A STUDY OF TREATMENT-BY-SITE INTERACTION IN MULTISITE CLINICAL TRIALS

by

Kaleab Zenebe Abebe

B.A. Mathematics, Goshen College, Goshen, IN 2003

M.A. Statistics, University of Pittsburgh, Pittsburgh, PA 2006

Submitted to the Graduate Faculty of

the Arts & Sciences in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

June 25, 2009

UNIVERSITY OF PITTSBURGH

ARTS & SCIENCES

This dissertation was presented

by

Kaleab Zenebe Abebe

It was defended on

June 25, 2009

and approved by

Satish Iyengar, Ph.D., Professor, Statistics

Leon J. Gleser, Ph.D., Professor, Statistics

Allan R. Sampson, Ph.D., Professor, Statistics

David A. Brent, M.D., Professor, Psychiatry

Dissertation Director: Satish Iyengar, Ph.D., Professor, Statistics

A STUDY OF TREATMENT-BY-SITE INTERACTION IN MULTISITE CLINICAL TRIALS

Kaleab Zenebe Abebe, PhD

University of Pittsburgh, June 25, 2009

Currently, there is little discussion about methods to explain treatment-by-site interaction in multisite clinical trials, so investigators are left to explain these differences post-hoc with no formal statistical tests in the literature. Using mediated moderation techniques, three significance tests used to detect mediation are extended to the multisite setting. Explicit power functions are derived and compared.

In the two-site case, the mediated moderation framework is utilized to test two difference-in-coefficients and one product-of-coefficients type tests. The test in the latter group is based on the product of two independent standard normal variables, which is a modified Bessel function of the second kind. Because the alternative distribution does not have a closed form expression, power is approximated using Gauss-Hermite quadrature. This test suffers from an inflated type I error, so two modifications are proposed: a combination of intersection-union and union-intersection tests; and one based on a variance stabilizing transformation. In addition, a modification of one of the difference-in-coefficients tests is proposed.

The tests are also extended to deal with multiple sites in the ANOVA and logistic regression models, and the groundwork has been laid to account for multiple mediators as well.

The contribution of this is a group of formal significance tests for explaining treatment-by-site interaction in the multisite clinical trial setting. This will serve to inform the design of future clinical trials by accounting for this site-level variability. The proposed methodology is illustrated in the analysis of the Treatment of SSRI-Resistant Depression in Adolescents

study conducted across six sites coordinated at the University of Pittsburgh.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

I am indebted to Dr. Iyengar for being a wonderful advisor and mentor throughout my graduate career. I would like to thank Dr. Gleser for his constructive criticism and feedback in his courses as well as this dissertation. I've also appreciated his many stories. To Dr. Sampson, thank you for believing in me and for pushing me to always do better. I would also like to thank Dr. Brent for allowing me to be apart of the TORDIA group.

The department of statistics has been a wonderful "family" to me over the past five years. I want to thank Mary and Kim for all their hard work. Without them, the department wouldn't function. Appreciation goes out to all of the students I've known over the years. To Ghideon and Scott, thanks for being the "older, wiser" students that I could solicit advise from.

To my family, thanks for all of your encouragement, love, and unending support. To the Becks, thanks for always having a warm fireplace to sit by and for always having enough food to eat. To Teshome, Solomon, Tamene, Assege, Surafel, Yetu, and Nitsuh, you have all been outstanding role models for me growing up. To my brother, thanks for always being a best friend. To my mother and father, I can't begin to express my gratitude and appreciation for you both. You laid the foundation for all of this to be possible, and for that, I'm eternally grateful.

Finally to my wife, Alyssa. Your patience, understanding, and constant support (financially and otherwise) of what was only supposed to be a two year master's program is beyond appreciated. I can't begin to put into words what that means to me.

1.0 INTRODUCTION

1.1 STATEMENT OF THE PROBLEM

In power and sample size calculations for randomized clinical trials, the current process is fairly straightforward. The investigators from different academic and/or industrial sites come together and agree on a common treatment protocol. They identify an effect size of the treatment of interest and specify type I and II errors (and therefore, power) *a priori*. The investigators then turn to their favorite sample size calculator (or favorite statistician) to obtain the overall sample size needed (N). Of interest is whether or not N can be obtained at one particular site, or if several sites are needed. Usually institutional affiliations or previous collaborations dictate how many sites can be recruited to participate, rather than explicit methodological considerations. As more sites are involved in the clinical trial, the inherent differences among them can build up and take a toll on the power to detect treatment effects. As a result, a treatment-by-site interaction can appear in the analysis stage of the clinical trial, which can temper the true effect of treatment. Investigators are left to discern the differences post-hoc.

This issue was recently investigated by Vierron and Giraudeau who incorporated a pre-specified intraclass correlation (ICC) into the typical sample size equation for a two-way mixed ANOVA model without interaction [49] . They found that for a fixed overall sample size, as the ICC and the number of sites varied, the estimated power did not deviate too much from the nominal power of 0.80. Their recommendation was to avoid recruiting a large number of sites relative to the overall sample size due to costs associated with more sites.

If the number of patients per site does not cause the power to decrease, then what does? The intent of this dissertation is to identify those sources of site heterogeneity that do, as well

as their impact on estimates of treatment effect. Among others, an example is a therapist at a particular site delivering cognitive behavior therapy in a different manner than a therapist at another site. By identifying sources at the design stage of a clinical trial, differences leading to a treatment-by-site interaction can hopefully be minimized. Yet, in order to identify said sources, one must have an idea of how to tackle the issue of treatment-by-site interaction at the analysis stage. This will inform the designs of future multisite clinical trials.

1.2 MOTIVATION

The motivation behind this research proposal stems from the meta-analysis done by Bridge et al. [8] that weighed the efficacy and risk of suicidal ideation in children and adolescents taking antidepressants. The study synthesized the results from 27 placebo-controlled trials of antidepressants in subjects suffering from major depressive disorder, obsessive compulsive disorder, or non-OCD anxiety. Studies ranged in size from single-site trials of 40 subjects to multisite trials of 396 subjects, with the maximum number of sites being 59.

By using the DerSimonian-Laird random effects model, pooled risk differences in response were obtained for each of the three disorders. Although there was an increased risk in suicidal thoughts across all disorders, the risk differences within each disorder group were not statistically significant. The conclusion was that the benefits outweighed the risks in each of the three disorders.

Upon examination of the potential moderators of clinical response, the authors found that the estimated effect size decreased as the number of sites in each study increased. This finding tempers one of the principal advantages of a multisite clinical trial (as opposed to a single site) which is that larger sample sizes result in higher power. Dr. David Brent, of the Department of Psychiatry, posed the problem of understanding

> "...the impact of increasing number of sites on power taking into account the increase in 'noise' due to differences in assessment and treatment procedure."

This gave rise to the question of whether explicit sources of site heterogeneity could be identified at the design stage (such as outcome reliability, patient severity, patient characteristics)

as well as their quantitative impact on the degradation of treatment effect.

In the next chapter, we give a summary of multisite clinical trials, including commonly used methods of analysis. Also, the two primary estimators of treatment effect, weighted and unweighted, are compared and contrasted in the context of several scenarios that occur in multisite trials: disparate sample sizes across sites and the presence of treatment-by-site interaction. Finally, hierarchical linear models are introduced and their properties regarding sources of site heterogeneity are discussed.

In Chapter 3, the relationship between treatment effect size and the number of sites is investigated as well as the identification of potential sources of site heterogeneity from the Treatment of SSRI-Resistant Depression in Adolescents (TORDIA) multisite clinical trial.

In Chapter 4, several statistical methods to identify sources in regression and ANOVA models are shown using an idea called "mediated moderation" applied to the multisite clinical trial setting. Three significance tests popular in investigating mediation were extended. For each, their respective test statistics are described and power analyses are performed. In addition, the tests are illustrated on the TORDIA dataset.

Multisite mediated moderation (MMM) is extended to the logistic regression models in Chapter 5. A simulation study to estimate power is conducted, and the significance tests are applied on the TORDIA dataset.

Finally, the last chapter presents a discussion and lays down the foundation for future work.

2.0 LITERATURE REVIEW

2.1 MULTISITE CLINICAL TRIALS

2.1.1 Introduction

Meinert defines a multi-center clinical trial as one that has at least two clinics (or centers), a common treatment protocol, and a centralized unit to receive and process the study data [38]. Multisite trials are preferred to their single site counterparts for several reasons, Kraemer suggests [27]. First, the multisite trial has the ability to recruit many more subjects than the single site trial, resulting in higher power. The time it could take for a single site trial to accrue the same amount of patients as a corresponding multisite trial could be substantially longer.

The second advantage is generalizability. Several single site trials can be designed to address the same question yet yield varying results. This may be due to different patient characteristics in certain geographic regions or substantially different treatment protocols between sites. On the other hand, bringing those different patient populations together in one multisite trial makes it easier to study treatment effects on patients in general.

Finally, multisite trials have the ability to bring together experts with widely varying viewpoints concerning the treatment protocol. In single site trials, centers that tend toward a particular philosophy may have results that are affected by that philosophy. For example, if a single-site psychiatric trial for treatment of depression is based at an academic center that adheres to the use of selective serotonin re-uptake inhibitors (as opposed to cognitive behavioral therapy), then the resulting effect of treatment may shortchange cognitive behavioral therapy.

2.1.2 Model / Analysis

Whereas single-site trials can focus on a single treatment effect, multisite trials have the added difficulty of dealing with site effects. Despite the fact that the sites are expected to follow a common protocol, their estimated treatment effects are not guaranteed to be similar. This is due to site heterogeneity, as will be explained in detail in the next chapter.

The classic analytic approach in multisite studies analysis is to include in the model an effect for site. For instance, a fixed effects model for comparing a response Y_{ijk} between two treatments across J sites is

$$Y_{ijk} = \mu + \tau_i + \varsigma_j + \epsilon_{ijk} \tag{2.1}$$

where $i = 1, 2$, $j = 1, ..., J$, and $k = 1, ..., n_{ij}$. The usual model constraint is that $\sum_{i=1}^{2} \tau_i = \sum_{j=1}^{J} \varsigma_j = 0$. The treatment, τ_i, and site main effects, ς_j, are nonrandom and the replication errors, ϵ_{ijk}, are i.i.d. $N(0, \sigma^2)$ variates. The model assumes that the effect of treatment is constant across sites [17]. Because each of the sites is expected to follow the common study protocol and accrue patients independently, the assumption of no interaction is a desirable one in multisite clinical trials. In fact, the International Conference on Harmonisation (ICH) E9 guideline on statistical principles strongly recommends the non-interaction model, (2.1), for analysis [23]. With regard to this, Gallo (2000) states: "Rarely is a trial undertaken with a clear expectation regarding the nature of different effects expected in different centers." [17]

On the other hand, due to differences in underlying patient populations as well as subtle protocol deviations, the treatment effects can easily differ across sites. Under this assumption, the above model is modified in the following way:

$$Y_{ijk} = \mu_{ij} + \epsilon_{ijk} = \mu + \tau_i + \varsigma_j + \gamma_{ij} + \epsilon_{ijk} \tag{2.2}$$

where γ_{ij} is a fixed effect. The usual constraints in addition to that of (2.1) are $\sum_{i=1}^{2} \gamma_{ij} = \sum_{j=1}^{J} \gamma_{ij} = 0$.

2.1.3 Interaction

The presence of treatment-by-site interaction makes it difficult to interpret the main effect of treatment. Even trying to detect the phenomenon is difficult because tests for interaction typically have low power [27, 14, 33, 45, 17, 46, 51]. The reason for this is that most trials have power only to detect main effects, such as treatment effect, but adequate power to detect an interaction requires a much larger sample size [10]. Due to this lack of power, falsely rejecting the hypothesis of no interaction, risks having biased estimates of main effects [14]. A common approach used by statisticians, although not optimal, is to use model (2.2) and remove the interaction term when a significance test for interaction results in a p-value larger than .1 or .2 [45, 17, 51].

2.1.4 Treatment Effect

After choosing the type of model for multisite studies, the most important step is examining the true effect of treatment, since evaluating this effect is the main reason for conducting the trial in the first place. For simplicity, suppose that each site has the same number of subjects taking each of the two treatments ($n_{1j} = n_{2j} = n_j$; $n_1 = ... = n_j$). The true treatment effect (or treatment difference) at a particular site j is $\delta_j = \mu_{1j} - \mu_{2j}$ is estimated by $\bar{d}_j = \bar{Y}_{1j.} - \bar{Y}_{2j.}$ [48]. Then, $\delta = \dfrac{\sum_{j=1}^{J} \delta_j}{J}$ is the true average treatment effect across sites and is estimated by $\bar{d} = \dfrac{\sum_{j=1}^{J} \bar{d}_j}{J} = \bar{Y}_{1..} - \bar{Y}_{2..}$. The interpretation of this is quite straightforward.

