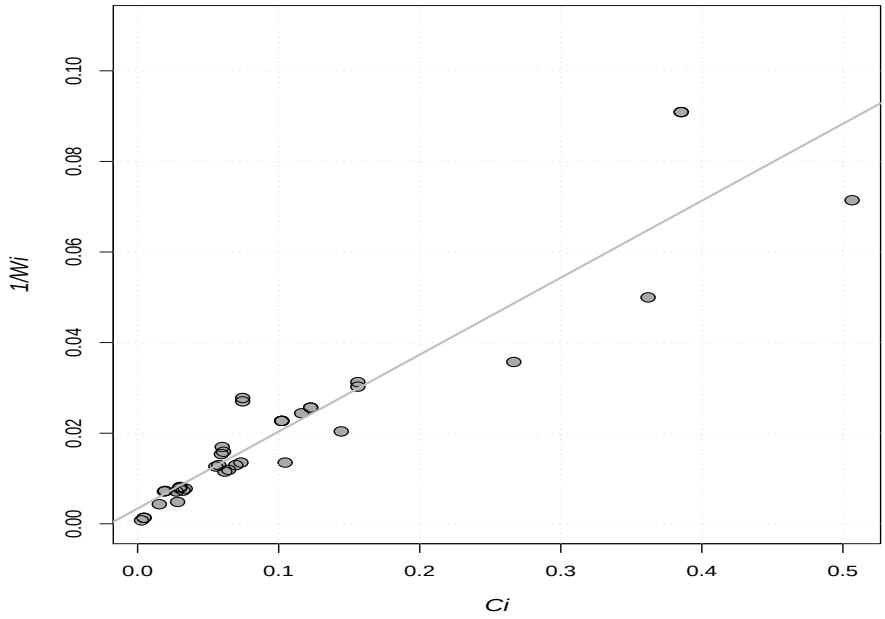

Figure 3.3: Scatter Plot of $(c, 1/w)$

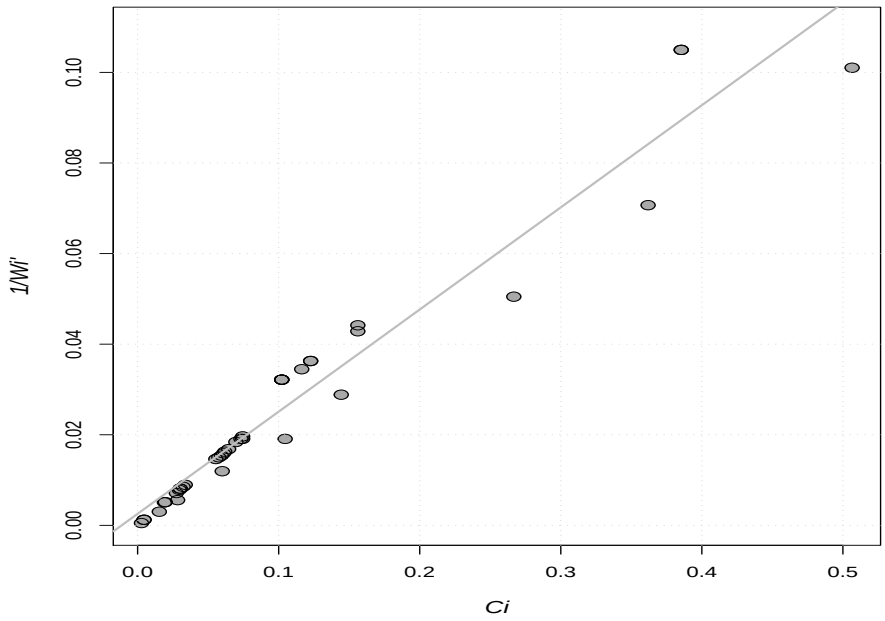

Figure 3.4: Scatter Plot of $(c, 1/w')$

Figure (3.3) compares w_is to c_is and Figure (3.4) compares w_i's to c_is. In both scatter plots, the points approximately gather around a straight line. This suggests, if the assumptions we made in this section are reasonable, then both weights used by SBRCMB also seem reasonable. From these two plots, we would expect the w_is and w_i's to perform similarly.

3.3.2 Correlation of the Contrasts

The WLS approach is appropriate when the response variables are independent. However, this is not the case for the SBRCMB data. As mentioned before, 14 of the 23 trials have more than 2 arms. So, if two contrasts are from the same trial, then the estimated treatment effect on the clinical endpoint from these two contrasts are correlated, because the two contrasts share the same control arm.

For example, let Y_1 and Y_2 be two estimated treatment effects on the clinical endpoint from the same three-arm trial. Then, $Y_1 = \log \frac{R_{a1}}{R_c}$ and $Y_2 = \log \frac{R_{a2}}{R_c}$, where R_{a1}, R_{a2} and R_c are the observed annualized relapse rates in the first active arm, the second active arm and the control arm respectively. Because R_{a1} and R_{a2} are from different arms with different patients, we assume they are independent. Then:

$$Cov(Y_1, Y_2) = Cov(\log \frac{R_{a1}}{R_c}, \log \frac{R_{a2}}{R_c}) = Var(\log R_c). \tag{3.19}$$

Now, it is clear that Y_1 and Y_2 are correlated. An immediate way to address this correlation in the regression model is to use generalized least squares. However, from the last section, we know that:

$$Var(\log R_c) \approx Var(\log R_c^{true}) + \frac{\phi}{TN_c} E(\frac{1}{R_c^{true}}) \approx Var(\log R_c^{true}) + \frac{1.10\phi}{TN_c}. \tag{3.20}$$

To make use of generalized least squares, we need to estimate the covariance between any two correlated Y_i and Y_j. But the estimate of that covariance requires

an estimate of $Var(\log R_c^{true})$, the variance of the logarithm of the true annualized relapse rate across all the trials, and the unknown parameter ϕ. These two quantities are difficult to estimate without assuming a more complicated model. We will address this question in the next chapter by developing a more comprehensive model.

3.3.3 Influence of Estimation Errors

As mentioned at the beginning of this chapter, the relationship of real interest is between Y_i^{true} and X_i^{true}. However, we cannot observe Y_i^{true} and X_i^{true} directly, but can only use Y_i and X_i to estimate them. Suppose the true surrogacy relationship is:

$$E(Y_i^{true}|X_i^{true}) = \alpha + \beta X_i^{true}. \tag{3.21}$$

Then, the question is: when we use Y_i and X_i in place of Y_i^{true} and X_i^{true} to estimate α and β as was done by SBRCMB, how good are these estimators?

In this section, we consider X_i^{true} as random rather than fixed. We think it is a reasonable assumption for the SBRCMB dataset. Since all the patients included in the study received treatments that are considered to be of the same type, it is then natural to think of all the true treatment effects from the different trials as coming from a single probability distribution. To simplify the discussion, we assume Y_i^{true} and X_i^{true} are bivariate normally distributed. The conditional expectation of Y_i^{true} given X_i^{true} is given in (3.21), and the conditional variance of Y_i^{true} given X_i^{true} is denoted as τ^2. Also, let μ_X and σ_X^2 represent the expectation and the variance of X_i^{true}. Then, the bivariate normal distribution of Y_i^{true} and X_i^{true} is:

$$\begin{pmatrix} Y_i^{true} \\ X_i^{true} \end{pmatrix} \sim N_2 \left(\begin{pmatrix} \alpha + \beta\mu_X \\ \mu_X \end{pmatrix}, \begin{pmatrix} \beta^2\sigma_X^2 + \tau^2 & \beta\sigma_X^2 \\ \beta\sigma_X^2 & \sigma_X^2 \end{pmatrix} \right). \tag{3.22}$$

If we could observe Y_i^{true} and X_i^{true}, then the OLS estimators based on Y_i^{true} and X_i^{true} for β and α are unbiased and consistent. Because:

$$\hat{\beta} = \frac{\sum (X_i^{true} - \bar{X}^{true})(Y_i^{true} - \bar{Y}^{true})}{\sum (X_i^{true} - \bar{X}^{true})^2}, \tag{3.23}$$

where $\bar{X}_i^{true}$ and $\bar{Y}_i^{true}$ are the averages of all X_i^{true}s and Y_i^{true}s included in the study, then:

$$E(\hat{\beta}) = E(E(\hat{\beta}|X^{true})) = E(\beta) = \beta, \tag{3.24}$$

where X^{true} represents the collection of all X_i^{true}s. For consistency, note that $\sum (X_i^{true} - \bar{X}^{true})^2/n \xrightarrow{p} Var(X_i^{true}) = \sigma_X^2$ and $\sum (X_i^{true} - \bar{X}^{true})(Y_i^{true} - \bar{Y}^{true})/n \xrightarrow{p} Cov(Y_i^{true}, X_i^{true}) = \beta\sigma_X^2$, where n is the number of contrasts. So, $\hat{\beta} \xrightarrow{p} \frac{\beta\sigma_X^2}{\sigma_X^2} = \beta$.

A similar argument can be made for $\hat{\alpha}$. Note that:

$$\hat{\alpha} = \bar{Y}^{true} - \hat{\beta}\bar{X}^{true}. \tag{3.25}$$

Then it is clear that $E(\hat{\alpha}) = E(E(\hat{\alpha}|X^{true})) = E(\alpha + \beta\bar{X}^{true} - \beta\bar{X}^{true}) = E(\alpha) = \alpha$, and $\hat{\alpha} \xrightarrow{p} (\alpha + \beta\mu_X) - \beta\mu_X = \alpha$.

However, if we can only observe Y_i and X_i, then the OLS estimator for β becomes:

$$\tilde{\beta} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}, \tag{3.26}$$

where $\bar{X}$ and $\bar{Y}$ are the average of all X_is and Y_is included in the study. Consequently $\tilde{\alpha} = \bar{Y} - \tilde{\beta}\bar{X}$. Are these estimators still unbiased and consistent?

Consider the following simple model. Let e_i and f_i represent the estimation errors on X_i^{true} and Y_i^{true} respectively:

$$X_i = X_i^{true} + e_i \quad \text{and} \quad Y_i = Y_i^{true} + f_i. \tag{3.27}$$

We assume $e_i \overset{iid}{\sim} N(0, \delta^2)$ and $f_i \overset{iid}{\sim} N(0, \sigma^2)$. Furthermore, we assume that e_i and f_i are independent and are independent of X_i^{true} and Y_i^{true} for all i. As a result, we obtain the joint distribution for the estimated treatment effects:

$$\begin{pmatrix} Y_i \\ X_i \end{pmatrix} \sim N_2 \left(\begin{pmatrix} \alpha + \beta\mu_X \\ \mu_X \end{pmatrix}, \begin{pmatrix} \beta\sigma_X^2 + \tau^2 + \sigma^2 & \beta\sigma_X^2 \\ \beta\sigma_X^2 & \sigma_X^2 + \delta^2 \end{pmatrix} \right). \tag{3.28}$$

It follows that:

$$\begin{aligned} E(Y_i|X_i) &= \alpha + \beta\mu_X + \frac{\beta\sigma_X^2}{\sigma_X^2 + \delta^2}(X_i - \mu_X) \\ &= \left(\alpha + \beta\mu_X - \frac{\beta\sigma_X^2}{\sigma_X^2 + \delta^2}\mu_X \right) + \left(\beta \cdot \frac{\sigma_X^2}{\sigma_X^2 + \delta^2} \right) X_i. \end{aligned} \tag{3.29}$$

Analogous to (3.23) and (3.24), we now have $E(\tilde{\beta}) = \beta \cdot \frac{\sigma_X^2}{\sigma_X^2+\delta^2}$, which means $\tilde{\beta}$ is not an unbiased estimator of β. For consistency, it is also clear that $\tilde{\beta} = \frac{S_{xy}}{S_{xx}} \xrightarrow{p} \beta\frac{\sigma_X^2}{\sigma_X^2+\delta^2}$. So, $\tilde{\beta}$ is also not a consistent estimator of β. Similar conclusions hold for $\tilde{\alpha}$.

Note that the coefficient $\frac{\sigma_X^2}{\sigma_X^2+\delta^2}$ is always less than 1 unless $\delta^2 = 0$. Hence, under this model, when there exist estimation errors in the regressor, the expectation of the OLS estimator is always smaller than its true value. This is called the attenuation effect in regression. As demonstrated, this effect does not disappear even when the sample size goes to infinity. So, when the estimation error is not negligible (i.e. δ^2 is not very small relative to σ_X^2), the OLS estimator is not a good estimator. On the other hand, we see the estimation errors in the response variable don't affect the unbiasedness and consistency property of the OLS estimator.

For more complex situations such as when the estimation errors are not identically distributed, or the X_i^{true} is fixed rather than random, it can be shown that the OLS estimator is still biased and inconsistent. The WLS estimator can also

be shown to be biased and inconsistent when there exist estimation errors in the regressor, no matter what kind of weights are applied to the data. For the SBRCMB dataset, since some trials included only a modest number of patients, non-negligible estimation errors must exist in the estimated treatment effects from those trials. Therefore, the OLS (WLS) estimator will tend to underestimate the true regression coefficient.

Furthermore, using simple linear regression may lead to incorrect assessment of the surrogacy relationship. For example, in the above model, if no estimation errors exist, then the coefficient of determination R^2 is the square of the sample correlation coefficient between Y_i^{true} and X_i^{true}. From (3.22), we have:

$$R^2 = \frac{[\sum (X_i^{true} - \bar{X}^{true})(Y_i^{true} - \bar{Y}^{true})]^2}{\sum (X_i^{true} - \bar{X}^{true})^2 \sum (Y_i^{true} - \bar{Y}^{true})^2} \xrightarrow{p} \frac{\beta^2 \sigma_X^4}{\sigma_X^2(\beta^2 \sigma_X^2 + \tau^2)}. \tag{3.30}$$

However, if estimation errors exist, and (3.28) is assumed, the coefficient of determination becomes

$$\tilde{R}^2 = \frac{[\sum (X_i - \bar{X})(Y_i - \bar{Y})]^2}{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2} \xrightarrow{p} \frac{\beta^2 \sigma_X^4}{(\sigma_X^2 + \delta^2)(\beta^2 \sigma_X^2 + \tau^2 + \sigma^2)}. \tag{3.31}$$

When estimation errors exist, σ^2 and δ^2 are always larger than 0, so the coefficient of determination tends to underestimate the square of the correlation coefficient between Y_i^{true} and X_i^{true}, which may lead to a false conclusion about the surrogacy relationship. The coefficient of determination is 65% from SBRCMB with $w_i'' \equiv 1$. However, the correlation between the true treatment effects on the clinical endpoint and on the surrogate endpoint may be higher, which means a better surrogacy relationship.