Although it is simple, the completely balanced case shown above is unrealistic. Since randomization in multisite trials is done at the site level, it is not uncommon to have nearly identical sample sizes across treatments, but not across sites. Having unequal number of subjects per site is more typical in multisite trials [46].

In the case of unequal sample sizes, the use of the above estimator, $\bar{d}$, raises the question of whether larger sites add more to the effect of treatment despite being weighted the same as much smaller sites. Also, in the presence of sample size disparities and/or treatment-by-site interaction, what type of estimator is most easily interpretable?

Two estimators, introduced by Fleiss (1985), have led to much discussion about how

best to estimate the overall effect of treatment [14]. The weighted (or type II) estimator is defined as follows:

$$\bar{d}_w = \frac{\sum_{j=1}^{J} w_j \bar{d}_j}{\sum_{j=1}^{J} w_j}, \tag{2.3}$$

where $\bar{d}_j$ is described above and $w_j = \frac{n_{1j} n_{2j}}{n_{1j}+n_{2j}}$ are the weights at each site j. Each weight function is the harmonic mean of the sample sizes associated with a particular site j. Also, the weights are inversely proportional to the variance of the response variable, so larger sites get larger weights. The interpretability of $\bar{d}_w$ is clear unless there is treatment-by-site interaction. This is well illustrated by showing that the underlying parameter estimated by $\bar{d}_w$ differs under models (2.1) and (2.2). Under the full model (2.2),

$$E(\bar{d}_w) = \frac{\sum_{j=1}^{J} w_j E(\bar{d}_w)}{\sum_{j=1}^{J} w_j} = \frac{\sum_{j=1}^{J} w_j E(\bar{Y}_{1j.} - \bar{Y}_{2j.})}{\sum_{j=1}^{J} w_j}. \tag{2.4}$$

Since $E(\bar{Y}_{ij.}) = \mu_{ij} = \mu_{..} + \tau_i + \varsigma_j + \gamma_{ij}$, we have

$$E(\bar{d}_w) = \frac{\sum_{j=1}^{J} w_j \left(\tau_1 + \gamma_{1j} - \tau_2 - \gamma_{2j}\right)}{\sum_{j=1}^{J} w_j} = (\tau_1 - \tau_2) + \frac{\sum_{j=1}^{J} w_j \left(\gamma_{1j} - \gamma_{2j}\right)}{\sum_{j=1}^{J} w_j}. \tag{2.5}$$

As evident from above, the estimate $\bar{d}_w$ is unbiased for the true treatment difference when the treatment-by-site interaction is absent, or when the sample size weights are identical at each site.

The unweighted (or type III) estimator is:

$$\bar{d}_u = \frac{\sum_{j=1}^{J} \bar{d}_j}{J}, \tag{2.6}$$

where $\bar{d}_j$ is as before. In this case, all sites get equal weight, regardless of sample sizes. Since $\bar{d}_u$ is just the unweighted average across all sites, it is always interpretable – even in the presence of interaction – because

$$E(\bar{d}_u) = \frac{\sum_{j=1}^{J} E(\bar{d}_u)}{J} = \frac{\sum_{j=1}^{J} (\tau_1 + \gamma_{1j} - \tau_2 - \gamma_{2j})}{J} = \tau_1 - \tau_2, \tag{2.7}$$

under the usual model restrictions that require $\sum_{i=1}^{I} \gamma_{ij} = \sum_{j=1}^{J} \gamma_{ij} = 0$.

Several authors have attempted to tackle the issue of which estimator to use in which cases, namely disparate sample sizes across sites and treatment-by-site interaction. First, for unequal sample sizes, the consensus is that the weighted estimator is superior to the unweighted in the sense that the variance of $\bar{d}_u$ is always at least as big as the variance of $\bar{d}_w$ [33, 45, 17, 46].

A commonly proposed solution to the problem of disparate sample sizes is to pool smaller centers into a larger one (usually until a maximum sample size per center is reached), which is explained in detail in Lin, Gallo, and Worthington [33, 17, 51]. This has been met with criticism by some authors who argue that it results in loss of power as well as the potential introduction of bias (especially in the unweighted case) [33, 17]. When combining small centers, the assumption that they are similar to each other is not guaranteed.

Secondly, others say that treatment-by-site interaction eliminates $\bar{d}_w$ as a possibility due to the lack of interpretability described above [45, 46]. On the other hand, Gallo claimed that as long as there was no systematic relationship between site sample size and within-site effect, the parameters that $\bar{d}_w$ and $\bar{d}_u$ estimate are identical [17].

2.1.5 Fixed versus Random Sites

The issue of whether sites should be modeled as fixed or random effects is an intriguing one that deserves discussion. According to the textbook definition [31], a factor should be considered random if interest in its effect on the response variable extends beyond the factor levels used in the analysis. On the other hand, if the factor levels are the only levels of interest, then the factor is clearly fixed. Some factors, such as gender, are inherently

fixed. However, there seems to be agreement that the effect of site should be considered fixed [46, 51]. Among the reasons given are that sites are not usually chosen in a "random" manner, but rather based on previous collaborations. For example, academic institutions that have worked together on previous clinical trials usually develop good relationships which facilitate finding sites for future studies.

In the case of random site effects, (2.1) and (2.2) are modified in the following way:

$$Y_{ijk} = \mu + \tau_i + S_j + \epsilon_{ijk} \tag{2.8}$$

and

$$Y_{ijk} = \mu + \tau_i + S_j + G_{ij} + \epsilon_{ijk}, \tag{2.9}$$

where $S_j \overset{iid}{\sim} N(0, \sigma_S^2)$ and mutually independent of $G_{ij} \overset{iid}{\sim} N(0, \sigma_G^2)$.

Senn (1998) gives a detailed overview of both sides of the random versus fixed argument [46]. Some advantages of the fixed approach include the following. First, there is better precision of the estimate of effect because the variance is smaller. Second, it is the only realistic option in the presence of very few sites. An example of this would be studies of very rare diseases, where it may be that only a handful of sites specialize in it. Third, to regard sites as random is unrealistic due to the fact that actual random sampling rarely occurs.

In defense of the random approach, the purpose of developing treatments is to say something about their effects on patients in general. By adding the site variability, the scope of prediction is broadened to include patients from different geographic regions. Second, if interest is about a given site, the fixed approach leaves little alternative but to use the results from that site only.

Another interesting point is whether it is appropriate to assume that random effects follow a normal distribution. If the underlying distribution of the effect is highly skewed, then bias may occur in the estimation of the effect [2]. Ways to remedy this include assuming a skew-normal or skew-t distribution, which is beyond the scope of this proposal but is discussed in Azzalini and Capitanio [4, 5].

2.1.6 Hierarchical Linear Models

There are a number of data structures that are naturally hierarchical in their organization. In educational studies, students may be nested within a class with the same teacher, the teachers in turn nested in particular schools, and so on. In longitudinal studies, a subject's measurements over time are nested within that subject. Multisite clinical trials are no exception to this. Because randomization is conducted at the level of the individual sites, patients are expected to be more closely related to those within their site.

Raudenbush and Bryk give an account of the theory and applications of hierarchical linear models (HLMs) [41]. which are able to model at the level of the study sites as well as the level of the subjects within them. Three main features of HLMs that the authors emphasize are as follows. First, hierarchical linear models allow improved estimation of the effects at the subject-level. This is due, in part, to the fact that individual sites have their own separate regression equations that "borrow strength" from other sites with similar estimates. Second, besides investigating effects at a particular level, HLMs allow the examination of effects across levels. In multisite trials, this can be likened to the effect of treatment across sites, or treatment-by-site interaction. Third, the use of variance-covariance components facilitates estimation in unbalanced designs.

Examples of HLMs are one-way random effects AN(C)OVA, means-as-outcomes regression, random coefficients regression, and coefficients-as-outcomes regression. The rest of this section will restrict its attention to the latter two.

The random coefficients model is set up as follows [42]. The subject-level model is

$$Y_{jk} = \beta_{0j} + \beta_{1j} X_{jk} + r_{jk} \tag{2.10}$$

where $j = 1, ..., J$, $k = 1, ..., n_j$, and r_{jk} are i.i.d. $N(0, \sigma^2)$. Notice that the subscript i has been suppressed. X_{jk} is a treatment contrast for subject k in site j taking a value of 1 for treatment and -1 for control subjects. The intercept β_{0j} represents the mean response for site j, while the slope β_{1j} is the effect due to a subject's particular treatment. The site-level model is

$$\beta_{0j} = \alpha_{00} + \alpha_{0j} \tag{2.11}$$

$$\beta_{1j} = \alpha_{10} + \alpha_{1j} \tag{2.12}$$

where α_{00} and α_{10} are the grand mean and average treatment effect, respectively. α_{0j} and α_{1j} are random effects, independent of r_{jk}, that are distributed as

$$\begin{pmatrix} \alpha_{0j} \\ \alpha_{1j} \end{pmatrix} \overset{iid}{\sim} N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} \eta_{00} & \eta_{01} \\ \eta_{01} & \eta_{11} \end{bmatrix} \right).$$

As can be seen above, both the intercept and slope have their own respective random effects which account for the variability in the mean response and mean treatment effect across sites. In addition, the parameter η_{01} denotes the covariance between the mean and the treatment effect at a particular site. When (2.10), (2.11), and (2.12) are combined, the resulting full model is

$$Y_{jk} = \alpha_{00} + \alpha_{10}X_{jk} + \alpha_{0j} + \alpha_{1j}X_{jk} + r_{jk}. \tag{2.13}$$

The variance of a particular observation is

$$\text{Var}(Y_{jk}) = \sigma_Y^2 = \eta_{00} + \eta_{11} + 2X_{jk}\eta_{01} + \sigma^2, \tag{2.14}$$

which depends on the treatment. The covariance between two different observations in the same site and taking the same treatment is denoted by

$$\text{Cov}(Y_{jk}, Y_{jk'}) = \sigma_{kk'} = \eta_{00} + X_{jk}X_{jk'}\eta_{11} = \eta_{00} + \eta_{11}. \tag{2.15}$$

For two different observations that also differ in treatment (denoted by the missing subscript i), (2.15) becomes

$$\text{Cov}(Y_{jk}, Y_{jk'}) = \sigma_{ii'} = \eta_{00} - \eta_{11}. \tag{2.16}$$

Finally, the covariance between two observations neither from the same site nor taking the same treatment is

$$\text{Cov}(Y_{j'k}, Y_{jk'}) = 0. \tag{2.17}$$

By adding a subscript for treatment ($i = 1, 2$), (2.13) is nearly identical to the full mixed-effects model, (2.9).

$$Y_{ijk} = \alpha_{00} + \alpha_{10i} + \alpha_{0j} + \alpha_{1ij} + r_{ijk}. \tag{2.18}$$

The main difference between the two models is that in (2.18), α_{0j} and α_{1ij} are independent of each other for a particular treatment, i.

The coefficients-as-outcomes model takes one step further and adds site-level predictors, W_j, to model said variability. The site-level equations become

$$\beta_{0j} = \alpha_{00} + \alpha_{01} W_j + \alpha_{0j} \tag{2.19}$$

$$\beta_{1j} = \alpha_{10} + \alpha_{11} W_j + \alpha_{1j}. \tag{2.20}$$

One important thing to notice is the fact that the predictor is used in both the intercept and slope models. Raudenbush and Liu point out that due to the correlated nature of the errors, α_{0j} and α_{1j}, failure to specify the errors in the intercept model may lead to an inaccurate estimate of the predictor's contribution to the treatment effect. The authors also note that large treatment-by-site variance, η^2_{11}, signals the need to add site-level predictors to model the extra variation [42].

There are three main issues in multilevel analysis that HLMs deal with: 1) aggregation bias, 2) misestimation of standard errors, and 3) heterogeneity of regression [41]. The first issue occurs when the relationship between variables takes on different meanings at different levels. For example, two variables may be highly correlated in one direction among an unnested group of subjects. When nesting occurs and mean values of the variables are computed, the correlation could be in the opposite direction or diminished entirely. This is also known as the ecological fallacy [44]. Hierarchical models address this by decomposing the full model to show relationships within and across all levels.

Another common problem in multilevel analysis is not taking into account the correlated (dependent) nature of the grouped observations. By including a random effect for each site, HLMs use this added information when estimating the standard errors [42].

The third problem, heterogeneity of regression, exists when relationships between predictors and outcomes vary across sites. This is dealt with by estimating separate sets of regression coefficients for each site. The variability at the second level can then be accounted for by other predictors [42].

3.0 SOURCES OF SITE HETEROGENEITY

The previous chapter outlined some of the common methods of analyzing a multisite clinical trial as well as the issues statisticians are likely to face while doing so. The purpose of Chapter 3 is to investigate the effect that the number of sites chosen has on the effect of treatment. If there is no relationship, then there are other sources at work that need to be identified.

3.1 RELATIONSHIP BETWEEN TREATMENT EFFECT SIZE AND THE NUMBER OF SITES

In investigating the relationship between effect size of treatment and the number of sites, 4 different clinical trials were simulated with 200 participants each. The trials varied from one another by the number of sites each had, ranging from 2 sites of 100 each to 10 sites of 20 participants. Within each site, the number of treatment and control subjects were equal, making the trials completely balanced ($n_{1j} = n_{2j} = n$ for all sites, $j = 1, ..., J$).

In the initial simulation, we considered a continuous outcome variable, Y_{ijk}, coming from the following distribution

$$Y_{ijk} \sim N(\mu_{..} + \tau_i + \varsigma_j, \sigma^2), \tag{3.1}$$

where μ, τ_i, and σ^2 denote the overall mean, treatment effect, and error variance respectively. While μ was arbitrarily chosen to be 0, the values of τ and σ were chosen such that they would yield a treatment effect size of 0.4, which was close to the weighted average of effect sizes in the Bridge et al. meta-analysis [8]. The site effects $\varsigma_1, ..., \varsigma_J$ were generated from

a $N(0, 0.5)$ distribution subject to the constraint $\sum_{j=1}^{J} \varsigma_j = 0$. This was accomplished by obtaining $\varsigma_1, ..., \varsigma_{J-1}$ and then setting $\varsigma_J = \sum_{j=1}^{J-1} -\varsigma_j$. The 0.5 variance was arbitrarily chosen to represent noise.

The simulated data were then fit using the following one-way ANOVA comparing treatment and placebo:

$$Y_{ik} = \mu_{..} + \tau_i + \epsilon_{ik}. \tag{3.2}$$

For each site number iteration, the mean of 1000 effect sizes was computed using Hedges's g:

$$\widehat{g} = \frac{\bar{Y}_{1..} - \bar{Y}_{2..}}{\sqrt{(MSE)}}, \tag{3.3}$$

where the pooled standard deviation is the square root of the mean square error obtained from the above model:

$$MSE = \frac{\sum_{i=1}^{2}\sum_{j=1}^{J}\sum_{k=1}^{n}(Y_{ijk} - \bar{Y}_{i..})^2}{2Jn - 2} = \frac{\sum_{i=1}^{2}(Jn-1)s_i^2}{2Jn-2}. \tag{3.4}$$

Because the sample sizes for each treatment-site combination are equal, (3.4) reduces to

$$MSE = \frac{s_1^2 + s_2^2}{2}, \tag{3.5}$$

which is nothing more than the average variance across treatment and placebo.

Regarding the MSE, Kraemer et al. states that the power to detect treatment effects is inversely related to the number of sites, for a fixed sample size N [29]. This is only the case when the MSE from a two-way ANOVA with treatment and site (and possibly, their interaction) is used. This would be a misspecification because the standard deviation should be that of the raw outcome scores [37, 19]. For instance McGaw and Glass give an example of a study comparing treatment and control conditions (factor A) across gender (factor B) [37]. The model is a two-way ANOVA with both main effects as well as their interaction

(factor AB). To calculate the appropriate effect size in this example, the denominator should be

$$\hat{\sigma} = \sqrt{\frac{SSB + SSAB + SSE}{df_B + df_{AB} + df_E}}. \tag{3.6}$$

Because $SSB + SSAB + SSE$ and $df_B + df_{AB} + df_E$ are equivalent to the SSE and df_E, respectively, from a balanced one-way ANOVA with only an effect for treatment suggests that the denominator used in (3.3) is correct.

Returning to the simulation, the mean effect size was then plotted against the number of sites, which did not reveal any relationship between the two. Since the MSE is distributed as $\frac{\sigma^2}{2Jn-2}\chi^2_{2Jn-2}$ and $\bar{Y}_{1..} - \bar{Y}_{2..}$ is independent of $\sqrt{MSE}$, (3.3) can be rewritten as

$$E(\widehat{g}) = E\left(\frac{\bar{Y}_{1..} - \bar{Y}_{2..}}{\sigma}\sqrt{\frac{\sigma^2}{MSE}}\right) = \frac{\mu_1 - \mu_2}{\sigma} E\left(\sqrt{\frac{df}{\chi^2_{df}}}\right), \tag{3.7}$$

with degrees of freedom $df = 2Jn - 2$. It can be shown that the expectation of the random variable is equivalent to $C(p)$, where

$$C(p) = \frac{\sqrt{p}\Gamma\left(\frac{p-1}{2}\right)}{\sqrt{2}\Gamma\left(\frac{p}{2}\right)}. \tag{3.8}$$

In fact, Hedges and Olkin showed that the estimator for Hedges's g was upwardly biased without the correction factor $C(p)$ [20, 11]. So, (3.7) reduces to

$$E(\widehat{g}) = \frac{\mu_1 - \mu_2}{\sigma} C(df). \tag{3.9}$$

While this is indeed a function of the degrees of freedom (and hence, the number of sites J), it is a constant because the error degrees of freedom do not change because the overall sample size is $N = 2Jn = 200$. Returning to Kraemer's statement in the previous paragraph, using the MSE from the two-way ANOVA changes df from $N-2$ to $N-2J$. So in this case, $C(df)$ is in fact a decreasing function of the number of sites, although the decrease is very slow.

This analytical result suggests that there is nothing inherently heterogeneous about the number of sites chosen. Rather, specific sources of site heterogeneity are solely responsible for the degradation of treatment effect.

3.2 TREATMENT OF SSRI-RESISTANT DEPRESSION IN ADOLESCENTS

As was shown in the previous section, there is nothing inherently heterogeneous about the number of sites chosen in multisite clinical trial. Therefore, there are other sources to blame. A motivating example is the following study.

The Treatment of SSRI-Resistant Depression In Adolescents (TORDIA) clinical trial was an NIMH funded multisite study that sought to evaluate the efficacy of four treatment strategies in depressed youths [7]. A sample of 334 patients, across 6 sites, with major depressive disorder who were not responding to an initial 2-month selective serotonin re-uptake inhibitor (SSRI) treatment were randomized to one of the following regimens

1. switch to a second, different SSRI (paroxetine, citalopram, or fluoxetine),
2. switch to a different SSRI plus cognitive behavior therapy (CBT),
3. switch to venlafaxine, or
4. switch to venlafaxine plus CBT.

The primary dichotomous outcome, clinical response, was defined as the combination of the Clinical Global Impressions score ≤ 2 and a change in the Children's Depression Rating Scale-Revised of $\geq 50\%$. CBT plus medication (CBT-MED) showed a higher rate of response than medication (MED) alone (54.8% vs. 40.5%, p=0.009). On the other hand, there were no differences between medication response rates regardless of CBT use (48.2% vs. 47.0%, p=0.83). In addition, the effect of CBT-MED treatment versus MED alone was heterogeneous across the sites, ranging from a 35.3% difference favoring MED to a 45.0% difference favoring CBT-MED.

This finding led to a subsequent paper by Spirito et al. which investigated potential causes for this treatment-by-site interaction [47]. The paper discussed two potential causes of site heterogeneity: participant clinical characteristics, and treatment protocol consistency. The process of identifying these causes was straightforward. The authors first examined whether particular variables were related to site. If found to be significant, those variables were examined to see if they were significant predictors of the response variable.

In the case of clinical characteristics, the stratification variable that measured suicidality (BDI item 9) differed across sites as well as being significantly related to outcome. In addition, three of the variables measured at baseline that were significant across sites were also related to outcome: duration of depression, hopelessness, and family conflict. This implies that baseline clinical characteristics are a potential source of site heterogeneity. A recursive partitioning method was then used to determine optimal subgroups where site variability was minimal [26]. Subjects with low hopelessness and low family conflict comprised the optimal subgroup, which had a clinical response rate of 67.8% compared with 47.6% from the original 334 subjects. In addition, the treatment-by-site interaction from above didn't exist in this subgroup. Conversely, subjects with high family conflict and high hopelessness scores had a clinical response rate of 37.0%. Finally, there was a significant imbalance of the number of subjects from the optimal subgroup across the six sites in the MED-only group. This imbalance was not present in the CBT-MED group.

With regard to treatment protocol consistency, several of the variables included fidelity to treatment, ancillary pharmacotherapy and protocol attrition. Treatment fidelity refers to either fidelity to therapy or fidelity to medication. The former was measured by the Cognitive Therapy Rating Scale (CTRS), which differed across sites but was not related to outcome. The same held for the Pharmacotherapy Rating Scale (PTRS), which measured medication therapy.

4.0 IDENTIFYING SOURCES OF SITE HETEROGENEITY

4.1 BACKGROUND

As was mentioned in the introduction, when the effect of treatment differs across sites (i.e. treatment-by-site interaction), the investigators are left to explain the differences post-hoc. Currently, there is little discussion of the impact of these differences on clinical trials [47]. In addition, the only publication to date that attempts to explain the methodology of identifying sources of site heterogeneity is Spirito et al [47].

The process of identification of site differences involves two phenomena: moderation and mediation (see Kraemer et al. [28], Aiken & West [3], and Baron and Kenny [6]). A moderator is a baseline variable, uncorrelated with treatment, that identifies subgroups of patients who have different effect sizes [30]. In the case of a continuous moderator, the effect size of a particular treatment is a function of the moderator variable. For example, consider the following equation where there are two levels of treatment, $\tau = \pm 1$; an outcome, y; and a moderator variable, mo:

$$y = \beta_0 + \beta_1 \tau + \beta_2 mo + \beta_3 (\tau * mo) + \epsilon. \tag{4.1}$$

Showing that the coefficient for the interaction term is significant is sufficient to demonstrate moderation. On the other hand, there are a set of causal relationships that must be shown in order to demonstrate mediation. First, the treatment must significantly predict the outcome. Second, a mediator must be significantly predicted by treatment. Third, it must significantly predict outcome after accounting for treatment. The result is a change in the relationship between the treatment and outcome [6]. Consider the basic case where there are two levels

of treatment, $\tau = \pm 1$; an outcome, y; and a mediator, m. There are 3 equations involved in the testing of mediation in this scenario:

$$y = \beta_{00} + \beta_{01}\tau + \epsilon_0 \tag{4.2}$$

$$m = \beta_{10} + \beta_{11}\tau + \epsilon_1 \tag{4.3}$$

$$y = \alpha_{00} + \alpha_{01}\tau + \alpha_{12}m + \epsilon_2 \tag{4.4}$$

where the ϵ_i are Gaussian with mean 0, and variance σ_i^2. Equation 4.2 shows the direct effect of the treatment on the response (β_{01}). Equation 4.3 shows the relationship between the mediator and treatment (β_{11}). Finally, the residual effect of treatment on response, after accounting for the effect of the mediator, is shown in equation (α_{01}) 4.4. This is shown graphically in Figure 1 as a path diagram.