In the next chapter, we will re-analyze the surrogacy relationship with a more comprehensive approach to take into account the existence of estimation errors and the correlated contrasts in the SBRCMB dataset.

Chapter 4

Lesion counts as a Surrogate Endpoint in RRMS: A More Comprehensive Approach

In this chapter, we use the SBRCMB dataset to re-analyze the surrogacy relationship between the MRI lesion count and the annualized relapse rate at the trial level. We start with modeling the true treatment effects (the surrogacy relationship) in the single-contrast clinical trials and develop the conditional distribution of the observed endpoints given the true endpoints to account for the estimation errors. Similar models are then generalized to the multiple-contrast trials to address the issue of the correlated contrasts. Once all components of the model are constructed, the model parameters are estimated based on "normal estimating equations". The results are then compared with those obtained from the SBRCMB approach and the estimated surrogacy relationship is evaluated as well as its usefulness in practice.

In each arm, we define the true clinical endpoint as the true annualized relapse rate, which is the expected value of the observed annualized relapse rate. In fact, every patient in the same arm has his/her own observed annualized relapse rate,

and we assume they all have the same probabilistic distribution whose expectation is the true annualized relapse rate (as defined in Section 3.3.1). Similarly, we define the true surrogate endpoint as the true MRI lesion count per scan per patient, which is the expected value of the observed MRI lesion count per scan per patient. So, corresponding to the estimated treatment effects defined through the observed endpoints, we define the true treatment effects on the endpoints as the log ratio between the true endpoints in the active arm and in the control arm. We aim to assess the relationship between these true treatment effects.

4.1 Model for the Single-contrast Clinical Trials

4.1.1 Model for the True Treatment Effects

In the SBRCMB dataset, there are 9 single-contrast trials. For each of these 9 trials, let R_a^{true} and R_c^{true} denote the true annualized relapse rates in the active arm and in the control arm, and let M_a^{true} and M_c^{true} denote the true MRI lesion counts per scan per patient in the active arm and in the control arm. Then the true treatment effect on the clinical endpoint is defined as $Y^{true} = \log \frac{R_a^{true}}{R_c^{true}}$ and the true treatment effect on the surrogate endpoint is defined as $X^{true} = \log \frac{M_a^{true}}{M_c^{true}}$. We assume the following bivariate normal model for these two true treatment effects:

$$\begin{pmatrix} Y^{true} \\ X^{true} \end{pmatrix} \sim N_2 \left(\begin{pmatrix} \mu_Y \\ \mu_X \end{pmatrix}, \begin{pmatrix} \sigma_Y^2 & \sigma_{YX} \\ \sigma_{YX} & \sigma_X^2 \end{pmatrix} \right). \tag{4.1}$$

Since different trials consist of different patients, we assume that the true treatment effects are independent across trials. The model (4.1) is assumed to be true for all the contrasts from all the single-contrast trials. This is reasonable because all the trials in the dataset are included to examine the effects of treatments with similar mechanisms of action and therefore we hope to see a similar relationship between the true treatment effect on the the clinical endpoint and on the surrogate

endpoint across all the trials with this type of treatment. We omit the subscript i for the ith trial in our notation throughout the following development.

The distribution in (4.1) is specified in an unstructured form. To express the surrogacy relationship, we represent the moments of the conditional distribution of Y^{true} on X^{true} as:

$$E(Y^{true} \mid X^{true}) = \alpha + \beta X^{true} \text{ and } Var(Y^{true} \mid X^{true}) = \tau^2. \tag{4.2}$$

The parameter β is our primary interest, as it measures the strength of the surrogacy relationship. If β is 0, then the MRI lesion count is not a surrogate for the annualized relapse rate for this type of treatment at the trial level, since knowledge of the true treatment effect on the MRI lesion count doesn't help to predict the true treatment effect on the annualized relapse rate. The parameter α is also of interest and we expect it to be small. If α is not 0, there is a part of the true treatment effect on the annualized relapse rate that is unexplained by the true treatment effect on the MRI lesion count per patient per scan. The parameter τ^2 represents the precision of this linear relationship; that is, how precisely we can predict the true treatment effect on the annualized relapse rate given the true treatment effect on the MRI lesion count.

The Prentice definition (1.1) describes a prefect surrogate relationship: no treatment effect on the surrogate endpoint implies no treatment effect on the clinical endpoint and vice versa. In our context, (1.1) requires both α and τ^2 to be 0, while β must not be 0; that is, the relationship between Y^{true} and X^{true} is deterministic and multiplicative: $Y^{true} = \beta X^{true}$. However, such a perfect surrogacy relationship will seldom be realized in practice.

Using the parametrization specified in (4.2), we can rewrite (4.1) as:

$$\begin{pmatrix} Y^{true} \\ X^{true} \end{pmatrix} \sim N_2 \left(\begin{pmatrix} \alpha + \beta\mu_X \\ \mu_X \end{pmatrix}, \begin{pmatrix} \beta^2\sigma_X^2 + \tau^2 & \beta\sigma_X^2 \\ \beta\sigma_X^2 & \sigma_X^2 \end{pmatrix} \right). \tag{4.3}$$

4.1.2 Model for the Observed Annualized Relapse Rate and MRI Lesion Count Per Patient Per Scan

Let R_a, R_c and M_a, M_c denote the observed annualized relapse rates and the observed MRI lesion counts per patient per scan on the active and control arms. To derive the probability distribution of R_a and R_c, we use the same assumptions used in Section 3.3.1 and follow the notation used there (except we use ϕ_1 now instead of ϕ). As a result, we have:

$$\log R_a | R_a^{true} \approx N(\log R_a^{true}, \frac{\phi_1}{TN_a R_a^{true}}), \tag{4.4}$$

$$\log R_c | R_c^{true} \approx N(\log R_c^{true}, \frac{\phi_1}{TN_c R_c^{true}}). \tag{4.5}$$

Similarly, for the observed MRI lesion count, let G_j denote the cumulative number of MRI lesions of the jth patient from the active arm on the K scans obtained for this patient during the follow-up time T. (As in SBRCMB, we assume the follow-up time for the MRI data is the same as the follow-up time for the relapse data, all the patients in a trial have the same follow-up time T, and all the patients in a trial have the same number of scans K.) We then assume:

$$E(G_j | M_a^{true}) = KM_a^{true}, \quad Var(G_j | M_a^{true}) = \phi_2 \cdot KM_a^{true}, \tag{4.6}$$

where ϕ_2 is a dispersion parameter describing how the variance of the MRI lesion count is related to its expectation. As for ϕ_1, we assume that ϕ_2 is the same for all the patients in all the trials. Thus, ϕ_2 has neither subscript j nor subscript i. Then:

$$E(\frac{G_j}{K}|M_a^{true}) = M_a^{true}, \quad Var(\frac{G_j}{K}|M_a^{true}) = \phi_2 \cdot \frac{M_a^{true}}{K}. \tag{4.7}$$

By definition, the observed MRI lesion count per patient per scan is:

$$M_a = \frac{G_1 + G_2 + ... + G_{N_a}}{KN_a}. \tag{4.8}$$

Then, by the delta method and the Central Limit Theorem, we obtain the following approximation to the conditional distribution of $\log M_a$:

$$\log M_a|M_a^{true} \approx N(\log M_a^{true}, \frac{\phi_2}{KN_aM_a^{true}}). \tag{4.9}$$

Similarly, for the control arm, we have:

$$\log M_c|M_c^{true} \approx N(\log M_c^{true}, \frac{\phi_2}{KN_cM_c^{true}}). \tag{4.10}$$

4.1.3 Model for the Estimated Treatment Effects

From (4.4), it is clear that R_a and R_a^{true} are not independent, which is reasonable since the observed clinical endpoint should depend on the true clinical endpoint. Now, we assume that given R_a^{true}, the conditional distribution of $\log R_a$ is independent of R_c^{true}, M_a^{true} and M_c^{true}; that is , if we already know R_a^{true}, the additional information of R_c^{true}, M_a^{true} and M_c^{true} does not help to predict $\log R_a$.

It is natural to think that R_c^{true} and M_c^{true} affect neither R_a nor R_a^{true}. The patients in the active arm and in the control arm are distinct, and the patients in the

active arm received the treatment while the patients in the control arm did not, so it seems obvious that the behavior of the patients in the control arm should not affect the behavior of the patients in the active arm. For M_a^{true}, we could think that if it affects R_a, that effect would be only through R_a^{true}. Therefore, instead of (4.4), we make the stronger assumption that:

$$\log R_a|U^{true} = \log R_a|R_a^{true} \approx N(\log R_a^{true}, \frac{\phi_1}{TN_aR_a^{true}}), \tag{4.11}$$

where $U^{true} = (R_a^{true}, R_c^{true}, M_a^{true}, M_c^{true})^T$. The same argument leads to the corresponding results for $\log R_c|U^{true}, \log M_a|U^{true}$ and $\log M_c|U^{true}$.

Furthermore, we make the additional model assumption that $\log R_a, \log R_c, \log M_a$ and $\log M_c$ are conditionally independent, given U. The motivation for this assumption is the intuitive notion that each observed quantity is only affected by the corresponding true quantity. So if all the true quantities are given, the observed quantities are supposed to not affect each other. Then, if $U = (R_a, R_c, M_a, M_c)^T$, we have:

$$\log U|U^{true} \approx N_4\left(\begin{pmatrix} \log R_a^{true} \\ \log R_c^{true} \\ \log M_a^{true} \\ \log M_c^{true} \end{pmatrix}, \begin{pmatrix} \frac{\phi_1}{TN_aR_a^{true}} & 0 & 0 & 0 \\ 0 & \frac{\phi_1}{TN_cR_c^{true}} & 0 & 0 \\ 0 & 0 & \frac{\phi_2}{KN_aM_a^{true}} & 0 \\ 0 & 0 & 0 & \frac{\phi_2}{KN_cM_c^{true}} \end{pmatrix}\right). \tag{4.12}$$

Let $Y = \log \frac{R_a}{R_c}$ and $X = \log \frac{M_a}{M_c}$ denote the estimated treatment effects on the clinical outcome and on the surrogate outcome respectively. We can express Y and X in terms of U:

$$\begin{pmatrix} Y \\ X \end{pmatrix} = A \log U, \text{ where } A = \begin{pmatrix} 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix}. \tag{4.13}$$

Combining (4.3) and (4.12), we obtain the approximations to the first two moments of the estimated treatment effects:

$$E\begin{pmatrix} Y \\ X \end{pmatrix} = E(A\log U) = E(E(A\log U \mid U^{true})) \approx E(A\log U^{true}) = \begin{pmatrix} \alpha+\beta\mu_X \\ \mu_X \end{pmatrix}. \tag{4.14}$$

$$Var\begin{pmatrix} Y \\ X \end{pmatrix} = Var(A\log U) = Var(E(A\log U \mid U^{true})) + E(Var(A\log U \mid U^{true}))$$

$$\approx \begin{pmatrix} (\beta^2\sigma_X^2+\tau^2) + \frac{\phi_1}{TN_a}E(\frac{1}{R_a^{true}}) + \frac{\phi_1}{TN_c}E(\frac{1}{R_c^{true}}) & \beta\sigma_X^2 \\ \beta\sigma_X^2 & \sigma_X^2 + \frac{\phi_2}{KN_a}E(\frac{1}{M_a^{true}}) + \frac{\phi_2}{KN_c}E(\frac{1}{M_c^{true}}) \end{pmatrix}. \tag{4.15}$$

Unlike these marginal moments, the marginal distribution of the estimated treatment effects is difficult to derive. In fact, to obtain the marginal distribution of $(Y,X)^T$, we need to make additional distributional assumptions about U^{true}. On the other hand, as N_a and N_c, the number of patients in the active arm and in the control arm increase, the influence of the estimation errors become small. As a result, the observed endpoints approach the true endpoints and the estimated treatment effects approach the true treatment effects. Since in (4.1) we assume that the true treatment effects follow a joint normal distribution, we may think the normal distribution with moments given by (4.14) and (4.15) is a reasonable approximation to the true distribution of $(Y,X)^T$ for large N_a and N_c.

4.2 Model for the Multiple-contrast Clinical Trials

Besides the 9 single-contrast trials, there are 12 two-contrast trials, 1 three-contrast trial, and 1 four-contrast trial. In each of the two-contrast trials, there is a control

arm, a high dose arm and a low dose arm. For each of the 12 two-contrast trials, let R_{a1}^{true} and R_{a2}^{true} represent the true annualized relapse rate in the high dose arm and in the low dose arm respectively, and let M_{a1}^{true} and M_{a2}^{true} represent the true MRI lesion count per patient per scan in the high dose arm and in the low dose arm respectively. Then, the true treatment effects from the high dose versus control contrast can be expressed as $Y_1^{true} = \log \frac{R_{a1}^{true}}{R_c^{true}}$ and $X_1^{true} = \log \frac{M_{a1}^{true}}{M_c^{true}}$, and the true treatment effects from the low dose versus control contrast can be expressed as $Y_2^{true} = \log \frac{R_{a2}^{true}}{R_c^{true}}$ and $X_2^{true} = \log \frac{M_{a2}^{true}}{M_c^{true}}$. Here, we also omit the subscript i for the ith trial.

To take into account the fact that these two pairs of true treatment effects, (Y_1^{true}, X_1^{true}) and (Y_2^{true}, X_2^{true}), are correlated, we assume a joint normal distribution for them. Focusing on (Y_1^{true}, X_1^{true}) or (Y_2^{true}, X_2^{true}) individually, the marginal distributions of both of these pairs should be the bivariate normal distribution (4.3). This is because we are examining the effects of treatments with similar mechanism of action; whether two contrasts are from one trial or from different trials, they should reflect the same surrogacy relationship. However, to determine the joint distribution of these four quantities, we also need to specify the covariance structure between (Y_1^{true}, X_1^{true}) and (Y_2^{true}, X_2^{true}).