These relationships can be written in terms of the joint distributions of y and m.

$$\begin{pmatrix} y \\ m \end{pmatrix} \sim \mathrm{N}\left\{ \begin{pmatrix} \beta_{00} + \beta_{01}\tau \\ \beta_{10} + \beta_{11}\tau \end{pmatrix}, \begin{bmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{01} & \sigma_1^2 \end{bmatrix} \right\}$$

The final equation is based on the conditional distribution of y given τ and m:

$$(y|m,\tau) \sim \mathrm{N}\left(\beta_{00} + \beta_{01}\tau + \frac{\sigma_{01}}{\sigma_1^2}\left(m - \beta_{10} - \beta_{11}\tau\right), \sigma_0^2 - \frac{\sigma_{01}^2}{\sigma_1^2}\right),$$

and the conditional mean can be rewritten as

$$\begin{aligned} \mathrm{E}(y|m,\tau) &= \left(\beta_{00} - \frac{\sigma_{01}}{\sigma_1^2}\beta_{10}\right) + \left(\beta_{01} - \frac{\sigma_{01}}{\sigma_1^2}\beta_{11}\right)\tau + \frac{\sigma_{01}}{\sigma_1^2}m \\ &= \alpha_{00} + \alpha_{01}\tau + \alpha_{12}m. \end{aligned} \tag{4.5}$$

Of course there are redundancies, so the parameters are constrained. For instance,

$$\beta_{01} - \alpha_{01} = \frac{\sigma_{01}}{\sigma_1^2}\beta_{11} = \alpha_{12}\beta_{11}. \tag{4.6}$$

In the context of the path diagram, (4.6) says that the difference between the direct and residual effect of treatment on response is equivalent to the product of the two indirect paths. The significance test for mediation, according to Baron & Kenny [6] is made up of three separate tests conducted in succession:

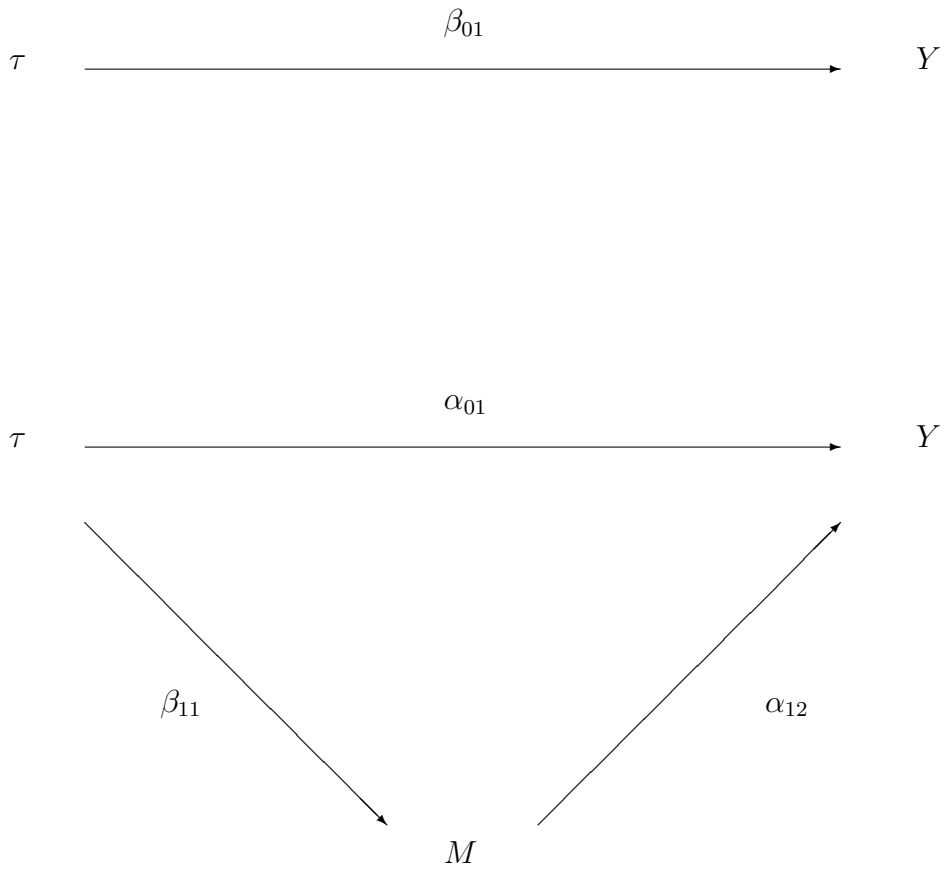

Figure 1: Mediation Path Diagram

- $H_0 : \beta_{01} = 0$ with test statistic $t^* = \dfrac{b_{01}}{\hat{\sigma}_{b_{01}}} \sim t_{n-2}$,
- $H_0 : \beta_{11} = 0$ with test statistic $t^* = \dfrac{b_{11}}{\hat{\sigma}_{b_{11}}} \sim t_{n-2}$,
- $H_0 : \alpha_{12} = 0$ with test statistic $t^* = \dfrac{a_{12}}{\hat{\sigma}_{a_{12}}} \sim t_{n-3}$,

where n is the sample size. So, if each of the above hypothesis tests result in significance, then mediation exists. [6].

While the testing of moderation is relatively straightforward [28], there has been much discussion on how best to test mediation. MacKinnon et al. used simulation to compare 14 different significance tests (including Baron and Kenny's approach above) for mediation grouped into three types: causal steps, difference-in-coefficients, and product-of-coefficients [35]. The authors concluded that widely used causal methods, such as those proposed by Judd and Kenny [25] and Baron and Kenny [6], had low power. The best causal steps test was the following significance tests proposed by MacKinnon et al.:

- $H_0 : \beta_{11} = 0$ with test statistic $t_b^* = \dfrac{b_{11}}{\hat{\sigma}_{b_{11}}} \sim t_{n-2}$,
- $H_0 : \alpha_{12} = 0$ with test statistics $t_a^* = \dfrac{a_{12}}{\hat{\sigma}_{a_{12}}} \sim t_{n-3}$,

where n is the sample size. In the difference-in-coefficients group, the following test proposed by Freedman and Schatzkin had the greatest power and most accurate type I error: [16]

- $H_0 : \beta_{01} - \alpha_{01} = 0$ with test statistic
  $t^* = \dfrac{b_{01} - a_{01}}{\sqrt{\hat{\sigma}_{b_{01}}^2 + \hat{\sigma}_{a_{01}}^2 - 2\hat{\sigma}_{b_{01}}\hat{\sigma}_{a_{01}}\sqrt{1 - \hat{\rho}_{\tau m}^2}}} \sim t_{n-2}$

where $\hat{\rho}_{\tau m}^2$ is the correlation between the mediator and treatment. In the last group of tests, product-of-coefficients, a test introduced by MacKinnon et al. [35] was superior with regard to power:

- $H_0 : \beta_{11}\alpha_{12} = 0$ with test statistic $w^* = \dfrac{b_{11}}{\hat{\sigma}_{b_{11}}}\dfrac{a_{12}}{\hat{\sigma}_{a_{12}}}$, a product of two standard normals under H_0.

While the aforementioned gives good insight into how best to deal with moderation and mediation separately, site heterogeneity in multisite clinical trials involves dealing with both simultaneously. For instance, moderation involves differing treatment effects across sites. In other words, the effect of treatment differs as you move across the levels of site. Moreover,

site moderating treatment is equivalent to treatment moderating site (i.e. site effects differ by treatment regimen), so the two concepts are interchangeable.

On the other hand, mediation involves an outside variable influencing the relationship between an independent variable and the response. For example, exercise has been shown to be a significant predictor of glucose level in women who are at risk for diabetes [50]. However, after adjusting for BMI, the association is still significant yet reduced. The obvious reason is that women who exercise more tend to have a lower BMI, which is also related to glucose level. Here, BMI is mediating the relationship between glucose level and exercise. Therefore, returning to the context of multisite clinical trials, the main objective in identifying sources of site heterogeneity is to pinpoint particular variables that mediate the moderation of site.

Muller et al. [39] describes the methodology of "moderated mediation" and "mediated moderation", with the latter being of interest here. Mediated moderation occurs when an underlying mediation process is responsible for the overall moderation that exists; and by accounting for that process, the magnitude of moderation is reduced [39]. The previous basic mediation equations ((4.2), (4.3), and (4.4)) can be extended to a two-site scenario in the following way:

$$y = \beta_{00} + \beta_{01}\tau + \beta_{02}s + \beta_{03}(\tau * s) + \epsilon_0 \tag{4.7}$$

$$m = \beta_{10} + \beta_{11}\tau + \beta_{12}s + \beta_{13}(\tau * s) + \epsilon_1 \tag{4.8}$$

$$y = \alpha_{00} + \alpha_{01}\tau + \alpha_{02}s + \alpha_{03}(\tau * s) + \alpha_{12}m + \epsilon_2, \tag{4.9}$$

where $s = \pm 1$ is the effect of site. In terms of joint and conditional distributions,

$$\begin{pmatrix} y \\ m \end{pmatrix} \sim \mathrm{N}\left\{ \begin{pmatrix} \beta_{00} + \beta_{01}\tau + \beta_{02}s + \beta_{03}(\tau * s) \\ \beta_{10} + \beta_{11}\tau + \beta_{12}s + \beta_{13}(\tau * s) \end{pmatrix}, \begin{bmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{01} & \sigma_1^2 \end{bmatrix} \right\},$$

and

$$(y|m, \tau, s) \sim N\left(\mu_{y|m,\tau,s}, \sigma_0^2 - \frac{\sigma_{01}^2}{\sigma_1^2}\right),$$

where the conditional mean is

$$\begin{aligned}
\mu_{y|m,\tau,s} &= \beta_{00} + \beta_{01}\tau + \beta_{02}s + \beta_{03}(\tau * s) + \frac{\sigma_{01}}{\sigma_1^2}\left(m - \beta_{10} - \beta_{11}\tau - \beta_{12}s - \beta_{13}(\tau * s)\right) \\
&= \left[\beta_{00} - \frac{\sigma_{01}}{\sigma_1^2}\beta_{10} + \beta_{01}\tau - \frac{\sigma_{01}}{\sigma_1^2}\beta_{11}\tau\right] + \left[\beta_{02} - \frac{\sigma_{01}}{\sigma_1^2}\beta_{12} + \beta_{03}\tau - \frac{\sigma_{01}}{\sigma_1^2}\beta_{13}\tau\right] s \\
&\quad + \frac{\sigma_{01}}{\sigma_1^2}m \\
&= (\alpha_{00} + \alpha_{01}\tau) + (\alpha_{02} + \alpha_{03}\tau)s + \alpha_{12}m \qquad (4.10)
\end{aligned}$$

One thing to note is that Muller et al. allowed the partial effect of the mediator to be moderated, which added another term to (4.9) [39]. This is not a necessary condition for mediated moderation to hold, and we will assume that this partial effect is not moderated (see Appendix A for details). The direct effect of site, which is a function of treatment, is $\beta_{02} + \beta_{03}\tau$, and the residual effect is $\alpha_{02} + \alpha_{03}\tau$. Just as in the basic mediator setup, the following identities hold:

$$(\beta_{02} + \beta_{03}\tau) - (\alpha_{02} + \alpha_{03}\tau) = (\beta_{12} + \beta_{13}\tau) * (\alpha_{12}) \qquad (4.11)$$

$$\beta_{03} - \alpha_{03} = \beta_{13}\alpha_{12}. \qquad (4.12)$$

According to Muller et al., mediated moderation can only occur when 1) overall moderation exists ($\beta_{03} \neq 0$), 2) both paths are statistically significant, and 3) there a decrease in moderation after adjusting for the mediator [39]. In the context of multisite clinical trials where a treatment-by-site interaction is detected, the first criteria is already met. Also, the decrease in moderation can be measured by the magnitude of the interaction parameter(s). We will refer to the extension of mediated moderation to multisite clinical trials as "multisite mediated moderation" (MMM) from here on.

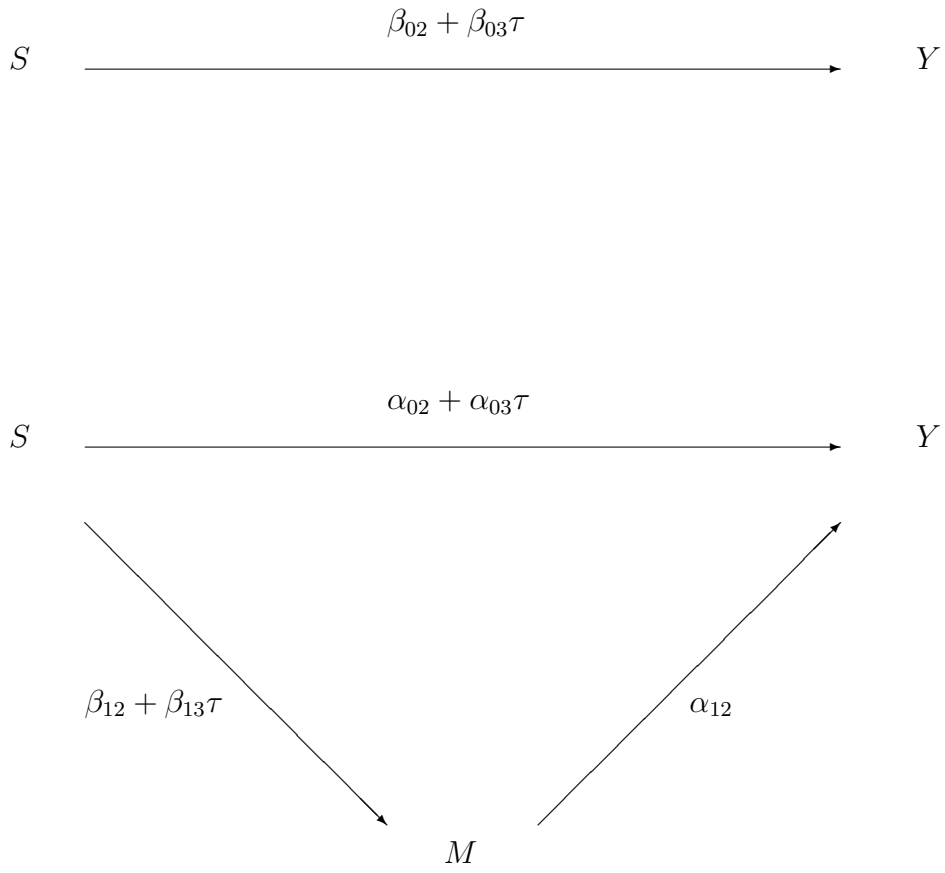

Figure 2: Multisite Mediated Moderation Path Diagram: 1 Mediator

4.2 MMM IN THE 2-SITE CASE

4.2.1 Significance Testing with 1 Mediator

Three significance tests were selected from MacKinnon et al. in order to see how the MMM idea could be applied. From the difference-in-coefficients group, the Freedman & Schatzkin and Olkin & Finn tests were chosen, while the Product of Standardized Coefficients test was chosen from the product-of-coefficients group. For each of the significance tests, their test statistics were extended to the multisite case and their power functions were derived. All significance tests throughout this dissertation were two-sided as well as based on large sample approximations.

4.2.1.1 Freedman & Schatzkin Test For the first test, by Freedman and Schatzkin, the null hypothesis, test statistic, and power function are as follows: $H_0 : \beta_{03} - \alpha_{03} = 0$ with test statistic $t^* = \dfrac{b_{03} - a_{03}}{\sqrt{\hat{\sigma}^2}} \sim t_{n-4}$ where

$$\begin{aligned} \hat{\sigma}^2 &= \hat{\sigma}^2_{b_{03}} + \hat{\sigma}^2_{a_{03}} - 2\hat{\sigma}_{b_{03}}\hat{\sigma}_{a_{03}}\sqrt{1-R^2} \\ &= \frac{MSE_0}{n} + \frac{MSE_2}{(1-R^2)n} - 2\frac{\sqrt{MSE_0}\sqrt{MSE_2}}{n}. \end{aligned} \tag{4.13}$$

Above MSE_i are the estimates of the corresponding errors in equations (4.7), (4.8), and (4.9), and R^2 is the multiple correlation squared when the interaction τs is regressed on τ, s, m, and $m\tau$. The power function is

$$\text{power} = P\left(|t^*| > t_{n-4,1-\alpha/2} \| H_1\right) = P\left(|t_{n-4,\psi}| > t_{n-4,1-\alpha/2}\right) \tag{4.14}$$

where $t_{n-4,\psi}$ is a non-central t distribution with non-centrality parameter $\psi = \dfrac{\Delta}{\sqrt{\sigma^2}}$ for the alternative value $\Delta = \beta_{03} - \alpha_{03}$.

4.2.1.2 Olkin & Finn Test The Olkin and Finn test is based on the difference between $\rho_{ys}(\tau)$, the point-biserial correlation between outcome and site at a particular level of treatment, and $\rho_{ys.m}(\tau)$, the partial point-biserial correlation between the two after accounting for the effect of the mediator. $H_0 : \{\rho_{ys}(\tau) - \rho_{ys.m}(\tau) = 0; \tau = T, C\}$ with joint test statistics $z_T^* = \dfrac{f_T(r)}{\sqrt{\hat{\sigma_T^2}}}$ and $z_C^* = \dfrac{f_C(r)}{\sqrt{\hat{\sigma_C^2}}}$, where $f_i(r) = r_{ysi} - \dfrac{r_{ysi} - r_{ymi}r_{smi}}{\sqrt{1 - r_{ymi}^2}\sqrt{1 - r_{smi}^2}}$ for treatment and control, respectively. We assumed large samples and the multivariate delta method to obtain $\hat{\sigma_i}^2 = \mathbf{a}_i \Phi_i \mathbf{a'}_i$, where $\mathbf{a}_i = \left(\dfrac{\partial f_i(r)}{\partial r_{ysi}}, \dfrac{\partial f_i(r)}{\partial r_{ymi}}, \dfrac{\partial f_i(r)}{\partial r_{smi}} \right)$ and Φ_i is the covariance matrix of the zero-order correlations described in Olkin and Siotani [40]. By definition, this is a union-intersection test, so the null hypothesis is that there is no difference between correlations in both the treatment and control groups. The alternative hypothesis states that there is a significant difference between correlations in at least one of the groups [9]. On the other hand, the hypotheses of a intersection-union test would be defined as follows. The null hypothesis is that there is no difference between correlations in at least one of the two treatment groups, while the alternative states that there is a significant difference in both groups [9].

The critical values, u_1 and u_2, are chosen with type I error such that

$$\alpha = P_{H_0}(|z_T^*| > u_1) + P_{H_0}(|z_C^*| > u_2) - P_{H_0}(|z_T^*| > u_1) P_{H_0}(|z_C^*| > u_2). \quad (4.15)$$

If u_1 and u_2 are both chosen to be $z_{1-\alpha_0/2}$ where $\alpha_0 = 0.0253$, this yields a test of size 0.05. So, the power is

$$\begin{aligned} \text{power} = & P\left(|z| > z_{.98735} - \frac{\Delta_T}{\sqrt{\hat{\sigma_T^2}}}\right) + P\left(|z| > z_{.98735} - \frac{\Delta_C}{\sqrt{\hat{\sigma_C^2}}}\right) \\ & - P\left(|z| > z_{.98735} - \frac{\Delta_T}{\sqrt{\hat{\sigma_T^2}}}\right) P\left(|z| > z_{.98735} - \frac{\Delta_C}{\sqrt{\hat{\sigma_C^2}}}\right). \end{aligned} \quad (4.16)$$

4.2.1.3 Product of Standardized Coefficients Test Finally, the details of the product of standardized coefficients (PSC) test by MacKinnon et al are: $H_0 : \{\beta(\tau)\alpha = 0; \tau = T, C\}$ where $\beta(\tau) = \beta_{12} + \beta_{13}\tau$ and $\alpha = \alpha_{12}$ are independent with respective test statistics $w_T^* = \frac{b_T}{\hat{\sigma}_{b_T}}\frac{a}{\hat{\sigma}_a}$ and $w_C^* = \frac{b_C}{\hat{\sigma}_{b_C}}\frac{a}{\hat{\sigma}_a}$. The standard errors, $\hat{\sigma}_{bi}$ and $\hat{\sigma}_a$, can be explicitly written as $\sqrt{\frac{2MSE_1}{n}}$ and $\sqrt{\frac{MSE_2}{(n-1)s_m^2\left(1-R_{(m)}^2\right)}}$, respectively, where $R_{(m)}^2$ is the multiple correlation from (4.8). Craig showed that the product of two standard normals has pdf $\pi^{-1}K_0(|x|)$, where $K_0(|x|)$ is a modified Bessel function of the second kind with order zero, and provided tables for critical values [12]. Therefore critical values, u_1 and u_2, are chosen with type I error such that

$$\alpha = P_{H_0}\left(|w_T^*| > u_1\right) + P_{H_0}\left(|w_C^*| > u_2\right) - P_{H_0}\left(|w_T^*| > u_1\right) P_{H_0}\left(|w_C^*| > u_2\right). \quad (4.17)$$

The power function follows

$$\begin{aligned} \text{power} &= P_{H_1}\left(|w_T^*| > d_{1-\alpha_0/2}\right) + P_{H_1}\left(|w_C^*| > d_{1-\alpha_0/2}\right) \\ &\quad - P_{H_1}\left(|w_T^*| > d_{1-\alpha_0/2}\right) P_{H_1}\left(|w_C^*| > d_{1-\alpha_0/2}\right), \end{aligned} \quad (4.18)$$

where $d_{1-\alpha_0/2}$ is the appropriate critical value and α_0 is the significance level chosen to achieve an overall size α test. Under the alternative hypothesis, $|w_i^*|$ is distributed as the product of two independent normal random variables with non-null means and unit variance. Closed form expressions are not available for this distribution, so we used Gauss-Hermite quadrature with 32 knots to approximate the probability function (see Appendix B) [1].

A difficulty arises such that the test statistics are distributed as modified Bessel functions under the null hypothesis only when both α and β are zero. As a consequence, the true type I error will be inflated. Craig derived the mean and variance of the product of two independent normal variables divided by their respective standard deviations [12]:

$$\mathrm{E}\left(\frac{ba}{\sigma_\beta\sigma_\alpha}\right) = \frac{\beta\alpha}{\sigma_\beta\sigma_\alpha} \quad (4.19)$$

$$\mathrm{Var}\left(\frac{ba}{\sigma_\beta\sigma_\alpha}\right) = \frac{\beta^2}{\sigma_\beta^2} + \frac{\alpha^2}{\sigma_\alpha^2} + 1. \quad (4.20)$$

It should be evident from the above central moments that as one of the means grows, the variability of the product also grows. While the sample mean is still symmetric around the origin, the tails become fatter thus increasing the type I error [12]. The simulation study of the product of standardized coefficients in MacKinnon et al. validated this [35].

4.2.1.4 Combination Test One remedy is to rewrite the product of coefficients hypothesis as a combination of intersection-union and union-intersection tests (referred to from now on as the combo test). The null hypothesis would be that at least one of the indirect paths is non-significant in both treatment and control. The alternative is that both indirect paths are significant in at least one of the treatment groups. $H_0 : \{\beta(\tau) = 0 \text{ or } \alpha = 0; \tau = T, C\}$ with test statistics $t^*_{bi} = \dfrac{b_i}{\hat{\sigma}_{bi}}$ and $t^*_a = \dfrac{a}{\hat{\sigma}_a}$ for $i = T, C$, and where $\hat{\sigma}_{bi}$ and $\hat{\sigma}_a$ are described previously. Then, critical values, u_i, with an overall type I error can be chosen such that

$$\begin{aligned} \alpha &= P_{H_0}\left(|t^*_{bT}| > u_1\right) P_{H_0}\left(|t^*_a| > u_2\right) + P_{H_0}\left(|t^*_{bC}| > u_3\right) P_{H_0}\left(|t^*_a| > u_4\right) \\ &\quad - P_{H_0}\left(|t^*_{bT}| > u_1\right) P_{H_0}\left(|t^*_a| > u_2\right) P_{H_0}\left(|t^*_{bC}| > u_3\right) P_{H_0}\left(|t^*_a| > u_4\right). \end{aligned} \tag{4.21}$$

In the worst case scenario that two of the parameters, say β_T and β_C, have non-null means, then the overall type I error rate is still determined by the choice of u_2 and u_4 [9]. Therefore, all critical values should be chosen to be $t_{df,0.98735}$ in order to achieve an overall type I error rate of 0.05, where $df = n - 5$ for u_2 and u_4 and $df = n - 4$ for u_1 and u_3.

4.2.1.5 Variance Stabilizing Transformation Test Another modification of the product of standardized coefficients test is to use a variance stabilizing transformation [32]. Since we know

$$\sqrt{n}\left(\frac{b_i}{\sigma_1} - \frac{\beta_i}{\sigma_1}\right) \Longrightarrow N(0,1) \quad \text{and} \quad \sqrt{n}\left(\frac{a}{\sigma_2} - \frac{\alpha}{\sigma_2}\right) \Longrightarrow N(0,1),$$

where $\sigma_1 = 2\sigma$ and $\sigma_2 = \sqrt{\dfrac{n\sigma^2}{(n-1)s^2_m\left(1 - R^2_{(m)}\right)}}$, then by the multivariate delta method [32]

$$\sqrt{n}\left(\frac{b_i a}{\sigma_1\sigma_2} - \frac{\beta_i\alpha}{\sigma_1\sigma_2}\right) \Longrightarrow N\left(0, \frac{\alpha^2}{\sigma^2_2} + \frac{\beta^2_i}{\sigma^2_1}\right).$$

So, a variance stabilizing transformation is a function, $f(\theta)$, that satisfies

$$f'(\theta) = \frac{c}{\sqrt{\tau^2(\theta)}}, \tag{4.22}$$

where $\tau^2(\theta)$ is the variance as a function of θ. If we let $\theta = \frac{\beta_i \alpha}{\sigma_1 \sigma_2}$, then we can reparametrize and let $\frac{\theta}{w} = \frac{\beta_i}{\sigma_1}$ and $w = \frac{\alpha}{\sigma_2}$. Now, $\tau^2(\theta) = \frac{\theta^2}{w^2} + w^2$ and it is evident that the variance is proportional to the square of the mean. At first glance, this would suggest a logarithmic transformation [21]. If we set $w = 1$ and integrate both sides of equation 4.22 with $c = 1$, we get the variance stabilizing function to be

$$f(\theta) = \sinh^{-1}(\theta) = \ln\left(\theta + \sqrt{1+\theta^2}\right). \tag{4.23}$$

Therefore,

$$\sqrt{n}\left(f(\hat{\theta}) - f(\theta)\right) \Longrightarrow N(0,1), \tag{4.24}$$

where $\hat{\theta} = \frac{b_i a}{\hat{\sigma}_1 \hat{\sigma}_2}$. Since we parametrized $w = 1$, then we can fix $\alpha = \sigma_2$ and the hypothesis test becomes $H_0 : \{\theta(\tau) = 0; \tau = T, C\}$ with test statistics $z_i^* = \sqrt{n}\left\{ f\left(\hat{\theta}\right) - f\left(\theta\right) \right\}$ where $i = T, C$ is defined above.