Assuming independence among the true endpoints from different arms, we have:

$$Cov(Y_1^{true}, Y_2^{true}) = Cov(\log \frac{R_{a1}^{true}}{R_c^{true}}, \log \frac{R_{a2}^{true}}{R_c^{true}}) = Var(\log R_c^{true}), \quad (4.16)$$

$$Cov(X_1^{true}, X_2^{true}) = Cov(\log \frac{M_{a1}^{true}}{M_c^{true}}, \log \frac{M_{a2}^{true}}{M_c^{true}}) = Var(\log M_c^{true}), \quad (4.17)$$

$$Cov(Y_1^{true}, X_2^{true}) = Cov(\log \frac{R_{a1}^{true}}{R_c^{true}}, \log \frac{M_{a2}^{true}}{M_c^{true}}) = Cov(\log R_c^{true}, \log M_c^{true}) \quad (4.18)$$

$$Cov(Y_2^{true}, X_1^{true}) = Cov(\log \frac{R_{a2}^{true}}{R_c^{true}}, \log \frac{M_{a1}^{true}}{M_c^{true}}) = Cov(\log R_c^{true}, \log M_c^{true}) \quad (4.19)$$

In principle, these covariances represent 3 new parameters in the joint distribution of $(Y_1^{true}, X_1^{true}, Y_2^{true}, X_2^{true})^T$ in addition to the parameters $\alpha, \beta, \mu_X, \sigma_X^2, \tau^2$

that appear in (4.3). However, note that, $Var(Y_1^{true}) = Var(\log R_{a1}^{true}) + Var(\log R_c^{true})$, where $Var(\log R_{a1}^{true})$ represents the variability of the log of the true annualized relapse rate in the high dose arm across trials and $Var(\log R_c^{true})$ represents the variability of the log of the true annualized relapse rate in the control arm across trials. So, even though in a given trial, $\log R_{a1}^{true}$ and $\log R_c^{true}$ may be quite different due to the treatment effect, the two variabilities across trials may not differ too much. To simplify our model, we assume $Var(\log R_{a1}^{true}) = Var(\log R_c^{true})$. Under this assumption, from (4.3), we obtain:

$$Cov(Y_1^{true}, Y_2^{true}) = Var(\log R_c^{true}) = \frac{1}{2} Var(Y_1^{true}) = \frac{1}{2}(\beta^2 \sigma_X^2 + \tau^2). \quad (4.20)$$

The assumption that $Var(\log M_{a1}^{true}) = Var(\log M_c^{true})$ similarly leads to:

$$Cov(X_1^{true}, X_2^{true}) = Var(\log M_c^{true}) = \frac{1}{2} Var(X_1^{true}) = \frac{1}{2}\sigma_X^2. \quad (4.21)$$

At the same time, note that $Cov(Y_1^{true}, X_1^{true}) = Cov(\log \frac{R_{a1}^{true}}{R_c^{true}}, \log \frac{M_{a1}^{true}}{M_c^{true}}) = Cov(\log R_{a1}^{true}, \log M_{a1}^{true}) + Cov(\log R_c^{true}, \log M_c^{true})$, where $Cov(\log R_{a1}^{true}, \log M_{a1}^{true})$ measures how closely the two true endpoints on the high dose arm are related across trials, and $Cov(\log R_c^{true}, \log M_c^{true})$ measures how closely the two true endpoints on the control arm are related across trials. Even though the true relationship between the two endpoints on the high dose arm may be quite different from that on the control arm, the two measures of closeness may not differ too much. Thus, to simplify our model, we assume $Cov(\log R_{a1}^{true}, \log M_{a1}^{true}) = Cov(\log R_c^{true}, \log M_c^{true})$. Under this assumption, from (4.3), we obtain:

$$\begin{aligned} Cov(Y_1^{true}, X_2^{true}) = Cov(Y_2^{true}, X_1^{true}) = Cov(\log R_c^{true}, \log M_c^{true}) \\ = \tfrac{1}{2} Cov(Y_1^{true}, X_1^{true}) = \tfrac{1}{2}\beta\sigma_X^2. \end{aligned} \quad (4.22)$$

All these assumptions lead to the joint distribution of the true treatment effects

in a two-contrast trial:

$$\begin{pmatrix} Y_1^{true} \\ X_1^{true} \\ Y_2^{true} \\ X_2^{true} \end{pmatrix} \sim N_4 \left(\begin{pmatrix} \alpha+\beta\mu_X \\ \mu_X \\ \alpha+\beta\mu_X \\ \mu_X \end{pmatrix}, \begin{pmatrix} \beta^2\sigma_X^2+\tau^2 & \beta\sigma_X^2 & \frac{1}{2}(\beta^2\sigma_X^2+\tau^2) & \frac{1}{2}\beta\sigma_X^2 \\ \beta\sigma_X^2 & \sigma_X^2 & \frac{1}{2}\beta\sigma_X^2 & \frac{1}{2}\sigma_X^2 \\ \frac{1}{2}(\beta^2\sigma_X^2+\tau^2) & \frac{1}{2}\beta\sigma_X^2 & \beta^2\sigma_X^2+\tau^2 & \beta\sigma_X^2 \\ \frac{1}{2}\beta\sigma_X^2 & \frac{1}{2}\sigma_X^2 & \beta\sigma_X^2 & \sigma_X^2 \end{pmatrix} \right). \tag{4.23}$$

To derive the probabilistic structure of the estimated treatment effects in a two-contrast trial, we first focus on the conditional distribution of the observed endpoints given the true endpoints. Let $\tilde{U} = (R_{a1}, R_{a2}, R_c, M_{a1}, M_{a2}, M_c)^T$ and $\tilde{U}^{true} = (R_{a1}^{true}, R_{a2}^{true}, R_c^{true}, M_{a1}^{true}, M_{a2}^{true}, M_c^{true})^T$ represent the observed and true endpoints respectively. We assume that $\log\tilde{U}|\tilde{U}^{true}$ has the same stochastic behavior as $\log U|U^{true}$ in the single-contrast trials. Then, as in (4.12), we have:

$$\log\tilde{U}|\tilde{U}^{true} \approx N_6(\tilde{U}^{true},\ \text{diag}\left(\frac{\phi_1}{TN_1R_{a1}^{true}}, \frac{\phi_1}{TN_2R_{a2}^{true}}, \frac{\phi_1}{TN_cR_c^{true}}, \frac{\phi_2}{TN_1M_{a1}^{true}}, \frac{\phi_2}{TN_2M_{a2}^{true}}, \frac{\phi_2}{TN_cM_c^{true}}\right)), \tag{4.24}$$

where "diag" indicates a diagonal matrix.

Then, combining (4.23) and (4.24), the estimated treatment effects, $Y_1 = \log\frac{R_{a1}}{R_c}$, $Y_2 = \log\frac{R_{a2}}{R_c}$, $X_1 = \log\frac{M_{a1}}{M_c}$ and $X_2 = \log\frac{M_{a2}}{M_c}$, have the following approximations to their first two moments:

$$(Y_1, X_1, Y_2, X_2)^T \approx \left(\begin{pmatrix} \mu \\ \mu \end{pmatrix}, \begin{pmatrix} \Sigma_1 & \Sigma_3 \\ \Sigma_3 & \Sigma_2 \end{pmatrix} \right), \tag{4.25}$$

where

$$\mu = \begin{pmatrix} \alpha+\beta\mu_X \\ \mu_X \end{pmatrix}$$

and

$$\Sigma_1 = \begin{pmatrix} (\beta^2\sigma_X^2+\tau^2)+\frac{\phi_1}{TN_{a1}}E(\frac{1}{R_{a1}^{true}})+\frac{\phi_1}{TN_c}E(\frac{1}{R_c^{true}}) & \beta\sigma_X^2 \\ \beta\sigma_X^2 & \sigma_X^2+\frac{\phi_2}{KN_{a1}}E(\frac{1}{M_{a1}^{true}})+\frac{\phi_2}{KN_c}E(\frac{1}{M_c^{true}}) \end{pmatrix},$$

$$\Sigma_2 = \begin{pmatrix} (\beta^2\sigma_X^2+\tau^2)+\frac{\phi_1}{TN_{a2}}E(\frac{1}{R_{a2}^{true}})+\frac{\phi_1}{TN_c}E(\frac{1}{R_c^{true}}) & \beta\sigma_X^2 \\ \beta\sigma_X^2 & \sigma_X^2+\frac{\phi_2}{KN_{a2}}E(\frac{1}{M_{a2}^{true}})+\frac{\phi_2}{KN_c}E(\frac{1}{M_c^{true}}) \end{pmatrix},$$

$$\Sigma_3 = \begin{pmatrix} \frac{1}{2}(\beta\sigma_X^2+\tau^2)+\frac{\phi_1}{TN_c}E(\frac{1}{R_c^{true}}) & \frac{1}{2}\beta\sigma_X^2 \\ \frac{1}{2}\beta\sigma_X^2 & \frac{1}{2}\sigma_X^2+\frac{\phi_2}{KN_c}E(\frac{1}{M_c^{true}}) \end{pmatrix}.$$

Similarly as in the single-contrast trial, the marginal distribution of the estimated treatment effects are difficult to derive, since we need to make additional distributional assumptions about $\tilde{U}^{true}$. As before, we may think the normal distribution with moments given by (4.25) is a reasonable approximation to the true distribution of $(Y_1, X_1, Y_2, X_2)^T$ for large N_{a1}, N_{a2} and N_c.

For the single three-contrast trial we have 6 estimated treatment effects, and for the single four-contrast trial we have 8 estimated treatment effects. Deriving the first two moments of those 6 and 8 estimated treatment effects proceeds analogously to the above development for the 4 estimated treatment effects in the two-contrast trial.

4.3 Parameter Estimation

From (4.25), we have approximations to the first two moments of the estimated treatment effects. In order to estimate the model parameters, we use the normal estimating equations: that is, we pretend the estimated treatment effects are multivariate normally distributed with the mean vector and variance covariance matrix

given by (4.25). Then maximum likelihood estimates (MLE) of the model parameters are obtained by maximizing the "normal likelihood".

In addition to the parameters of primary interest, $\alpha, \beta, \mu_X, \sigma_X^2, \tau^2, \phi_1$ and ϕ_2, there are several nuisance parameters in the covariance matrices that appear in this "likelihood" function, namely the expectations of the reciprocal of the true relapse rates and lesion counts such as $E(\frac{1}{R_c^{true}})$ and $E(\frac{1}{M_c^{true}})$ in (4.25). When fitting the model, to avoid too many parameters to be estimated in the maximization procedure, we treat these terms as known and replace them by estimates.

As mentioned in Section 3.3.1, we assume that all the R_a^{true}s in different contrasts have the same distribution and all the R_c^{true}s in different contrasts also have the same distribution. As a result:

$$E(R_a^{true}) = E(R_{a1}^{true}) = E(R_{a2}^{true}), \quad \text{for all the contrasts.} \tag{4.26}$$

Also, from (3.8), (3.9) and (3.10), we know that:

$$E(R_a) = E(E(R_a|R_a^{true})) = E(R_a^{true}). \tag{4.27}$$

From the delta method, we have the rough approximation:

$$E(\frac{1}{R_a^{true}}) \approx \frac{1}{E(R_a^{true})} = \frac{1}{E(R_a)}. \tag{4.28}$$

This means that we can use the observed annualized relapse rates to estimate the nuisance parameter $E(\frac{1}{R_a^{true}})$. From the total of 40 contrasts, we estimate $E(\frac{1}{R_a^{true}})$ by the inverse of the average value of the 40 observed annualized relapse rates on the active arms. We estimate $E(\frac{1}{R_c^{true}})$ similarly using the observed annualized relapse rates on the 23 control arms. By the same argument, we estimate $E(\frac{1}{M_a^{true}})$ and $E(\frac{1}{M_c^{true}})$ by using the observed MRI lesion counts per patient per scan from the 40 active arms and the 23 control arms respectively. As a result, we have

$\hat{E}(\frac{1}{R_a^{true}}) \approx 1.43$, $\hat{E}(\frac{1}{R_c^{true}}) \approx 1.10$, $\hat{E}(\frac{1}{M_a^{true}}) \approx 0.57$ and $\hat{E}(\frac{1}{M_c^{true}}) \approx 0.41$.

To maximize the "normal likelihood", we use the R function optim. The maximization procedure is based on the Nelder-Mead method [16]. The optimization process is "two-staged": after obtaining the optimized parameter estimates from each initial value, we set these as an initial value and run the optimization again to obtain a final result. The reason for doing the two-stages is that the first stage often converges to a local minimum.

To avoid negative estimates for σ_X and τ in the optimization, we re-parameterize them as $\eta_X = \log(\sigma_X)$ and $\eta = \log(\tau)$. The first set of initial values for $\hat{\alpha}, \hat{\beta}, \hat{\mu}_X, \hat{\eta}_X$, $\hat{\eta}, \hat{\phi}_1$ and $\hat{\phi}_2$ were -0.02, 0.55, -0.69, -0.04, -1.21, 1.5 and 1.5. The values for $\hat{\alpha}$ and $\hat{\beta}$ are from the SBRCMB result, the values for $\hat{\mu}_X, \hat{\eta}_X$ and $\hat{\eta}$ are based on the method of moments, and the values for $\hat{\phi}_1$ and $\hat{\phi}_2$ are chosen somewhat arbitrarily.

We then tried 999 different sets of random initial values, generating these initial values from independent uniform distributions. Specifically, we generate initial values for $\hat{\alpha}, \hat{\beta}, \hat{\mu}_X, \hat{\eta}_X, \hat{\eta}, \hat{\phi}_1$ and $\hat{\phi}_2$ uniformly on $(-0.5, 0.5)$, $(0,1)$, $(-2,0)$, $(-4.5, 0.5)$, $(-5,0)$, $(0.01, 10)$ and $(0.01, 20)$ respectively. Nearly all of these initial values led to convergence to a very similar optimization result. We choose the estimate which returned the smallest negative log "likelihood" as the final solution.

To calculate the standard errors of the parameter estimates based on the asymptotic normality of the MLE, we invert the negative hessian matrix of the log "likelihood" function and evaluate it at the parameter estimates. We also calculate standard errors for the parameter estimates based on the jackknife method, where we consider the 23 clinical trials as units and estimate the parameters after "leaving one out". We generate 23 different subsets of the original 23 clinical trials; the ith subset is without the ith clinical trial. If the estimate of β from the ith

subset is $\hat{\beta}_{(i)}$, then the jackknife estimate of the standard error of $\hat{\beta}$ is given by $[\frac{22}{23}\sum(\hat{\beta}_{(i)}-\hat{\beta}_{(.)})^2]^{0.5}$, where $\hat{\beta}_{(.)}$ is the average of all $\hat{\beta}_{(i)}$s [17]. Strictly speaking, this is not an appropriate application of the jackknife method, since different trials have different numbers of patients and different numbers of arms, which cause the estimation errors in different trials to be not identical. So the resulting estimated standard errors should be viewed as only "rough and ready" approximations.