In order to check the type I error and power of the VST test, we conducted a simulation study with the following steps.

- One million replicates of the following random variables were drawn: $a \sim N\left(\alpha, \frac{\sigma_2^2}{n}\right)$ and $b_i \sim N\left(\beta_i, \frac{\sigma_1^2}{n}\right)$ where $\alpha = \sigma_2$ is set.
- For each replicate, $\hat{\theta} = \frac{b_i a}{\hat{\sigma}_1 \hat{\sigma}_2}$ and $z^* = \sqrt{n}\left\{ f\left(\hat{\theta}\right) - f\left(\theta\right) \right\}$ are calculated.
- The true type I error and power are estimated as the number of $|z|$'s that exceed the critical value under the null or alternative, respectively.

4.2.1.6 d **Test** A fundamental issue regarding the tests that fall into the difference-in-coefficients category is the assumption of normality in the estimates of the coefficients. If we take a look at the basic mediator setup, we can see that β_{01} and α_{01} are estimated by

$$b_{01} = \frac{\sum_{i=1}^{n} y_i \tau_i}{n} \tag{4.25}$$

$$a_{01} = \frac{1}{n}\sum_{i=1}^{n} y_i \tau_i - \frac{b_{11}\sum_{i=1}^{n}(y_i - \bar{y})(m_i - \bar{m})}{(n-1)\hat{\sigma}_1^2}. \tag{4.26}$$

If we call the difference $\hat{d} = b_{01} - a_{01}$, then conditional on m and τ, the distribution of $\hat{d}$ is Gaussian with the following mean and variance

$$\mathrm{E}(\hat{d}|m,\tau) = \frac{b_{11}\sum_{i=1}^{n}\mathrm{E}(y_i - \bar{y})(m_i - \bar{m})}{(n-1)\hat{\sigma}_1^2} = \frac{b_{11}\beta_{01}\sum_{i=1}^{n} m_i\tau_i}{(n-1)\hat{\sigma}_1^2} \tag{4.27}$$

$$\mathrm{Var}(\hat{d}|m,\tau) = \left[\frac{b_{11}}{(n-1)\hat{\sigma}_1^2}\right]^2 \sum_{i=1}^{n}\mathrm{Var}(y_i - \bar{y})(m_i - \bar{m})^2 = \frac{b_{11}^2\sigma_0^2 s_m^2}{(n-1)(\hat{\sigma}_1^2)^2}. \tag{4.28}$$

Due to the equality (4.6), a significance test of $H_0 : d = 0$ should be equivalent (or nearly equivalent) to the Freedman-Schatzkin test.

Since m is random, our interest should be on the unconditional distribution of $\hat{d}$. The reason for this is that the unconditional variance is always larger than the conditional. Therefore, many of the analyses that treat the mediator as a fixed quantity when it is indeed random can severely underestimate the variance, which can overestimate the power.

The unconditional mean and variance of $\hat{d}$ are

$$\mathrm{E}(\hat{d}) = \mathrm{E}\left[\mathrm{E}(\hat{d}|m,\tau)\right] = \frac{\beta_{01}}{n-1}\mathrm{E}\left[\frac{\left(\sum_{i=1}^{n} m_i\tau_i\right)^2}{\hat{\sigma}_1^2}\right] \tag{4.29}$$

$$\begin{aligned}\mathrm{Var}(\hat{d}) &= \mathrm{Var}\left[\mathrm{E}(\hat{d}|m,\tau)\right] + \mathrm{E}\left[\mathrm{Var}(\hat{d}|m,\tau)\right] \\ &= \frac{\sigma_0^2}{(n-1)^2}\mathrm{E}\left[\frac{s_m^2\left(\sum_{i=1}^{n} m_i\tau_i\right)^2}{(\hat{\sigma}_1^2)^2}\right] + \frac{\beta_{01}^2}{n^2}\mathrm{Var}\left[\frac{\left(\sum_{i=1}^{n} m_i\tau_i\right)^2}{\hat{\sigma}_1^2}\right],\end{aligned} \tag{4.30}$$

but the distribution is complicated. We intend to investigate this further in the future with the hope of modifying the current difference-in-coefficients tests, but the remainder of this dissertation will focus on the conditional case.

As was mentioned previously, both the Freedman-Schatzkin test and the significance test of d should be equivalent with respect to their test statistics and standard errors. In the two-site case, $\hat{d} = b_{03} - a_{03} = b_{13}a_{12}$ is normally distributed with the following conditional mean and variance:

$$\begin{aligned} \mathrm{E}(\hat{d}|m,\tau,s) &= b_{13}\alpha_{12} && (4.31) \\ \mathrm{Var}(\hat{d}|m,\tau,s) &= b_{13}^2\mathrm{Var}(a_{12}) = \frac{b_{13}^2\sigma_2^2}{(1-R_{(m)}^2)\sum_{i=1}^{N}(m_i-\bar{m})}, && (4.32) \end{aligned}$$

where $\sigma_2^2 = \sigma_0^2 - \dfrac{\sigma_{01}^2}{\sigma_1^2} = \sigma_0^2(1-\rho_{ym}^2)$ and $R_{(m)}^2$ is the multiple squared correlation from (4.8). So, a t-test can be constructed with the following standard error: $\hat{\sigma}_d^2 = \dfrac{b_{13}^2 MSE_2}{(N-1)s_m^2(1-R_{(m)}^2)}$, where MSE_2 is the estimate of the error variance in the conditional distribution of y.

The explicit power function of the d significance test is as follows:

$$\text{power} = P\left(|t^*| > t_{n-5,1-\alpha/2} \| H_1\right) = P\left(|t_{n-5,\psi}| > t_{n-5,1-\alpha/2}\right), \quad (4.33)$$

where $t_{n-5,\psi}$ is a non-central t distribution with non-centrality parameter $\psi = \dfrac{\Delta}{\sqrt{\sigma_d^2}}$ for the alternative value $\Delta = \beta_{03} - \alpha_{03}$.

4.2.2 Power Analysis with 1 Mediator

Power analyses of each of the aforementioned 6 hypotheses were conducted using their explicit power functions and the following assumptions. First, the overall design was assumed to have two sites with two treatments at each site as well as equal sample sizes at each treatment-site combination. Second, each of the mediated moderation equations were assumed to have the same error variance of 1. As a consequence, (4.13) was simplified to

$$\hat{\sigma}^2 = \frac{MSE}{n}\left[\frac{R^2}{1-R^2}\right]. \quad (4.34)$$

For each of the power analyses, partial correlation effect sizes were chosen to correspond to small (.14), medium (.36), and large (.51). The sample sizes were chosen to be 50, 100, 200, 400, 500, and 1000. The overall type I error rate was chosen to be 0.05 for all hypothesis tests. For the Freedman-Schatzkin test, the term $\frac{R^2}{1-R^2}$ in the standard error is similar to an f^2 effect size with 0.02, 0.15, and 0.35 values for small, medium, and large, respectively [10]. The large effect of f^2, which corresponds to an R^2 of 0.51, was chosen to give a lower bound to power calculations. Similarly for the PSC test, $R^2_{(m)}$ was chosen to be 0.51 for the same reason. Since the true error variance for the d-test, σ_2^2, involves the correlation between the outcome and mediator variable (ρ^2_{ym}), this was set to 0.14^2 to exhibit a type of lower bound for the power.

4.2.2.1 Results The results of the two difference-in-coefficients tests – Freedman-Schatzkin and Olkin-Finn – were first examined (Table 1). Regarding the type I error, both tests achieved alpha levels of approximately 0.05 regardless of sample size. Nevertheless, it was the Freedman-Schatzkin test that attained greater power for a given effect size and sample size, which mimicked the results obtained by MacKinnon et al. in the basic mediation case [35].

Turning to the product-of-coefficients tests, we looked at the cases where both hypothesized parameters were identical ($\alpha = \beta$) as well as when they were not. In the first case, when the product of standardized coefficients test was examined, both tests seemed to do well in achieving an overall type I error of 0.05 although the combo test well underestimated it at 0.001. In the case where one of the parameters varies, the issue of the inflated type I error mentioned above appears (Table 2). For example as β increased for a given sample size, the type I error rate of the combo test tended to 0.05 while the type I error of the product of standardized coefficients test increased rapidly to 1. Despite the extremely low type I error rate for the combo test, the size of the test is still 0.05. In order to clearly see this, consider the first term of equation 4.21. Casella and Berger [9] showed that if one of probabilities is sent to 1 (say, the right one), then the probability of the product is determined by the values u_1 and u_3. So, in order to have an overall size of 0.05, u_1 and u_3 must be set to critical values corresponding with 0.053 probability.

Table 1: Type I error & power for MMM with 1 mediator

		Sample Size					
ES	Method	50	100	200	400	500	1000
0	Freedman-Schatzkin	0.050	0.050	0.050	0.050	0.050	0.050
	Olkin-Finn	0.050	0.050	0.050	0.050	0.050	0.050
	d	0.050	0.050	0.050	0.050	0.050	0.050
0.14	Freedman-Schatzkin	0.373	0.647	0.914	0.997	>0.999	>0.999
	Olkin-Finn	0.203	0.363	0.639	0.918	0.966	>0.999
	d	0.133	0.224	0.400	0.679	0.774	0.970
0.36	Freedman-Schatzkin	0.988	>0.999	>0.999	>0.999	>0.999	>0.999
	Olkin-Finn	0.857	0.993	>0.999	>0.999	>0.999	>0.999
	d	0.572	0.869	0.993	>0.999	>0.999	>0.999
0.51	Freedman-Schatzkin	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999
	Olkin-Finn	0.993	>0.999	>0.999	>0.999	>0.999	>0.999
	d	0.859	0.992	>0.999	>0.999	>0.999	>0.999

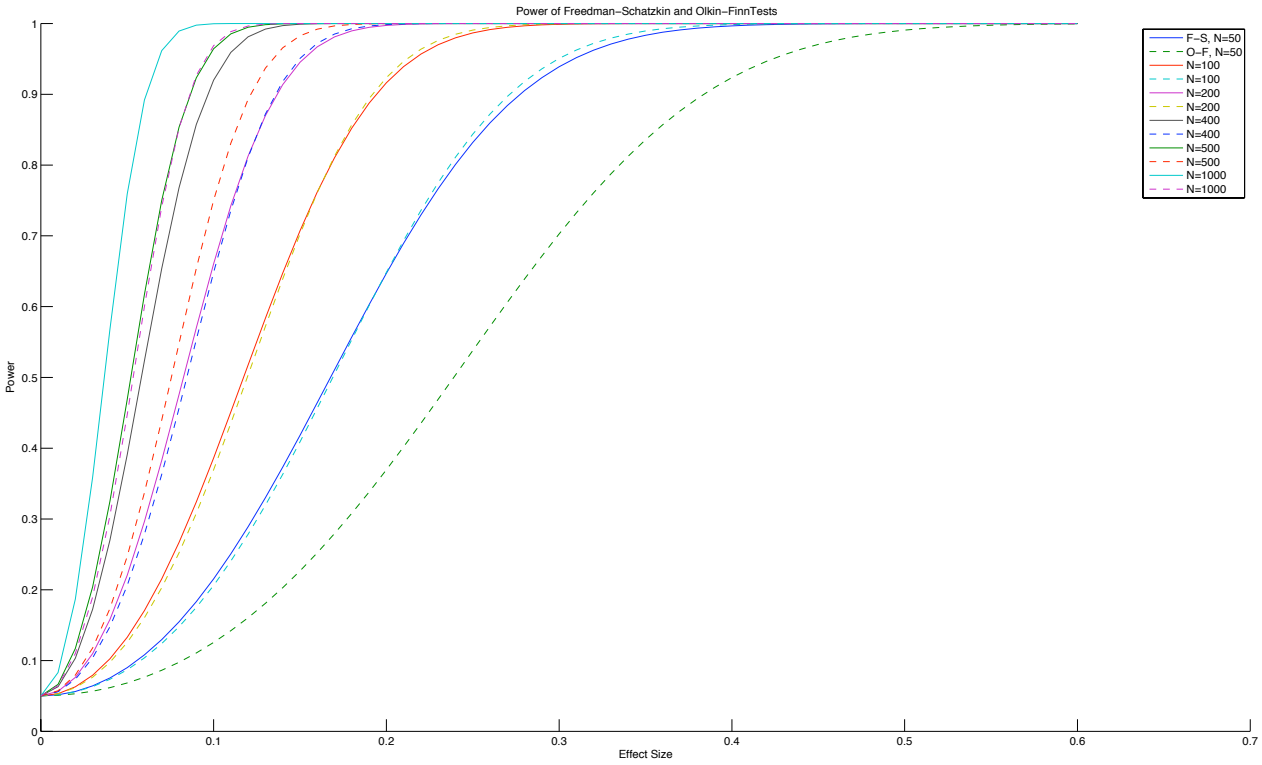

Figure 3: Power of Difference-in-Coefficients Tests

As can be seen in Table 3, the variance stabilizing transformation (VST) method has more accurate type I error as well as greater power than the combo method. Despite the fact that as either α or β grow large the combo test tends to a 0.05 size test, there could be many cases were neither one is large. As a consequence, the type I error will be underestimated making the null hypothesis difficult to reject. On the other hand, the VST will always have size 0.05 regardless of the value of the non-null parameter.

From the results of the power analysis, it is clear that both difference-in-coefficients tests (Freedman-Schatzkin and Olkin-Finn) performed well in terms of high power and accurate type I error. While the power of the product of standardized coefficients test was also large, it suffered from an inflated type I error. The combo test, a result of the modification of the product-of-coefficient's hypotheses, resulted in high power with the caveat of the type I error being underestimated.

Table 2: Type I error for MMM with 1 mediator and only one non-zero parameter

		Sample Size					
β	Method	50	100	200	400	500	1000
0	Prod Stand Coef	0.050	0.050	0.050	0.050	0.050	0.050
	Combo	0.001	0.001	0.001	0.001	0.001	0.001
0.14	Prod Stand Coef	0.107	0.162	0.264	0.428	0.491	0.687
	Combo	0.003	0.005	0.010	0.020	0.025	0.041
0.36	Prod Stand Coef	0.377	0.571	0.764	0.895	0.905	0.907
	Combo	0.016	0.031	0.046	0.050	0.050	0.050
0.51	Prod Stand Coef	0.572	0.765	0.895	0.907	0.907	0.963
	Combo	0.030	0.046	0.050	0.050	0.050	0.050

MacKinnon et al. gives a thorough overview of the best methods for testing basic mediation, while the goal of this dissertation is to extend those ideas to the multisite clinical trial setting. The Freedman-Schatzkin test, the Olkin-Finn test, and MacKinnon's product of standardized coefficients test were chosen not only due to their uniqueness, but also because they each had shown considerable power in the basic mediator setup. While the first two

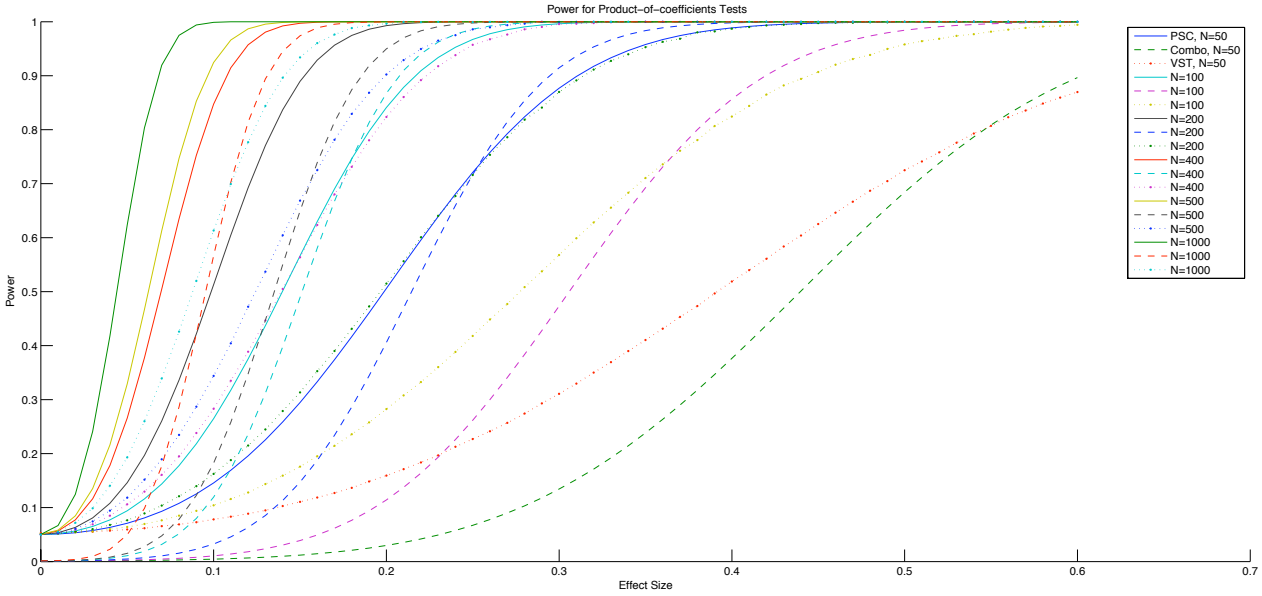

Figure 4: Power of Product-of-Coefficients Tests

tests performed well in terms of both type I error and power, the last test suffered from the same issue as in the basic mediator setup, inflated type I error. As a result, we rewrote the hypothesis in terms of a combination of an intersection-union and a union-intersection test. As the results showed, the level of the test was invariant to either of the parameters having a true non-zero mean.

Table 3: Type I error & power for MMM with 1 mediator

		Sample Size					
ES	Method	50	100	200	400	500	1000
0	Prod Stand Coef	0.050	0.050	0.050	0.050	0.050	0.050
	VST	0.054	0.052	0.051	0.051	0.050	0.050
	Combo	0.001	0.001	0.001	0.001	0.001	0.001
0.14	Prod Stand Coef	0.170	0.318	0.603	0.912	0.964	>0.999
	VST	0.103	0.159	0.279	0.505	0.605	0.896
	Combo	0.010	0.031	0.113	0.380	0.550	0.945
0.36	Prod Stand Coef	0.844	0.993	>0.999	>0.999	>0.999	>0.999
	VST	0.432	0.736	0.963	>0.999	>0.999	>0.999
	Combo	0.264	0.731	0.988	>0.999	>0.999	>0.999
0.51	Prod Stand Coef	0.993	>0.999	>0.999	>0.999	>0.999	>0.999
	VST	0.741	0.964	>0.999	>0.999	>0.999	>0.999
	Combo	0.712	0.988	>0.999	>0.999	>0.999	>0.999

We compared the power of the d significance test to that of the other two tests in the difference-in-coefficients group with the results in Table 1. While the power of the d test is clearly lower than the Freedman-Schatzkin and Olkin-Finn tests, an increase in ρ^2_{ym} to 0.51^2 increases the power by a factor 0.1511.

4.2.3 Significance Testing with K Mediators

It is plausible that more than one variable could be responsible for the mediation of treatment-by-site interaction. In this case, the mediated moderation setup as well as the three significance tests can be extended. The three necessary equations – (4.2), (4.3), and (4.4) – can be modified to reflect multiple mediators:

$$\begin{aligned} y &= \beta_{00} + \beta_{01}\tau + \beta_{02}s + \beta_{03}(\tau * s) + \epsilon_0 && (4.35)\\ m_1 &= \beta_{10} + \beta_{11}\tau + \beta_{12}s + \beta_{13}(\tau * s) + \epsilon_1 && (4.36)\\ &\vdots \\ m_K &= \beta_{K0} + \beta_{K1}\tau + \beta_{K2}s + \beta_{K3}(\tau * s) + \epsilon_K && (4.37)\\ y &= \alpha_{00} + \alpha_{01}\tau + \alpha_{02}s + \alpha_{03}(\tau * s) + \alpha_{12}m_1 + \cdots + \alpha_{K2}m_K + \epsilon_{K+1}. && (4.38) \end{aligned}$$

Based on the joint distribution of the outcome variable, y, and the mediators, we have the following:

$$\begin{pmatrix} y \\ m_1 \\ \vdots \\ m_K \end{pmatrix} \sim \mathrm{N}\left\{ \begin{pmatrix} \beta_{00} + \beta_{01}\tau + \beta_{02}s + \beta_{03}(\tau * s) \\ \beta_{10} + \beta_{11}\tau + \beta_{12}s + \beta_{13}(\tau * s) \\ \vdots \\ \beta_{K0} + \beta_{K1}\tau + \beta_{K2}s + \beta_{K3}(\tau * s) \end{pmatrix}, \begin{bmatrix} \sigma_0^2 & \sigma_{01} & \dots & \sigma_{0K} \\ \sigma_{10} & \sigma_1^2 & \dots & \sigma_{1K} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{K0} & \sigma_{K1} & \dots & \sigma_K^2 \end{bmatrix} \right\}.$$

If we partition y from the mediator variables such that the variance-covariance matrix is

$$\Sigma_{ym} = \begin{bmatrix} \sigma_0^2 & \Sigma_{0m} \\ \Sigma_{m0} & \Sigma_{mm} \end{bmatrix},$$

then the conditional distribution of y is

$$(y|m_1, \dots, m_K, \tau, s) \sim \mathrm{N}\left(\mu_y + \Sigma_{0m}\Sigma_{mm}^{-1}(\mathbf{m} - \boldsymbol{\mu}_m), \sigma_0^2 - \Sigma_{0m}\Sigma_{mm}^{-1}\Sigma_{m0}\right).$$

In the case of two mediators, it's straightforward to show (see Appendix C) that the conditional mean of y is

$$\begin{aligned} \mu_{y|m_1,m_2,\tau,s} &= (\beta_{00} - \alpha_{12}\beta_{10} - \alpha_{22}\beta_{20}) + (\beta_{01} - \alpha_{12}\beta_{11} - \alpha_{22}\beta_{21})\tau \\ &\quad + (\beta_{02} - \alpha_{12}\beta_{12} - \alpha_{22}\beta_{22})s + (\beta_{03}\tau - \alpha_{12}\beta_{13} - \alpha_{22}\beta_{23})(\tau * s) \\ &\quad + \alpha_{12}m_1 + \alpha_{22}m_2 \\ &= \alpha_{00} + \alpha_{01}\tau + \alpha_{02}s + \alpha_{03}(\tau * s) + \alpha_{12}m_1 + \alpha_{22}m_2. \qquad (4.39) \end{aligned}$$