The parameter estimates and the corresponding estimated standard errors are shown in Table 4.1 and the estimated asymptotic correlation matrix of $\hat{\alpha}, \hat{\beta}, \hat{\mu_X}$, $\hat{\sigma}_X^2, \hat{\tau}^2, \hat{\phi}_1$ and $\hat{\phi}_2$ based on the MLE method is:

$$\hat{R} = \begin{pmatrix} 1.000 & 0.776 & -0.056 & -0.394 & -0.002 & -0.442 & 0.468 \\ 0.776 & 1.000 & 0.108 & -0.479 & -0.007 & -0.414 & 0.444 \\ -0.056 & 0.108 & 1.000 & -0.106 & -0.002 & -0.158 & 0.194 \\ -0.394 & -0.479 & -0.106 & 1.000 & 0.003 & 0.215 & -0.410 \\ -0.002 & -0.007 & -0.002 & 0.003 & 1.000 & -0.001 & -0.004 \\ -0.442 & -0.414 & -0.158 & 0.215 & -0.001 & 1.000 & -0.557 \\ 0.468 & 0.444 & 0.194 & -0.410 & -0.004 & -0.557 & 1.000 \end{pmatrix} \tag{4.29}$$

Table 4.1: Results of the Model Fit

	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\mu_X}$	$\hat{\sigma_X}^2$	$\hat{\tau}^2$	$\hat{\phi}_1$	$\hat{\phi}_2$
Value	0.081	0.622	-0.713	0.521	< 0.001	0.825	37.427
Normal SE	0.084	0.074	0.156	0.167	< 0.001	0.383	19.932
Jackknife SE	0.105	0.150	0.179	0.198	0.003	0.498	33.496

Although that all the jackknife standard errors are larger than the corresponding MLE standard errors, the results of the statistical tests for significance of the estimates are consistent from these two methods (except for $\hat{\phi}_1$).

Recall that $Y^{true} = \log \frac{R_a^{true}}{R_c^{true}}$ and $X^{true} = \log \frac{M_a^{true}}{M_c^{true}}$. When a treatment has a beneficial effect, we expect a lower MRI lesion count and a smaller relapse rate, which means $Y^{true} < 0$ and $X^{true} < 0$. Therefore, an increase in the true treatment effect corresponds to a decrease in Y^{true} and in X^{true}. So, $\hat{\beta} = 0.622$ means that on average, a one unit increase in the true treatment effect on the MRI lesion count per patient per scan is associated with a 0.622 unit increase in the true treatment effect on the annualized relapse rate. Note this value is larger than the $\hat{\beta} = 0.55$ obtained with the SBRCMB approach (see Table 3.1). As the SBRCMB approach didn't take into account the estimation errors, their regression coefficient of 0.55 may underestimate the association between the true treatment effects due to the attenuation effect.

Although the value for $\hat{\alpha}$ of 0.081 is larger than the $\hat{\alpha} = -0.02$ from the SBRCMB approach, its approximate 95% confidence interval still covers 0. The estimate of $\hat{\alpha}$ being not significantly different from 0 is consistent with a good surrogacy relationship, since there is no strong indication of part of the true treatment effect on the annualized relapse rate being unexplained by the true treatment effect on the MRI lesion count per patient per scan. Finally, the value for $\hat{\tau}^2$ is almost 0, which suggests a nearly perfect linear relationship between the true treatment effects. One can predict the true treatment effect on the annualized relapse rate almost without error based on the true treatment effect on the MRI lesion count per patients per scan. As mentioned at the end of Section 4.1.1, the Prentice definition requires that $\alpha = 0$ and $\tau^2 = 0$. So, under our model assumptions, the MRI lesion count per patient per scan appears to be a very good surrogate endpoint.

Buyse et al. [13] suggest to use R^2_{trial} to evaluate the true surrogacy relationship. Analogous to (2.14) and (2.15), $\beta^2\sigma_X^2 + \tau^2$ represents the uncertainty of predicting the true treatment effect on the clinical endpoint without the information of the surrogate endpoint, and τ^2 represents the uncertainty with the information of the surrogate endpoint. Thus, the difference $\beta^2\sigma_X^2$ represents how much we

gain from using the surrogate. From Table 4.1, we have

$$\hat{R}^2_{trial} = \frac{\hat{\beta}^2 \hat{\sigma}^2_X}{\hat{\beta}^2 \hat{\sigma}^2_X + \hat{\tau}^2} \approx 1. \tag{4.30}$$

The estimate of R^2_{trial} of almost 1 suggests a very good surrogacy relationship. As a result, we can say that, at the trial level, the MRI lesion count per patient per scan has been validated as a surrogate endpoint for the annualized relapse rate in RRMS. However, the estimate of τ^2 being almost 0 or the estimate of R^2_{trial} being almost 1 may not guarantee a high precision in predicting the true treatment effect on the annualized relapse rate in a new trial. In Section 4.5, we will assess this using the estimated surrogacy relationship to make such predictions.

As noted earlier, the jackknife method may not be very appropriate since the 23 trials which we treat as units cannot be considered as a random sample. Of course, the standard errors calculated by the MLE method is also approximate, because we don't have the true likelihood. In the following sections, we use the standard errors based on the asymptotic normality of the MLE to develop our results.

4.4 Comparison between the Comprehensive Approach and the SBRCMB Approach

In a contrast from a new clinical trial (we use the subscript "0" to denote this new contrast), if we know the true treatment effect on the MRI lesion count per patient per scan, X_0^{true}, we can use it to predict the true treatment effect on the annualized relapse rate, Y_0^{true}. In practice, however, there are only a limited number of patients included in any trial and we only have the estimated treatment effect X_0. So, we need to use X_0 instead of X_0^{true} to predict Y_0^{true}; that is, we want to use the surrogacy relationship to predict the treatment effect on the clinical endpoint

based on the estimated treatment effect on the surrogate endpoint.

To identify the relationship between Y_0^{true} and X_0, first note that $Cov(Y_0^{true}, X_0) = E(Y_0^{true} X_0) - E(Y_0^{true})E(X_0)$. We assume this new trial has similar inclusion criteria and involves the same type of treatment as the 23 trials included in the SBR-CMB dataset. So, from (4.3) and (4.14), we have $E(X_0) \approx E(X_0^{true})$. Let $U_0^{true} = (R_{a0}^{true}, R_{c0}^{true}, M_{a0}^{true}, M_{c0}^{true})^T$ denote the true endpoints from the new contrast. Then, from (4.12), we have $E(Y_0^{true} X_0) = E(E(Y_0^{true} X_0 | U_0^{true})) \approx E(Y_0^{true} X_0^{true})$. Therefore:

$$Cov(Y_0^{true}, X_0) \approx E(Y_0^{true} X_0^{true}) - E(Y_0^{true})E(X_0^{true}) = Cov(Y_0^{true}, X_0^{true}). \quad (4.31)$$

As a result, we have the following approximation to the moment structure for Y_0^{true} and X_0:

$$\begin{pmatrix} Y_0^{true} \\ X_0 \end{pmatrix} \approx \left(\begin{pmatrix} \alpha + \beta\mu_X \\ \mu_X \end{pmatrix}, \begin{pmatrix} \beta^2\sigma_X^2 + \tau^2 & \beta\sigma_X^2 \\ \beta\sigma_X^2 & \sigma_X^2 + \frac{\phi_2}{K_0 N_{a0}} E(\frac{1}{M_{a0}^{true}}) + \frac{\phi_2}{K_0 N_{c0}} E(\frac{1}{M_{c0}^{true}}) \end{pmatrix} \right), \quad (4.32)$$

where K_0 is the total number of scans on each patient in the new trial, and N_{a0}, N_{c0} are the number of patients in the active and control arms in the new trial respectively.

The point prediction for Y_0^{true} can be based on $E(Y_0^{true}|X_0)$, but determination of a prediction interval for Y_0^{true} requires information on the conditional distribution of Y_0^{true} given X_0. To derive this distribution, we use the normal distribution with moments given by (4.32) as an approximation to the joint distribution of Y_0^{true} and X_0. The joint distribution is unknown, but as N_{a0} and N_{c0}, the number of patients included in this new trial becomes larger, the estimation error on the estimated treatment effect X_0, becomes smaller, and the estimated treatment effect approaches the true treatment effect X_0^{true}. We may think the bivariate normal distribution is a reasonable approximation for large N_{a0} and N_{c0}.

Under this bivariate normal approximation, we have:

$$E(Y_0^{true}|X_0) = \alpha + \beta\mu_X(1 - \frac{\sigma_X^2}{\sigma_X^2+H_0}) + \frac{\beta\sigma_X^2}{\sigma_X^2+H_0}X_0, \tag{4.33}$$

$$Var(Y_0^{true}|X_0) = \beta^2\sigma_X^2(1 - \frac{\sigma_X^2}{\sigma_X^2+H_0}) + \tau^2, \tag{4.34}$$

where $H_0 = \frac{\phi_2}{K_0N_{a0}}E(\frac{1}{M_{a0}^{true}}) + \frac{\phi_2}{K_0N_{c0}}E(\frac{1}{M_{c0}^{true}})$. So, the point prediction of Y_0^{true} from a future contrast, given the value of $X_0 = x_0$ from that contrast, is:

$$\hat{Y}_0^{true}(x_0) = \hat{E}(Y_0^{true}|X_0 = x_0) = \hat{\alpha} + \hat{\beta}\hat{\mu}_X(1 - \frac{\hat{\sigma}_X^2}{\hat{\sigma}_X^2+\hat{H}_0}) + \frac{\hat{\beta}\hat{\sigma}_X^2}{\hat{\sigma}_X^2+\hat{H}_0}x_0, \tag{4.35}$$

where $\hat{H}_0 = \frac{\hat{\phi}_2}{K_0N_{a0}}E(\frac{1}{M_{a0}^{true}}) + \frac{\hat{\phi}_2}{K_0N_{c0}}E(\frac{1}{M_{c0}^{true}})$. As earlier, $E(\frac{1}{M_{a0}^{true}})$ and $(E(\frac{1}{M_{c0}^{true}}))$ will be treated as known and replaced by the inverse of the average value of the 40 and 23 MRI lesion counts per patient per scan from the active and control arms in the SBRCMB dataset.

The prediction interval for Y_0^{true} given $X_0 = x_0$ can be based on the random variable:

$$W_0 = Y_0^{true}(x_0) - \hat{Y}_0^{true}(x_0). \tag{4.36}$$

Note that given $X_0 = x_0$, $Y_0^{true}(x_0)$ and $\hat{Y}_0^{true}(x_0)$ are independent, so $Var(W_0) = Var(Y_0^{true}(x_0)) + Var(\hat{Y}_0^{true}(x_0))$. From (4.34), we know that $Var(Y_0^{true}(x_0)) = \beta^2\sigma_X^2(1 - \frac{\sigma_X^2}{\sigma_X^2+H_0}) + \tau^2$. Furthermore, the delta method can be used to approximate $Var(\hat{Y}_0^{true}(x_0))$. Specifically, let Σ_W denote the asymptotic covariance matrix of $\hat{\alpha}, \hat{\beta}, \hat{\mu}_X, \hat{\sigma}_X^2$ and $\hat{\phi}_2$, and let g denote the partial derivatives of $E(Y_0^{true}|X_0 = x_0)$ with respect to $\alpha, \beta, \mu_X, \sigma_X^2$ and ϕ_2 (see Appendix *B*). Then:

$$Var(\hat{Y}_0^{true}(x_0)) \approx g^T \cdot \Sigma_W \cdot g. \tag{4.37}$$

As a result,

$$Var(W_0) \approx \beta^2 \sigma_X^2 (1 - \frac{\sigma_X^2}{\sigma_X^2 + H_0}) + \tau^2 + g^T \cdot \Sigma_W \cdot g. \tag{4.38}$$

Note that, W_0 is asymptotically normally distributed, so the approximate 95% prediction interval for $Y_0^{true}(x_0)$ can be given by:

$$\hat{Y}_0^{true}(x_0) \pm 1.96\sqrt{\hat{Var}(W_0)}, \tag{4.39}$$

where, $\hat{Var}(W_0) = \hat{\beta}^2 \hat{\sigma}_X^2 (1 - \frac{\hat{\sigma}_X^2}{\hat{\sigma}_X^2 + \hat{H}_0}) + \hat{\tau}^2 + \hat{g}^T \cdot \hat{\Sigma}_W \cdot \hat{g}$, and $\hat{g}, \hat{\Sigma}_W$ are the partial derivatives and the asymptotic variance covariance matrix of the parameter estimators evaluated at their estimated values.

Figure 4.1 shows the comparison between the SBRCMB results and the comprehensive results in predicting Y_0^{true} from X_0. Although the regression relationship modeled in the SBRCMB approach is between the two estimated treatment effects, for this purpose, we pretend it is between the true treatment effect on the clinical endpoint and the estimated treatment effect on the surrogate endpoint. The SBRCMB prediction line is $y = -0.02 + 0.55x$ while the prediction line for the comprehensive model is given by (4.35). To allow a specific illustration in the figure, we fixed K_0 at 6 (the median number of total scans among the 40 contrasts in the SBRCMB dataset) and N_{a0}, N_{c0} at 50 (the median number of patients among 23 placebo and 40 active arms in the SBRCMB dataset); for these values, $\hat{\alpha} + \hat{\beta}\hat{\mu}_X(1 - \frac{\hat{\sigma}_X^2}{\hat{\sigma}_X^2 + \hat{H}_0}) \approx 0$ and $\frac{\hat{\beta}\hat{\sigma}_X^2}{\hat{\sigma}_X^2 + \hat{H}_0} \approx 0.50$, so (4.35) becomes $y = 0.50x$. The points represent the 40 pairs of estimated treatment effects from the SBRCMB dataset.

From Figure 4.1, we can see that for X between -4 and 1 (the range of X in the SBRCMB dataset), the two prediction lines don't differ much: the point predictions for Y_0^{true} based on X_0 from these two approaches are close. However, when

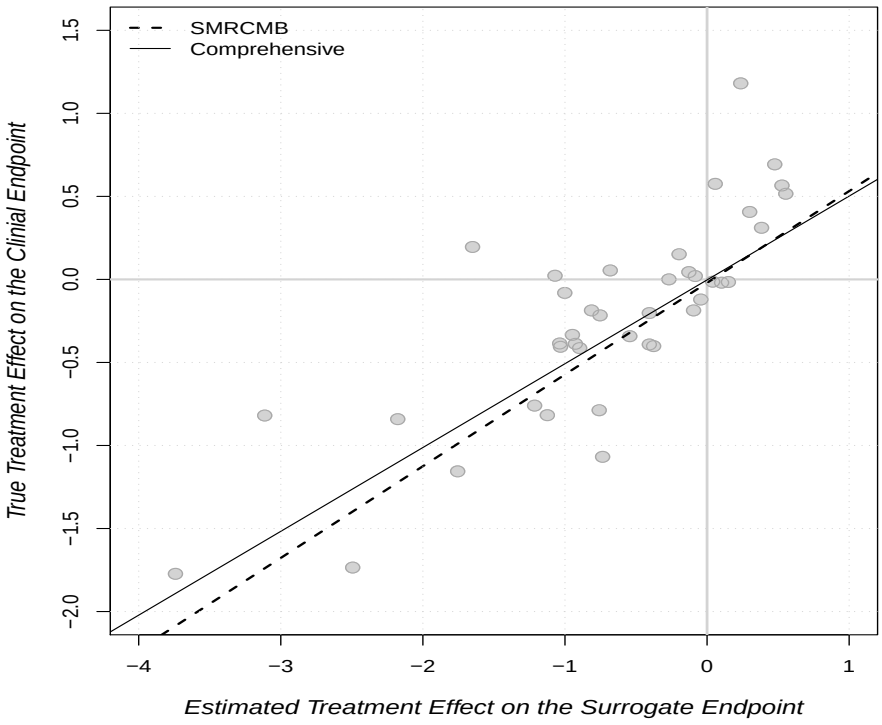

Figure 4.1: Regression Prediction Lines: the SBRCMB Approach ($y = -0.02 + 0.55x$) and the Comprehensive Approach with $K_0 = 6$ and $N_{a0} = N_{c0} = 50$ ($y = 0.50x$).

$X < 0$, the prediction line from the comprehensive approach is above that from the SBRCMB approach. Note that, when $X_0 < 0$, the treatment in the new trial shows a beneficial effect on the surrogate endpoint. When $Y_0^{true} < 0$, the true treatment effect on the clinical endpoint is beneficial, and more negative Y_0^{true} values represent greater beneficial effects. So, Figure 4.1 implies that for a future trial with moderate sample size (50 patients in each arm, for example) and a total of 6 scans, if the treatment shows a beneficial effect on the surrogate endpoint, the true treatment effect on the clinical endpoint predicted by the SBRCMB approach is always slightly greater than that predicted by the comprehensive approach. This means when prediction of the true treatment effect on the clinical endpoint is based on the estimated treatment effect on the surrogate endpoint (on which estimation er-

rors exist), the SBRCMB approach may slightly overestimate the true treatment effect on the clinical endpoint.

Figure 4.2 shows another comparison between the SBRCMB results and the comprehensive results in predicting Y_0^{true} from X_0^{true}. We pretend that the SBRCMB approach models the regression relationship between the two true treatment effects; the prediction line is $y = -0.02 + 0.55x$. The prediction line from the comprehensive model is also given by (4.35), but now we choose N_{a0} and N_{c0} to be infinity, to reflect the case that the future trial includes sufficient number of patients so that the observed treatment effect on the surrogate endpoint estimates the true treatment effect with negligible error. When N_{a0} and N_{c0} are infinity, (4.35) becomes $y = 0.08 + 0.62x$.

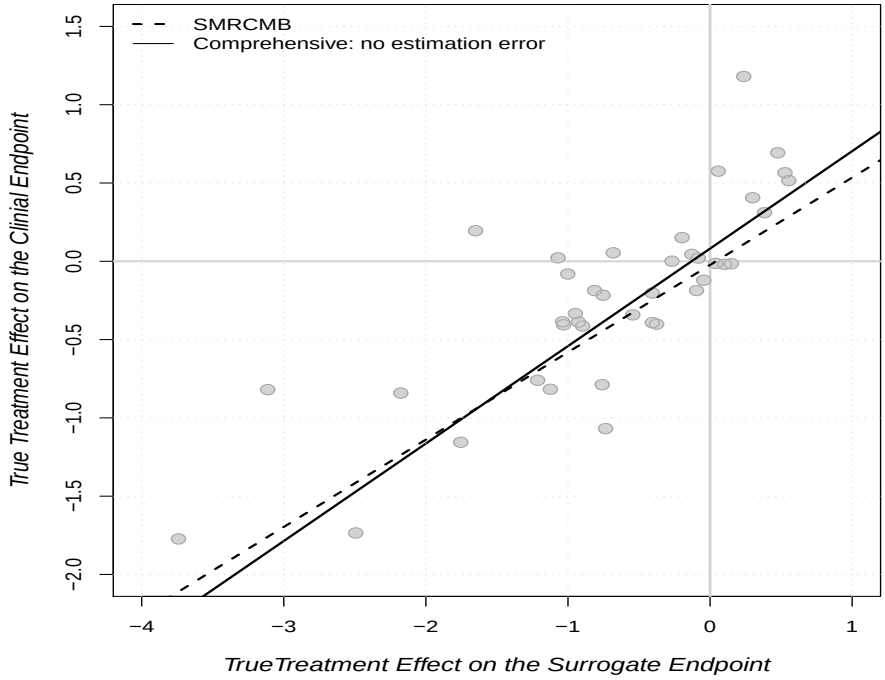

Figure 4.2: Regression Prediction Lines: the SBRCMB Approach ($y = -0.02 + 0.55x$) and the Comprehensive Approach with $K_0 = 6$ and $N_{a0} = N_{c0} = \infty$ ($y = 0.08 + 0.62x$).

From Figure 4.2, we see the two prediction lines intersecting at $X_0^{true} = -1.39$. Note that $\exp(X_0^{true}) = \frac{M_{a0}^{true}}{M_{c0}^{true}}$ and $\exp(-1.39) = 0.25$. So $X_0^{true} = -1.39$ means the treatment leads to a 75% reduction in MRI lesion count per patient per scan in the new trial, which is a large beneficial effect. Therefore, when the true treatment effect on the surrogate endpoint is available, the SBRCMB approach may underestimate/overestimate the true treatment effect on the clinical endpoint if the true treatment effect on the surrogate endpoint is larger/smaller than this value.

We can also compare the point predictions of the two approaches for the 40 contrasts included in the SBRCMB dataset. The SBRCMB approach still uses the prediction line $y = -0.02 + 0.55x$ to predict all of the Y_0^{true}s. But since each contrast has a different total number of scans and different numbers of patients, the comprehensive approach yields point predictions of the Y_0^{true}s that are no longer on a straight line.

Figure 4.3 and Figure 4.4 show the comparison between the SBRCMB results and the comprehensive results in predicting Y_0^{true} from X_0, for the 40 contrasts in the SBRCMB dataset. In Figure 4.3, the solid points represent the point predictions for the 40 contrasts from the comprehensive approach, and the transparent points represent the pairs of estimated treatment effects. In Figure 4.4, the point predictions from the comprehensive approach are plotted against the corresponding predictions from the SBRCMB approach.

From Figure 4.3 and 4.4, we can see that the point predictions for the true treatment effect on the clinical endpoints from the two approaches are generally very close. However, when $X_0 < 0$, all the predictions from the comprehensive approach are larger than the corresponding predictions from the SBRCMB approach. So, for those contrasts where the treatments show beneficial effects on the surrogate endpoint, the SBRCMB approach may overestimate the true treatment effects on the clinical endpoint. Again, this is because none of those trials

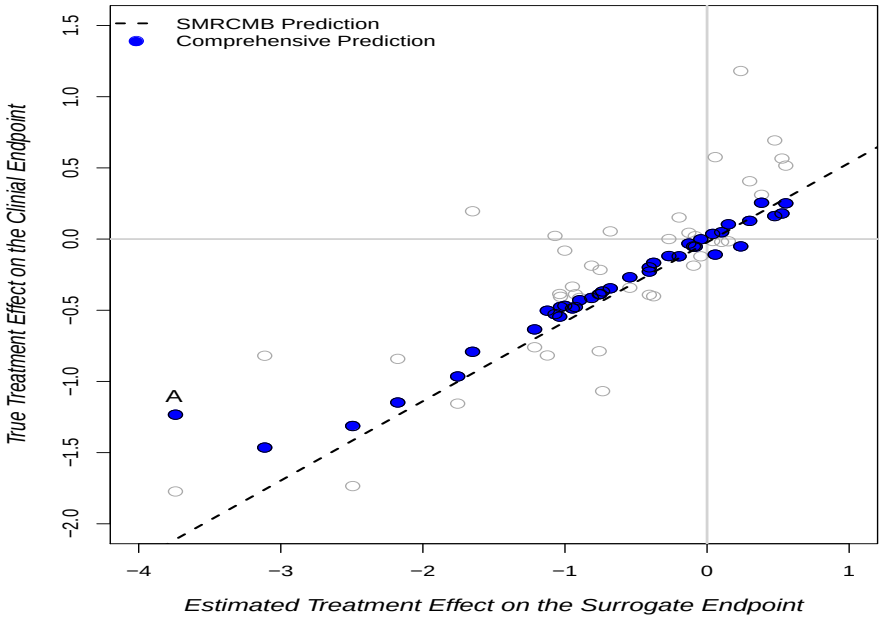

Figure 4.3: Point Predictions for the 40 Contrasts

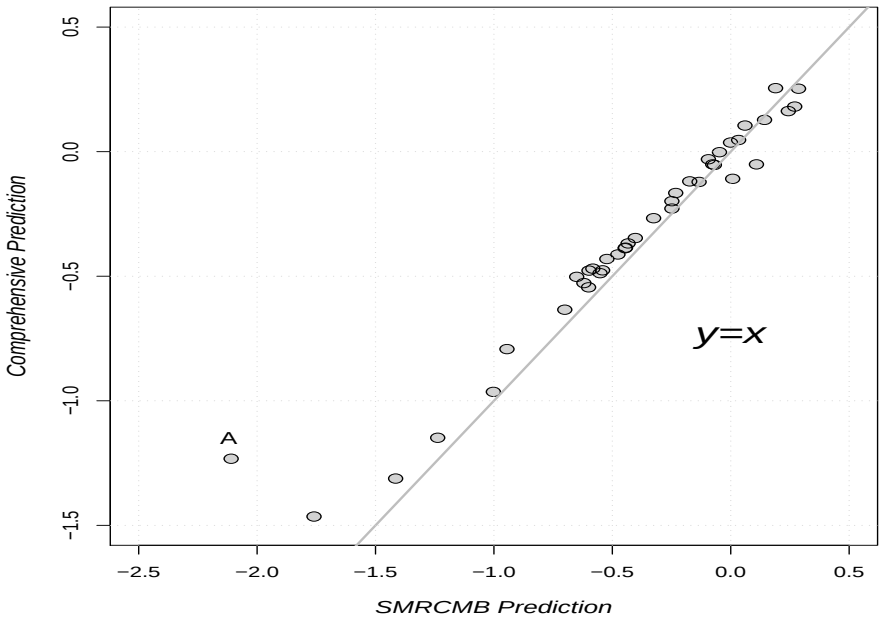

Figure 4.4: Comparison of Point Predictions for the 40 Contrasts

include infinite number of patients, so estimation error exists in the measurement of the treatment effect on the surrogate endpoint. The SBRCMB prediction may be a little more liberal due to its failure to take into account the estimation error. The point A in Figure 4.3 and 4.4 shows the effect of estimation error on predicting the true treatment effect on the clinical endpoint clearly. Note that this point deviates substantially from the remaining points. This point represents the single-contrast clinical trial which has only 10 patients in each arm. So the estimation error in the measurement of the treatment effect on the surrogate endpoint is very large. From (4.35), we know that when N_{a0} and N_{c0} are very small, $\frac{\hat{\beta}\hat{\sigma}_X^2}{\hat{\sigma}_X^2+\hat{H}_0}$ is much smaller than $\hat{\beta}$. This is why the point A deviates substantially from the rest of the points in the y direction. This means, with a large estimation error in the measurement of the treatment effect on the surrogate endpoint, a large estimated treatment effect on the surrogate endpoint may not be associated with a large true treatment effect on the clinical endpoint.

We can also compare the prediction intervals of the two approaches. The prediction interval for $Y_0^{true}(x_0)$ from the comprehensive approach can be calculated from (4.39), and the prediction interval from the SBRCMB approach can be calculated from the standard regression method. (To do so, we pretend the SBRCMB approach models the regression relationship between the true treatment effect on the clinical endpoint and the estimated treatment effect on the surrogate endpoint.) Table 4.2 shows the result of the approximate 95% prediction intervals of $\exp(Y_0^{true}(x_0))$ for the 40 contrasts included in the SBRCMB dataset. Note that $\exp(Y_0^{true}) = \frac{R_{a0}^{true}}{R_{c0}^{true}}$, which represents the true treatment effect on the annualized relapse rate in a future contrast, expressed as a percentage. Table 4.2 is ordered based on the magnitude of $\exp(X_0) = \frac{M_{a0}}{M_{c0}}$, the estimated percentage treatment effect on the surrogate endpoint. The first column is the ID of the contrast in the SBRCMB dataset (see Appendix *A*).

Table 4.2: Comparison of the Approximate 95% Prediction Intervals for $\exp(Y_0^{true}(x_0))$ for the SBRCMB and Comprehensive Approaches

Contrast ID	$\exp(X_0)$	SBRCMB		Comprehensive	
		Point	Interval	Point	Interval
3	0.02	0.12	(0.02, 0.60)	0.29	(0.12, 0.72)
29	0.04	0.17	(0.08, 0.35)	0.23	(0.13, 0.42)
20	0.08	0.24	(0.12, 0.49)	0.27	(0.18, 0.40)
21	0.11	0.29	(0.15, 0.58)	0.32	(0.22, 0.47)
28	0.17	0.37	(0.30, 0.45)	0.38	(0.29, 0.50)
15	0.19	0.39	(0.14, 1.09)	0.45	(0.28, 0.74)
25	0.30	0.50	(0.26, 0.95)	0.53	(0.38, 0.74)
4	0.32	0.52	(0.14, 1.95)	0.60	(0.33, 1.12)
14	0.34	0.54	(0.19, 1.53)	0.59	(0.37, 0.95)
8	0.35	0.55	(0.33, 0.91)	0.58	(0.43, 0.78)
40	0.36	0.55	(0.18, 1.68)	0.62	(0.34, 1.12)
1	0.37	0.56	(0.21, 1.48)	0.63	(0.35, 1.12)
26	0.39	0.58	(0.30, 1.12)	0.61	(0.43, 0.87)
27	0.40	0.58	(0.30, 1.14)	0.62	(0.44, 0.88)
2	0.41	0.59	(0.22, 1.59)	0.65	(0.36, 1.16)
36	0.44	0.62	(0.25, 1.51)	0.66	(0.44, 0.99)
10	0.47	0.64	(0.38, 1.09)	0.68	(0.49, 0.94)
24	0.47	0.64	(0.34, 1.21)	0.68	(0.49, 0.94)
6	0.48	0.65	(0.30, 1.40)	0.69	(0.34, 1.41)
38	0.51	0.67	(0.27, 1.63)	0.71	(0.47, 1.05)
7	0.58	0.72	(0.43, 1.20)	0.77	(0.57, 1.03)

Table 4.2: (continued)

Contrast ID	$\exp(X_0)$	SBRCMB		Comprehensive	
		Point	Interval	Point	Interval
5	0.67	0.78	(0.53, 1.16)	0.80	(0.49, 1.30)
33	0.67	0.78	(0.47, 1.30)	0.82	(0.58, 1.16)
18	0.69	0.79	(0.52, 1.20)	0.85	(0.66, 1.08)
9	0.76	0.84	(0.49, 1.44)	0.89	(0.64, 1.23)
19	0.82	0.88	(0.38, 2.04)	0.89	(0.56, 1.40)
30	0.88	0.91	(0.55, 1.51)	0.97	(0.73, 1.30)
39	0.91	0.93	(0.38, 2.26)	0.95	(0.63, 1.42)
11	0.92	0.93	(0.37, 2.35)	0.95	(0.62, 1.45)
23	0.96	0.95	(0.75, 1.21)	1.00	(0.71, 1.41)
32	1.04	1.00	(0.59, 1.70)	1.04	(0.72, 1.49)
16	1.06	1.01	(0.17, 5.97)	0.90	(0.48, 1.68)
37	1.11	1.03	(0.42, 2.53)	1.05	(0.70, 1.58)
22	1.16	1.06	(0.84, 1.35)	1.11	(0.78, 1.57)
17	1.27	1.11	(0.19, 6.60)	0.95	(0.51, 1.79)
13	1.35	1.15	(0.45, 2.98)	1.14	(0.73, 1.77)
31	1.47	1.21	(0.71, 2.07)	1.29	(0.94, 1.77)
34	1.61	1.27	(0.61, 2.68)	1.18	(0.71, 1.96)
35	1.69	1.31	(0.62, 2.79)	1.20	(0.71, 2.01)
12	1.74	1.33	(0.51, 3.45)	1.29	(0.82, 2.02)

From Table 4.2, we find that the lengths of the prediction intervals from the comprehensive approach are generally shorter than those obtained from the SBRCMB approach (34 out of 40 are shorter), which indicates that the comprehensive approach gives more precise prediction. This can be explained by the existence of estimation error in the measurement of the treatment effect on the clinical endpoint. Although we pretend that the SBRCMB approach can be used to predict Y_0^{true}, it actually predicts Y_0. Since in general, Y_0 is more variable than Y_0^{true}, it may not be surprising that the SBRCMB prediction intervals tend to be wider.

Figure 4.5 illustrates this information. The solid points and the solid lines represent the point predictions and the 95% prediction intervals from the SBRCMB approach, while the hollow points and the dashed lines represent those from the comprehensive approach. It is clear from the figure that most of the prediction intervals from the comprehensive approach are shorter than those from the SBRCMB approach.

The second column of Table 4.2 is the estimated percentage treatment effect on the surrogate endpoint. If $X_0 < 0$ or equivalently, $\exp(X_0) = \frac{M_{a0}}{M_{c0}} < 1$, then the treatment showed a beneficial effect on the surrogate endpoint in the contrast. Among the 40 contrasts, there are 30 contrasts where $\frac{M_{a0}}{M_{c0}} < 1$. For those contrasts, we expect to see beneficial true treatment effects on the clinical endpoint; that is, $\exp(Y_0^{true}) = \frac{R_{a0}^{true}}{R_{c0}^{true}} < 1$. However, based on the comprehensive approach, among those 30 contrasts, only 14 have 95% prediction intervals that don't contain 1. So for the other 16 contrasts, we get inconclusive prediction results for the true treatment effect on the clinical endpoint. The SBRCMB results are less definitive; only 7 contrasts have 95% prediction intervals that don't contain 1. In the next section, we will study how the magnitude of the estimated treatment effect on the surrogate endpoint and the number of patients influence the prediction interval of the true treatment effect on the clinical endpoint.

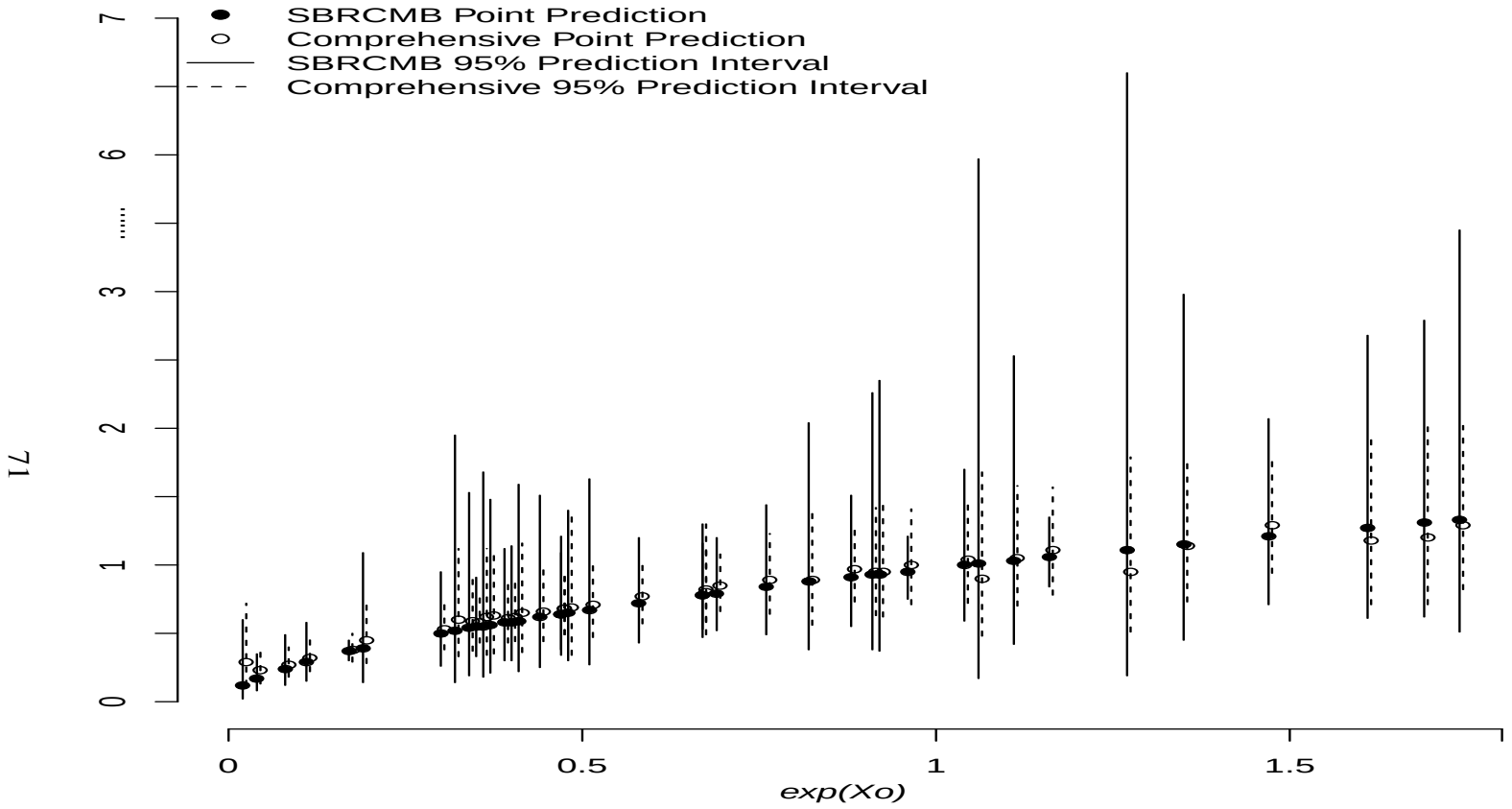

Figure 4.5: Comparison of the Approximate 95% Prediction Intervals for $\exp(Y_0^{true}(x_0))$ for the SBRCMB and Comprehensive Approaches

4.5 Assessment of the Estimated Surrogacy Relationship in Practice

For the MRI lesion count per patient per scan to be a useful surrogate endpoint in practice, it must provide precise enough information on the true treatment effect on the annualized relapse rate. Table 4.3 investigates the influence of the magnitude of X_0 (or $\exp(X_0)$) and the sample size N_{a0}, N_{c0} of the future contrast on the prediction interval for $Y_0^{true}(x_0)$ (or $\exp(Y_0^{true}(x_0))$) calculated from the comprehensive approach. When calculating the prediction intervals, we fix $K_0 = 6$. We set $N_{a0} = N_{c0} = N_0$ and vary N_0 from 10 to 600 (the number of patients in the arms in the SBRCMB dataset range from 8 to 627). We also vary $\exp(X_0)$ from 0.02 to 1.8 (the values of $\exp(X_0)$ in the SBRCMB dataset range from 0.024 to 1.742). The entries in Table 4.3 are the point predictions and approximate 95% prediction intervals for $\exp(Y_0^{true}(x_0))$.

From Table 4.3, first we note that, within each column (i.e., given the value of the estimated treatment effect on the surrogate endpoint), the length of the approximate 95% prediction interval for the true treatment effect on the clinical endpoint becomes shorter as N_0 increases. This is expected, since larger N_0 represents more information on the new contrast, and the prediction will be more precise. The last row in Table 4.3 represents the situation when a new trial includes infinite number of patients. In such a case, the estimation error in the measurement of the treatment effect on the surrogate endpoint becomes negligible. However, we see the prediction interval for $\exp Y_0^{true}(X_0)$ doesn't shrink to a point: even if we know the true treatment effect on the surrogate endpoint, we still cannot predict the true treatment effect on the clinical endpoint without error. From Table 4.1, we know that $\hat{\tau}^2 \approx 0$, which suggests a nearly perfect linear relationship between the true treatment effects. Therefore, the uncertainty in the last row of Table 4.3 is due to the fact that the surrogacy relationship is not estimated precisely enough (other parameters such as α and β are not estimated precisely enough).

Table 4.3: Influence of the Sample Size N_0 and the Magnitude of the Estimated Treatment Effect on the Surrogate Endpoint on the 95% Prediction Intervals for the True Treatment Effect on the Clinical Endpoint for Trials with $K_0 = 6$ Scans per Patient. The Entries are the Point Predictions and Approximate 95% Prediction Intervals for $\exp(Y_0^{true}(x_0))$.

	$\exp(X_0)$								
N_0	0.02	0.1	0.2	0.5	0.8	0.9	1.0	1.5	1.8
10	0.28 (0.11, 0.70)	0.44 (0.21, 0.93)	0.54 (0.27, 1.07)	0.70 (0.36, 1.35)	0.80 (0.41. 1.55)	0.83 (0.43, 1.61)	0.85 (0.44, 1.66)	0.96 (0.49, 1.89)	1.01 (0.51, 2.02)
20	0.20 (0.09, 0.44)	0.37 (0.20, 0.70)	0.49 (0.27, 0.87)	0.70 (0.41, 1.21)	0.84 (0.49. 1.46)	0.88 (0.51, 1.53)	0.92 (0.53, 1.60)	1.08 (0.61, 1.91)	1.16 (0.65, 2.07)
50	0.14 (0.08, 0.24)	0.31 (0.20, 0.50)	0.44 (0.29, 0.67)	0.70 (0.47, 1.05)	0.89 (0.60, 1.33)	0.94 (0.63, 1.41)	1.00 (0.66, 1.49)	1.22 (0.81, 1.86)	1.34 (0.88, 2.05)
100	0.12 (0.07, 0.19)	0.29 (0.20, 0.41)	0.42 (0.31, 0.58)	0.70 (0.52, 0.95)	0.91 (0.67, 1.25)	0.98 (0.71, 1.33)	1.03 (0.76, 1.42)	1.30 (0.93, 1.80)	1.44 (1.02, 2.02)
200	0.11 (0.07, 0.16)	0.27 (0.20, 0.36)	0.41 (0.32, 0.53)	0.70 (0.56, 0.89)	0.93 (0.73, 1.18)	1.00 (0.78, 1.27)	1.06 (0.82, 1.36)	1.34 (1.02, 1.77)	1.50 (1.12, 2.00)
600	0.10 (0.06, 0.15)	0.26 (0.20, 0.34)	0.40 (0.33, 0.49)	0.70 (0.60, 0.83)	0.94 (0.78, 1.13)	1.01 (0.83, 1.22)	1.07 (0.88, 1.31)	1.38 (1.09, 1.74)	1.54 (1.19, 1.99)
∞	0.10 (0.06, 0.15)	0.26 (0.20, 0.33)	0.40 (0.34, 0.46)	0.70 (0.63, 0.78)	0.94 (0.82, 1.09)	1.02 (0.87, 1.18)	1.08 (0.92, 1.23)	1.40 (1.13, 1.73)	1.56 (1.23, 1.98)

Recall that, $\exp(X_0) = \frac{M_{a0}}{M_{c0}}$ and $\exp(Y_0^{true}) = \frac{R_{a0}^{true}}{R_{c0}^{true}}$. So, when a new treatment is efficacious, we hope to observe $\exp(X_0) < 1$ and expect $\exp(Y_0^{true}) < 1$ (i.e., the upper bound of the approximate 95% prediction interval to be less than 1). On the other hand, when a new treatment has a negative effect, we hope to observe $\exp(X_0) > 1$ and expect $\exp(Y_0^{true}) > 1$ (i.e., the lower bound of the approximate 95% prediction interval to be larger than 1).

The last two columns of Table 4.3 represent the situation when the treatment shows medium or large negative effects on the surrogate endpoint (the treatment is 50% or 80% worse than the control in terms of the observed surrogate endpoint), so we hope to see the lower bound of the prediction interval larger than 1. This only happens when $N_0 \geq 200$ for $\exp(X_0) = 1.5$ and when $N_0 \geq 100$ for $\exp(X_0) = 1.8$. So for negative observed treatment effects on the surrogate endpoint to imply negative true treatments effects on the clinical endpoint, a new contrast needs to include a large number of patients. For those contrasts with a medium or small number of patients or with a less extreme observed treatment effect on the surrogate endpoint, conclusive predictions for the true treatment on the clinical endpoint will not be possible.

The $6th$ and $7th$ columns of Table 4.3 represent the situation when $\exp(X_0)$ is close to 1; that is, the estimated treatment effect on the surrogate endpoint is beneficial but the magnitude is small. We see all the prediction intervals within these two columns contain 1 even when N_0 is infinite. This suggests that when a new treatment shows only a small beneficial effect on the surrogate endpoint, we will not be able to determine if this treatment really has an effect on the clinical endpoint based on the estimated surrogacy relationship. In other words, the estimated surrogacy relationship is not very helpful in such a situation.

The $5th$ column of Table 4.3 shows the situation when $\exp(X_0) = 0.5$, which represents a medium beneficial estimated treatment effect on the surrogate end-

point (50% reduction in the observed surrogate endpoint). However, when $N_0 <$ 100, the prediction intervals all contain 1. So, when a new treatment shows a medium beneficial effect on the surrogate endpoint, we will only be able to conclude this treatment has an effect on the clinical endpoint if the new trial includes sufficient patients.

The first 3 columns of Table 4.3 represent the situation when $\exp(X_0)$ is close to 0; that is, the estimated treatment effect on the surrogate endpoint is beneficial and the magnitude is very large. When $N_0 \geq 20$, all the prediction intervals exclude 1. This means we are 95% sure that an observed beneficial treatment effect on the surrogate endpoint corresponds to a true beneficial treatment effect on the clinical endpoint. On the other hand, how precisely we can determine the magnitude of the true treatment effect on the clinical endpoint is also of interest. This precision is indicated by the length of the prediction interval. Note that when $N_0 \leq 50$, the lengths of all the prediction intervals are no less than 0.3 except for the case when $N_0 = 50$ and $\exp(X_0) = 0.02$. As $N_0 = 50$ is a typical size for a phase 2 clinical trial in RRMS, this suggests the prediction of the true treatment effect on the clinical endpoint may not be very precise for a phase 2 clinical trial of small or medium size. On the other hand, when $N_0 \geq 100$, all the lengths of the prediction intervals are smaller than 0.25 except for the case when $N_0 = 100$ and $\exp(X_0) = 0.2$. This indicates the prediction is relatively precise when a trial has a large number of patients.

We also investigate the relationship between N_0 and the value of $\exp(X_0)$ for which the prediction interval for $\exp(Y_0^{true})$ excludes 1 (we fix $K_0 = 6$). Burzykowsky and Buyse [18] introduced a similar concept called the "surrogate threshold effect". This value represents the least extreme value of the estimated treatment effect on the surrogate endpoint from which we can obtain a conclusive prediction for the true treatment effect on the clinical endpoint. In Figure 4.6 and Figure 4.7, we plot the "threshold value" of $\exp(X_0)$ against N_0. Figure 4.6 shows the re-

sult when a treatment shows a beneficial effect on the surrogate endpoint ($X_0 < 0$), and Figure 4.7 shows the result when a treatment shows a negative effect on the surrogate endpoint ($X_0 > 0$).

From Figure 4.6, we see that when the treatment shows a beneficial effect on the surrogate endpoint, the threshold value increases as N_0 increases. A larger threshold value represents a smaller estimated treatment effect on the surrogate endpoint. So, for a contrast with large number of patients, even though we observe only a relatively small treatment effect on the surrogate endpoint, we can still conclude that the treatment has a beneficial effect on the clinical endpoint. The threshold value for $N_0 = 50$ is $\exp(X_0) = 0.46$, which means in order to conclude that a new treatment has a beneficial effect on the clinical endpoint for a contrast with 50 patients in each arm, this treatment has to be observed to be at least $100\% - 46\% = 54\%$ better than the control on the surrogate endpoint. Similarly, for $N_0 = 10$, 20, 100, 200 and 600, the threshold values are 0.14, 0.30, 0.55, 0.61 and 0.67. Note that the asymptote for the curve is 0.71, which indicates the threshold value obtained when $N_0 = \infty$. So, when we try to predict the true treatment effect on the clinical endpoint based on the estimated surrogacy relationship, we require the new treatment to be at least 29% better than the control on the surrogate endpoint in order to conclude that there is a true beneficial treatment effect on the clinical endpoint.

From Figure 4.7, we see that when the treatment shows a negative effect on the surrogate endpoint, the threshold value decreases as N_0 increases. We can interpret Figure 4.7 in a similar way as Figure 4.6. For example, here the threshold value for $N_0 = 50$ is 2.39, which means in order to conclude that a new treatment has a negative effect on the clinical endpoint for a contrast with 50 patients in each arm, this treatment has to be observed to be 139% worse than the control on the surrogate endpoint. Note that the asymptote here is 1.19. So, when we try to predict the true treatment effect on the clinical endpoint based on the estimated

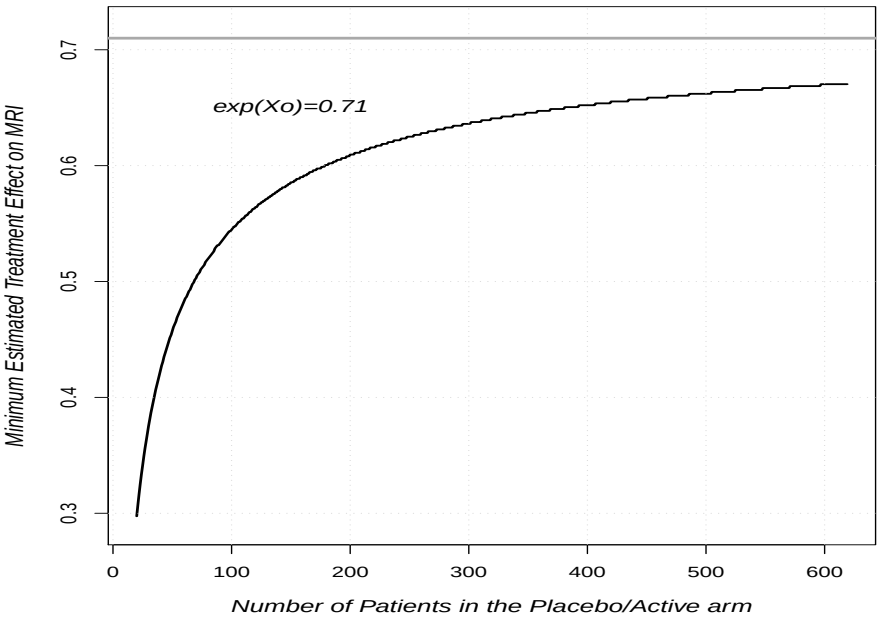

Figure 4.6: Threshold Value of $\exp(X_0)$ versus Sample Size N_0 when a Beneficial Treatment Effect is Observed on the Surrogate Endpoint

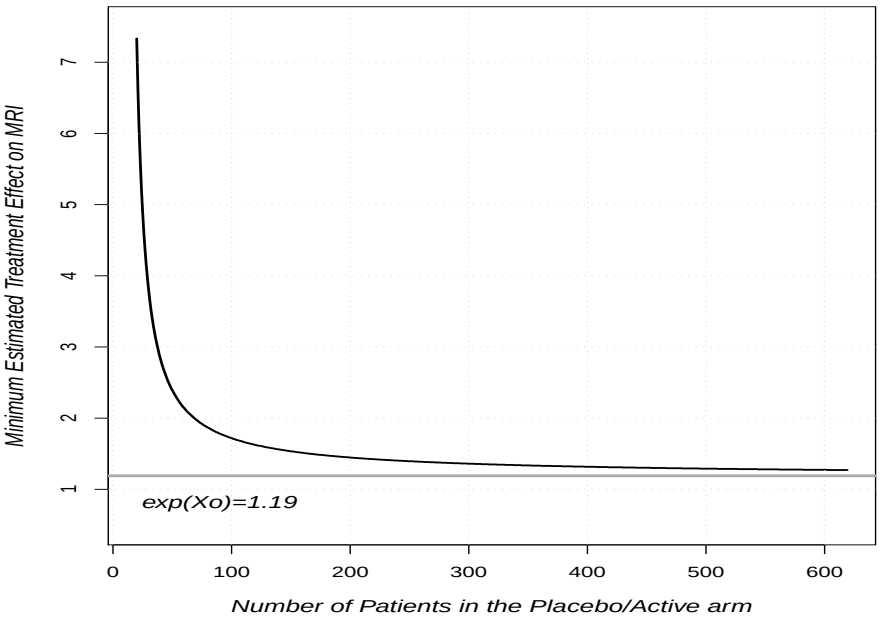

Figure 4.7: Threshold Value of $\exp(X_0)$ versus Sample Size N_0 when a Negative Treatment Effect is Observed on the Surrogate Endpoint

surrogacy relationship, we require the new treatment to be at least 19% worse than the control on the surrogate endpoint in order to conclude that there is a true negative treatment effect on the clinical endpoint.

In conclusion, the estimated surrogacy relationship is useful in predicting the true treatment effect on the clinical endpoint when the treatment shows a large effect on the surrogate endpoint and the number of patients in the contrast is large (e.g. $\exp(X_0) = 0.1$ and N_0 = 100). However, when the treatment shows a moderate beneficial effect on the surrogate endpoint (e.g. $\exp(X_0) = 0.5$), the prediction is not very precise (the prediction interval is wide). When the treatment only shows a small beneficial effect on the surrogate endpoint ($\exp(X_0) > 0.71$), using the estimated surrogacy relationship will lead to an inconclusive result for the true treatment effect on the clinical endpoint.

From (4.30), we know that the true surrogacy relationship may be very good or nearly perfect. Nevertheless, the surrogate endpoint may not be very useful in predicting the true treatment effect on the clinical endpoint unless the treatment shows a large effect on the surrogate endpoint. Furthermore, even if a new trial includes sufficient number of patients so that we can measure the treatment effect on the surrogate endpoint perfectly, we still cannot predict the true treatment effect on the clinical endpoint without error. These may be explained by the limited number of trials included in the SBRCMB dataset. Since we only have 23 trials, we may not estimate the true surrogacy relationship precisely. So, use of the estimated surrogacy relationship may not result in a very precise prediction.

Chapter 5

Conclusions and Discussion

In a clinical trial, a surrogate endpoint is used as a substitute for the clinical endpoint to assess the treatment effect. Using a surrogate endpoint instead of the clinical endpoint can shorten the period of a clinical trial, or reduce the number of patients needed in a clinical trial, and therefore reduce the cost. However, before a potential surrogate endpoint can be formally employed in practice, it must be validated. Use of an invalidated surrogate endpoint can lead to an incorrect conclusion about the treatment effect and thus use of the treatment in future may lead to ineffective or even harmful impact on patients.

A potential surrogate endpoint can be validated in a single clinical trial or in multiple clinical trials if the multiple trials study the same or similar treatments. When the validation is carried on in multiple trials, the validation process can be based on the summary information of each trial or on the individual patient data, depending on whether the individual patient level data is available. When individual patient level data is not available, we lose the possibility of examining how closely a surrogate is related to the clinical endpoint in individual patients, but retain the ability to evaluate the relationship between the treatment effects on the surrogate and the clinical endpoints.

In RRMS clinical trials, changes in MS brain lesion patterns determined by MRI reflect the underlying MS disease pathology and hence may be the best candidate for a surrogate endpoint. In this report, we studied whether the MRI lesion count per patient per scan can serve as a surrogate endpoint for the annualized relapse rate, which is the most commonly used clinical endpoint for RRMS clinical trials. The SBRCMB dataset only includes summary information from 23 clinical trials. Two different approaches (the SBRCMB approach and the comprehensive approach) are applied to the SBRCMB dataset to assess this potential surrogacy relationship.

The SBRCMB approach discussed in Chapter 3 uses simple linear regression with weighted least squares estimation, where the response and the explanatory variables are the estimated treatment effects on the clinical and the surrogate endpoints from each contrast, and the weights are chosen to account for the influence of different numbers of patients and different durations of contrasts. However, this approach treats the estimated treatment effects as the true treatment effects (doesn't take into account the estimation errors) and ignores the correlation structure among contrasts from the same trial.

The comprehensive approach discussed in Chapter 4 assumes a multivariate normal distribution for the true treatment effects to take into account the correlation structure among the contrasts from the same trial, and develops the conditional distribution of the estimated treatment effects given the true endpoints. The approximated marginal moments of the estimated treatment effects are then determined. To estimate the parameters related to the surrogacy relationship, we use the normal estimating equations.

The $\hat{\beta}$ from the comprehensive approach is 0.62, which is larger than 0.55 from the SBRCMB approach. So, the SBRCMB approach may underestimate the association between the true treatment effects. Neither of the $\hat{\alpha}$s from the two

approaches are significantly different from 0, which is consistent with a good surrogacy relationship, since there is no strong indication of part of the true treatment effect on the annualized relapse rate remaining unexplained by the true treatment effect on the MRI lesion count per patient per scan. The SBRCMB approach obtains a weighted $R^2 = 0.80$, and the comprehensive approach obtains $\hat{R}^2_{trial} \approx 1$. Both indicate a good surrogacy relationship. For the comprehensive approach, $\hat{R}^2_{trial} \approx 1$ is equivalent to $\hat{\tau}^2 \approx 0$, which indicates a negligible estimated conditional variance of the true treatment effect on the annualized relapse rate given the true treatment effect on the MRI lesion count per patient per scan. Under the assumptions of the comprehensive approach, the Prentice definition about a surrogate endpoint requires that $\alpha = 0$ and $\tau = 0$. So, the MRI lesion count per patient per scan appears to be a very good surrogate endpoint for the annualized relapse rate.

To assess how good this estimated surrogacy relationship is in practice, we predict the true treatment effect on the clinical endpoint for the 40 contrasts included in the SBRCMB dataset. The point predictions from the two approaches are very close, but those from the comprehensive approach are slightly larger than those from the SBRCMB approach for most contrasts. So, for those trials which showed beneficial treatment effects on the surrogate endpoint, the SBRCMB approach tends to predict slightly larger treatment effects than the comprehensive approach. The interval predictions from the two approaches are quite different however. The length of the prediction interval from the comprehensive approach is generally shorter (34 out of 40 are shorter), which indicates the comprehensive approach gives more precise prediction.

For the comprehensive approach, we also study how the number of patients per arm and the value of the estimated treatment effect on the surrogate endpoint affect the prediction interval for the true treatment effect on the clinical endpoint. For a new contrast with infinite number of patients in each arm (i.e. the estimation

error in the measurement of the treatment effect on the surrogate endpoint is negligible), we require the treatment to be observed to be at least 29% better or 19% worse than the control on the surrogate endpoint, in order to avoid inconclusive prediction for the true treatment effect on the clinical endpoint. For a new contrast with limited number of patients in each arm, we require the treatment to show more extreme effects. For a typical phase 2 clinical trial in RRMS with 50 patients in each arm and with 6 scans for each patient, we require the treatment is at least 54% better or 139% worse. Among the 30 contrasts included in the SBRCMB dataset where the treatments show beneficial effects on the surrogate endpoint, 20 show treatment effects greater than 54%, while among the 10 contrasts where the treatments show negative effects on the surrogate endpoint, only 4 treatments are 139% or more worse than the control. So, the estimated surrogacy relationship could be useful in prediction when a treatment shows an beneficial effect on the surrogate endpoint, but may not be useful in the contrary case. In addition, when the number of patients per arm is around 50, the prediction interval is wide and doesn't yield a precise prediction, unless the treatment shows a very large effect on the surrogate endpoint (e.g. $\geq 90\%$).

In conclusion, the comprehensive approach shows that the underlying surrogacy relationship may be very good. In a typical phase 2 with around 50 patients in each arm and with 6 scans for each patient, the estimated surrogacy relationship can give precise prediction for the true treatment effect on the clinical endpoint when the treatment displays a large effect on the surrogate endpoint. However, when the treatment displays only a modest or a small effect on the surrogate endpoint, the prediction may be inconclusive or not precise enough. The reason for this may be the limited number of trials included in the SBRCMB dataset: the parameters related to the surrogacy relationship may not be estimated precisely enough, which leads to a relatively wide prediction interval. To employ the surrogacy relationship to make predictions in practice, we may need information from more trials to estimate the surrogacy relationship more precisely.

The comprehensive approach we developed is in the spirit of Daniels and Hughes [2] (DH) and Korn et al. [3] (KAM). Both construct models to assess surrogacy relationships using summary results from multiple clinical trials. Both DH and KAM use multivariate normal distributions for the true treatment effects in their models to allow for correlated contrasts. However, DH starts with assumptions about the surrogacy relationship between the true treatment effects directly, while KAM starts with assumptions about the true endpoints, where the influence of the true surrogate endpoint on the true clinical endpoint is assumed to be the same regardless of the presence of the treatment. Building the model from endpoints requires a more detailed specification and we think the KAM assumptions may not be very appropriate in practice, so we started with assumptions about the true treatment effects. On the other hand, both papers assume the estimation errors in estimating the true treatment effects are independent from the true treatment effects. In contrast, we assume they are dependent and large true treatment effects are associated with small estimation errors. We think this dependence assumption is more reasonable in practice. However, not making assumptions about the true endpoints and the dependence estimation errors makes it difficult to obtain the marginal distribution the estimated treatment effects in our model. If one can find a reasonable assumption on the distribution of the true endpoints, then the marginal distribution can be obtained, and the surrogacy relationship could be re-estimated using the actual likelihood rather than the "approximated" likelihood. Furthermore, DH adopt a Bayesian approach to estimate the surrogacy relationship. By choosing appropriate priors for the parameters, we could also use a Bayesian approach to estimate the surrogacy relationship and compare the results to those obtained in this study.

The SBRCMB dataset only contains summary information from each trial but not the individual patient information. If the individual patient information is available, one can re-analyze the surrogacy relationship using the individual pa-

tient level data and compare the results for the estimated surrogacy relationship with those from this study. In principle, the estimated surrogacy relationship from the model with individual patient level data should be more precisely determined, since this model includes more information. However, if the two results are close, one may favor the model based on summary results. This is because it is much easier to collect the summary results of each trial than to collect the individual patient data from each trial, and the estimation process of the model with only summary results may be much less computational intensive. Despite this, if the individual patient information is available, one can assess how closely the surrogate endpoint is related to the clinical endpoint, (e.g. R_{ind} from Buyse et al. [13]), which is useful for patient management.

Bibliography

[1] M. P. Sormani, L. Bonzano, L. Roccatagliata, G. R. Cutter, G. L. Mancardi, and P. Bruzzi. Magnetic resonance imaging as a potential surrogate for relapses in multiple sclerorsis: A meta-analytic approach. *Annals of Neurology*, 65:268–275, 2009.

[2] M. J. Daniels and M. D. Hughes. Meta-analysis for the evaluation of potential surrogate markers. *Statistics in Medicine*, 16:1965–1982, 1997.

[3] E. L. Korn, P. S. Albert and L. M. McShane. Assessing surrogates as trial endpoints using mixed models. *Statistics in Medicine*, 24:163–182, 2005.

[4] T. Burzykowsky, G. Molenberghs and M. Buyse. *The Evaluation of Surrogate Endpoints*. Springer, New York, New York, 2005.

[5] R. L. Prentice. Surrogate endpoints in clinical trials: Definition and operational criteria. *Statistics in Medicine*, 8:431–440, 1989.

[6] H. F. McFarland, F. Barkhof, J. Antel, and D. H. Miller. The role of MRI as a surrogate outcome measure in multiple sclerosis. *Multiple Sclerosis*, 8: 40–51, 2002.

[7] T. R. Fleming and D. L. DeMets. Surrogate endpoints in clinical trials: Are we being misled? *Annals of Internal Medicine*, 125:605–613, 1996.

[8] M. Buyse and G. Molenberghs. Criteria for the validation of surrogate endpoints in randomized experiments. *Biometrics*, 54:1014–1029, 1996.

[9] V. W. Berger. Does the Prentice criterion validate surrogate endpoints? *Statistics in Medicine*, 23:1571–1578, 2004.

[10] L. S. Freedman and B. I. Graubard. Statistical validation of intermediate endpoints for chronic diseases. *Statistics in Medicine*, 11:167–178, 1992.

[11] G. Molenberghs, M. Buyse, H. Geys, D. Renard, T. Burzykowski, and A. Alonso. Statistical challenges in the evaluation of surrogate endpoints in randomized trials. *Cotrolled Clinical Trials*, 23:607–625, 2002.

[12] A. Alonso, G. Molenberghs, T. Burzykowski, D. Renard, H. Geys, Z. Shkedy, F. Tibaldi, J. C. Abrahantes, and M. Buyse. Prentice's approach and the meta-analytic paradigm: A reflection on the role of statisitcs in the evaluation of surrogate endpoints. *Biometrics*, 60:724–728, 2004.

[13] M. Buyse, G. Molenberghs, T. Burzykowsky, D. Renard, and H. Geys. The validation of surrogate endpoint in meta-analyses of rnadomized experiments. *Biometrics*, 1:49–67, 2000.

[14] R. E. Kass and A. E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90:773–795, 1995.

[15] A. J. Petkau, S. C. Reingold, U. Held, G. R. Cutter, T. R. Fleming, M. D. Hughes, D. H. Miller, H. F. McFarland, and J. S. Wolinsky. Magnetic resonance imaging as a surrogate outcome for multiple sclerosis relapses. *Multiple Sclerosis*, 14:770–778, 2008.

[16] J. A. Nelder and R. Mead. A simplex method for function minimization. *Computer Journal*, 7:308–313, 1965.

[17] F. Mosteller and J. W. Tukey. *Data Analysis and Regression, a Second Course in Statistics*. Addison-Wesley, Reading, Massachusetts, 1977.

[18] T. Burzykowsky and M. Buyse. Surrogate threshold effect: an alternative measure for meta-analytic surrogate endpoint validation. *Pharmaceutical Statistics*, 5:173–186, 2006.

Appendix A

The SBRCMB Dataset

In the table that follows, the last four columns represent the observed endpoints from each contrast: MRI = MRI lesion count per patient per scan; ARR = annualized relapse rate. The symbol "C" means "control arm" and the symbol "A" means "active arm". Unless otherwise noted, entries in columns 1, 2, 3, 4, 5, 11 and 12 are copied from the supplementary table accompanying the SBRCMB paper. Entries in the remaining columns are extracted or calculated from the original papers where the results of the corresponding clinical trials are reported.

Trial	Contrast ID	MRI Outcome	SBRCMB Weight	Follow-up (months)	# of Scans	# of Patients		MRI		ARR	
						C	A	C	A	C	A
1	1	Active T2 [a]	37	24	6	17	17	0.82	0.30	1.27	1.17
1	2	Active T2 [a]	36	24	6	17	17	0.82	0.33	1.27	0.84
2	3	Active T2 [b]	14	6	6	10	10	3.37	0.08	2.00	0.34
3	4	Active T2 [b]	20	6	6	14	14	4.22	1.37	1.29	0.57
4	5	Active T2 [b]	233	24	2	82	83	2.40	1.60	0.82	0.67
5	6	New T2	59	24	2	19	23	3.65	1.75	1.31	0.45
6	7	CUA [c]	138	24	10	66	64	1.55	0.90	1.28	0.91
6	8	CUA [c]	140	24	10	66	68	1.55	0.55	1.28	0.87
7	9	CUA [c]	123	12	6	97	87	1.70	1.30	1.08	1.08
7	10	CUA [c]	124	12	6	97	85	1.70	0.80	1.08	0.81
8	11	CUA [c]	41	6	6	43	44	1.48	1.37	0.98	1.00
8	12	CUA [c]	39	6	6	43	40	1.48	2.58	0.98	1.64
8	13	CUA [c]	39	6	6	43	40	1.48	2.00	0.98	1.47
9	14	New Gd	32	6	6	33	32	1.22	0.42	0.88	0.90
9	15	New Gd	33	6	6	33	32	1.22	0.23	0.88	1.07
10	16	New Gd	11	9	9	10	8	3.00	3.18	0.27	0.48
10	17	New Gd	11	9	9	10	8	3.00	3.80	0.27	0.88
11	18	New T2	207	9	9	120	119	1.52	1.04	1.21	0.81
12	19	CUA [c]	49	6	6	34	36	2.42	1.98	1.29	1.50

Trial	Contrast ID	MRI Outcome	SBRCMB Weight	Follow-up (months)	# of Scans	# of Patients		MRI		ARR	
						C	A	C	A	C	A
13	20	CUA [c]	74	6	6	71	68	1.62	0.13	0.51	0.09
13	21	CUA [c]	77	6	6	71	74	1.62	0.18	0.51	0.22
14	22	New T2	758	14	1	467	471	6.80	7.90	0.61	0.60
14	23	New T2	751	14	1	467	462	6.80	6.50	0.61	0.54
15	24	New T2	87	6	6	81	83	1.07	0.50	0.77	0.35
15	25	New T2	84	6	6	81	77	1.07	0.32	0.77	0.36
16	26	CUA [c]	79	9	7	61	61	2.68	1.04	0.81	0.58
16	27	CUA [c]	77	9	7	61	57	2.68	1.06	0.81	0.55
17	28	Active T2 [b]	1332	24	2	315	627	5.50	0.95	0.73	0.23
18	29	New Gd	74	6	4	35	69	1.12	0.05	0.84	0.37
19	30	New Gd	140 [d]	12	8	84	96	0.72	0.64	0.44	0.46
19	31	New Gd	128 [d]	12	8	84	87	0.72	1.06	0.44	0.60
20	32	New T2	129	9	4	102	98	2.40	2.50	0.77	0.76
20	33	New T2	136	9	4	102	106	2.40	1.60	0.77	0.52
21	34	CUA [c]	65	12	4	41	44	4.50	7.25	0.50	1.00
21	35	CUA [c]	63	12	4	41	42	4.50	7.62	0.50	0.88
22	36	New Gd	44 [d]	6	6	49	50	1.73	0.77	0.53	0.44
22	37	New Gd	44 [d]	6	6	49	50	1.73	1.91	0.53	0.52
22	38	New Gd	44 [d]	6	6	49	50	1.73	0.88	0.53	0.56
22	39	New Gd	44 [d]	6	6	49	50	1.73	1.57	0.53	0.44
23	40	New Gd	28	6	5	19	19	1.03	0.37	0.63	0.42

[a] new, recurrent and enlarging T2 lesions

[b] new and enlarging T2 lesions

[c] combined uniquely active lesions = recurrent and enlarging T2 lesions and new Gd enhancing lesions, avoiding double counting

[d] calculated from the original papers; these differ from those in the SBRCMB paper

Appendix B

Partial Derivatives of $E(Y_0^{true}|X_0 = x_0)$

From (4.34), we have:

$$E(Y_0^{true}|X_0 = x_0) = \alpha + \beta\mu_X(1 - \frac{\sigma_X^2}{\sigma_X^2 + H_0}) + \beta\frac{\sigma_X^2}{\sigma_X^2 + H_0}x_0,$$

where $H_0 = \phi_2[\frac{1}{K_0 N_{a0}}E(\frac{1}{M_{a0}^{true}}) + \frac{1}{K_0 N_{c0}}E(\frac{1}{M_{c0}^{true}})] = \phi_2 c_0$ say. Let $L_0 = \frac{\sigma_X^2}{\sigma_X^2 + H_0}$. Then:

$$E(Y_0^{true}|X_0 = x_0) = \alpha + \beta\mu_X(1 - L_0) + \beta L_0 x_0.$$

So:

$$\frac{\partial E}{\partial \alpha} = 1, \quad \frac{\partial E}{\partial \beta} = \mu_X(1 - L_0) + L_0 x_0, \quad \frac{\partial E}{\partial \mu_X} = \beta(1 - L_0),$$

$$\frac{\partial E}{\partial L_0} = -\beta\mu_X + \beta x_0, \quad \frac{\partial L_0}{\partial \sigma_X^2} = \frac{H_0}{(\sigma_X^2 + H_0)^2}$$

$$\frac{\partial E}{\partial \sigma_X^2} = \frac{\partial E}{\partial L_0}\frac{\partial L_0}{\partial \sigma_X^2} = (-\beta\mu_X + \beta x_0)\frac{H_0}{(\sigma_X^2 + H_0)^2},$$

$$\frac{\partial L_0}{\partial H_0} = \frac{-\sigma_X^2}{(\sigma_X^2 + H_0)^2}, \quad \frac{\partial H_0}{\partial \phi_2} = c_0,$$

$$\frac{\partial E}{\partial \phi_2} = \frac{\partial E}{\partial L_0}\frac{\partial L_0}{\partial H_0}\frac{\partial H_0}{\partial \phi_2} = (-\beta\mu_X + \beta x_0)\frac{-\sigma_X^2}{(\sigma_X^2 + H_0)^2}c_0.$$

The entries of the partial derivative of g is then given by:

$$g = (\frac{\partial E}{\partial \alpha}, \frac{\partial E}{\partial \beta}, \frac{\partial E}{\partial \mu_X}, \frac{\partial E}{\partial \sigma_X^2}, \frac{\partial E}{\partial \phi_2})^T.$$