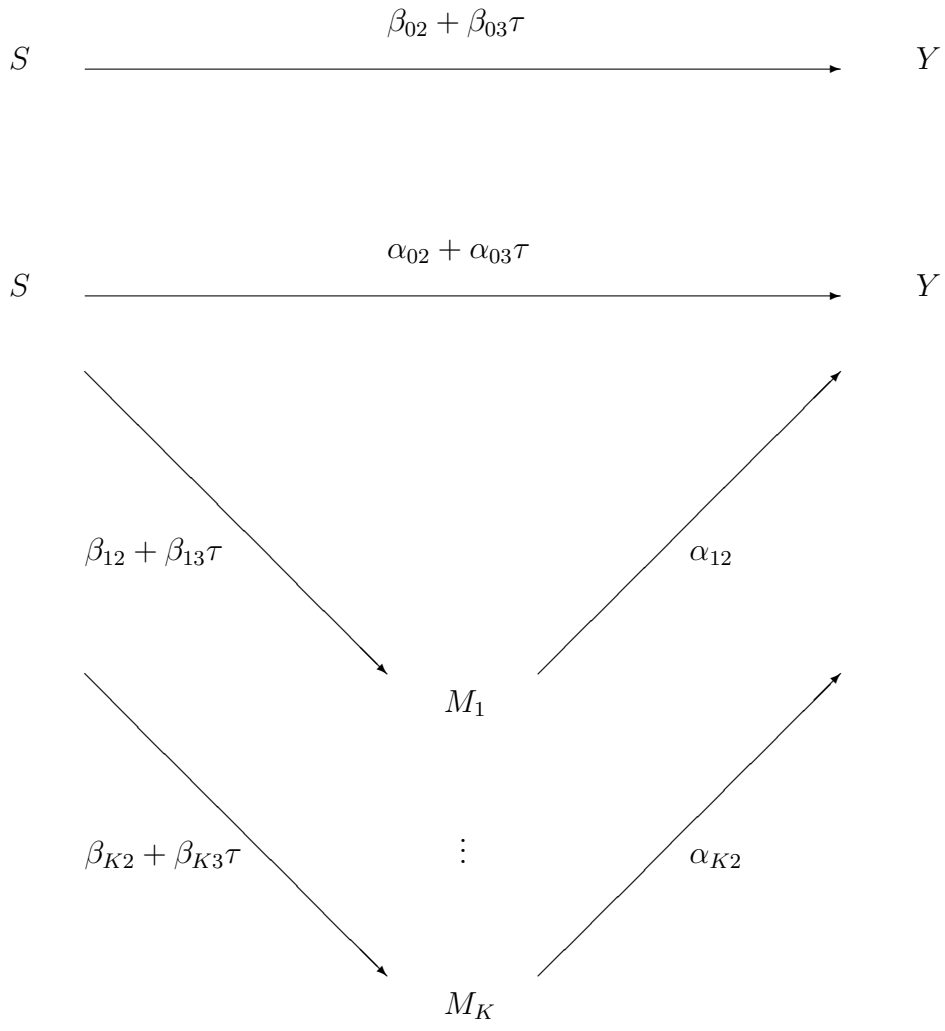

Figure 5: MMM Path Diagram: K Mediators

4.2.3.1 Freedman & Schatzkin Test The three original tests are modified in the following ways: $H_0 : \beta_{03} - \alpha_{03} = 0$ with test statistic $t^* = \dfrac{b_{03} - a_{03}}{\sqrt{\hat{\sigma}^2}} \sim t_{n-4}$ where

$$\begin{aligned} \hat{\sigma}^2 &= \hat{\sigma}^2_{b_{03}} + \hat{\sigma}^2_{a_{03}} - 2\hat{\sigma}_{b_{03}}\hat{\sigma}_{a_{03}}\sqrt{1-R^2} \\ &= \frac{MSE_0}{n} + \frac{MSE_{K+1}}{(1-R^2)n} - 2\frac{\sqrt{MSE_0}\sqrt{MSE_{K+1}}}{n}, \end{aligned} \tag{4.40}$$

where R^2 is the multiple correlation squared when τs is regressed on τ, s, and the mediators $m_1,\ldots,m_K$. Expressions for power are the same as in the single mediator case:

$$\text{power} = P\left(|t^*| > t_{n-4,1-\alpha/2} \| H_1\right) = P\left(|t_{n-4,\psi}| > t_{n-4,1-\alpha/2}\right) \tag{4.41}$$

where $t_{n-4,\psi}$ is a non-central t distribution with non-centrality parameter $\psi = \dfrac{\Delta}{\sqrt{\sigma^2}}$ for the alternative value $\Delta = \beta_{03} - \alpha_{03}$.

4.2.3.2 Olkin & Finn Test $H_0 : \rho_{ys}(\tau) - \rho_{ys.1\cdots K}(\tau) = 0$ with the same test statistics as before except that the functions f_i now become:

$$f_i(r) = r_{ys} - r_{ys.1\cdots K} \tag{4.42}$$

where the second term is the correlation between outcome and site at a particular treatment level after accounting for the effects of the K mediators (denoted $1\cdots K$). Despite the fact that higher order partial correlations can be decomposed into functions of zero-order partial correlations, depending on K, $f_i(r)$ can be very complicated to compute. Raveh (1985) showed that high-order partial correlations can be computed from the inverse correlation matrix of all zero-order correlations involved [43]. For example with inverse correlation $P = R^{-1}$, the partial correlation of two variables i and j holding a set of variables Δ constant is

$$r_{ij.\Delta} = \frac{-p_{ij}}{\sqrt{p_{ii}p_{jj}}}. \tag{4.43}$$

As was done in the single mediator case, the standard error of $f_i(r)$, $\mathbf{a}_i\mathbf{\Phi}_i\mathbf{a}'_i$, is computed where $\mathbf{a}$ is the vector of partial derivatives of $f_i(r)$ with respect to each pairwise correlation of zero-order correlations. In the case of K mediators, the dimensions of $\mathbf{a}$ and $\mathbf{\Phi}$ are 1-by-$(K+2)$ and $(K+2)$-by-$(K+2)$.

4.2.3.3 Product of Standardized Coefficients Test In the presence of K mediators, the third test is now: $H_0 : \sum_{q=1}^{K} \beta_q(\tau)\alpha_q = 0$ where $\beta_q(\tau) = \beta_{q2} + \beta_{q3}\tau$ and $\alpha_q = \alpha_{q2}$ are independent, and test statistics $L_T^* = \sum_{q=1}^{K} \left(\frac{b_{qT}}{\hat{\sigma}_{b_{qT}}} \frac{a_q}{\hat{\sigma}_{a_q}} \right)$ and $L_C^* = \sum_{q=1}^{K} \left(\frac{b_{qC}}{\hat{\sigma}_{b_{qC}}} \frac{a_q}{\hat{\sigma}_{a_q}} \right)$. The main interest is on the pdf of L_T^* and L_C^* under the null hypothesis, which is the sum of modified Bessel functions. Instead, each of the terms in the summation can be thought of as normal scale mixtures. If we let $W_q = \frac{b_q}{\hat{\sigma}_{b_q}} \frac{a_q}{\hat{\sigma}_{a_q}} = \mu_{b_q}\mu_{a_q}$ (regardless of treatment), then $W_q = (Z_{b_q} + \mu_{b_q})(Z_{a_q} + \mu_{a_q})$ where Z_{b_q} and Z_{a_q} are independent standard normal variables. After rearranging the terms, we get $W_q = Z_{a_q}Z_{b_q} + X_q$ where $X_q \sim \mathrm{N}(\mu_{a_q}\mu_{b_q}, \mu_{a_q}^2 + \mu_{b_q}^2)$. This is equivalent in distribution to $W_q = |Z_{a_q}|Z_{b_q} + X_s$, where $|Z_{a_q}|Z_{b_q}$ is a gaussian scale mixture. The following is a theorem and corresponding proof.

Theorem 1. *If $Y = Z_0Z_1$ and $W = |Z_0|Z_1$, where Z_0 and Z_1 are independent standard normal variables, then Y and W are equal in distribution. In this case, W is distributed as a scale mixture of a normal distribution with a chi-square mixing density.*

Proof. Assume that $W = |Z_0|Z_1$. Because $|Z_0|$ is a special case of the folded normal distribution [13], $|Z_0|$ can be written as $W = \chi_1 Z_1$, where χ_1 is a chi random variate with 1 degree of freedom. High order moments of W are given by the following:

$$\begin{aligned} \mathrm{E}(W^{2n}) &= E(\chi_1^{2n})E(Z_1^{2n}) = \left[\frac{2^n\Gamma(n+1/2)}{\sqrt{\pi}}\right]\left[\frac{(2n)!}{2^n n!}\right] \\ &= \left[\frac{(2n-1)!!\sqrt{\pi}}{\sqrt{\pi}2^n}\right]\left[\frac{(2n)!}{n!}\right], \end{aligned} \tag{4.44}$$

where $(2n-1)!!$ denotes the double factorial of $2n-1$. This is equivalent to $\frac{(2n)!}{2^n n!}$, so the above equation simplifies to:

$$\mathrm{E}(W^{2n}) = \left[\frac{(2n)!}{2^n n!}\right]^2, \tag{4.45}$$

which produces the exact same moments as the product of two standard normal variables. By Carleman's Condition (Chung 1974), if we can show that

$$\sum_{n=1}^{\infty} \frac{1}{[\mathrm{E}(W^{2n})]^{1/(2n)}} = +\infty,$$

then the sequence of moments produced by $\mathrm{E}(W^{2n})$ is unique. So, using Stirling's approximation we get

$$\mathrm{E}(W^{2n}) \approx \left[\frac{\sqrt{4\pi n}\left(\frac{2n}{e}\right)^{2n}}{2^n\sqrt{2\pi n}\left(\frac{n}{e}\right)^n}\right]^2 = \frac{2^{1+4n-2n}n^{2n}}{e^{2n}} \approx \left(\frac{2n}{e}\right)^{2n}, \tag{4.46}$$

and subsequently,

$$\sum_{n=1}^{\infty}\frac{1}{[\mathrm{E}(W^{2n})]^{1/(2n)}} = \frac{e}{2}\sum_{n=1}^{\infty}\frac{1}{n} = +\infty.$$

Therefore, since the distribution of W is uniquely defined by a sequence of moments shared by Y, then both Y and W must come from the same distribution. Hence Y and W are equal in distribution. Alternatively, this can be proven by showing the equivalence of characteristic functions. □

In addition, the covariance between any two Gaussian variates, Z_{b_q} and $Z_{b_{q'}}$, will be assumed to be $\sigma_{qq'}$. Conditioning on Z_{a_q}, $V_q = W_q|Z_{a_q} \sim \mathrm{N}(\mu_{b_q}\mu_{a_q}, z_{a_q}^2 + \mu_{a_q}^2 + \mu_{b_q}^2)$ with the following covariance:

$$\begin{aligned}
\mathrm{Cov}(V_q, V_{q'}) &= \mathrm{Cov}(|z_{a_q}|Z_{b_q} + X_q, |z_{a_{q'}}|Z_{b_{q'}} + X_{q'}) \\
&= |z_{a_q}z_{a_{q'}}|\mathrm{Cov}(Z_{b_q}, Z_{b_{q'}}) + |z_{a_q}|\mathrm{Cov}(Z_{b_q}, X_{q'}) + |z_{a_{q'}}|\mathrm{Cov}(Z_{b_{q'}}, X_q) + \mathrm{Cov}(X_q, X_{q'}) \\
&= |z_{a_q}z_{a_{q'}}|\sigma_{qq'} + |z_{a_q}|\mu_{a_{q'}}\sigma_{qq'} + |z_{a_{q'}}|\mu_{a_q}\sigma_{qq'} \\
&= \sigma_{qq'}\left(|z_{a_q}z_{a_{q'}}| + |z_{a_q}|\mu_{a_{q'}} + |z_{a_{q'}}|\mu_{a_q}\right)
\end{aligned} \tag{4.47}$$

Therefore, $L_i = \sum_{q=1}^{K} V_q$ is normally distributed with mean and variance:

$$\mathrm{E}(L_i) = \mathrm{E}\left(\sum_{q=1}^{K} V_q\right) = \sum_{q=1}^{K}\mu_{b_q}\mu_{a_q} \tag{4.48}$$

$$\begin{aligned}
\mathrm{Var}(L_i) &= \mathrm{Var}\left(\sum_{q=1}^{K} V_q\right) = \sum_{q=1}^{K}\mathrm{Var}(V_q) + 2\sum_{q\neq q'}^{K}\mathrm{Cov}(V_q, V_{q'}) \\
&= \sum_{q=1}^{K}(z_{a_q}^2 + \mu_{a_q}^2 + \mu_{b_q}^2) + 2\sum_{q\neq q'}^{K}\left[\sigma_{qq'}\left(|z_{a_q}z_{a_{q'}}| + |z_{a_q}|\mu_{a_{q'}} + |z_{a_{q'}}|\mu_{a_q}\right)\right],
\end{aligned} \tag{4.49}$$

where $|z_{a_q}|$ and $|z_{a_{q'}}|$ are folded normal random variables and $z_{a_q}^2$ is a chi squared random variable.

4.2.3.4 *d* **Test** Finally, the d significance test is extended to the K mediator case using the distribution of y conditional on the mediator variables shown in Appendix C. $H_0 : d = \beta_{03} - \alpha_{03} = 0$ with estimate $\hat{d} = b_{03} - a_{03} = \sum_{q=1}^{K} b_{q3}a_{q2}$. The conditional variance is derived as follows:

$$\begin{aligned}
\mathrm{Var}(b_{03} - a_{03}|m_1, \ldots, m_K, \tau, s) &= \sum_{q=1}^{K} b_{q3}^2 \mathrm{Var}(a_{q2}) + 2\sum_{q\neq q'}^{K} b_{q3}b_{q'3}\mathrm{Cov}(a_{q2}, a_{q'2}) \\
&= \sum_{q=1}^{K} b_{q3}^2 \mathrm{Var}(a_{q2}) + 2\sum_{q\neq q'}^{K} b_{q3}b_{q'3}\sqrt{\mathrm{Var}(a_{q2})}\sqrt{\mathrm{Var}(a_{q'2})}\rho_{qq'} \\
&= \sum_{q=1}^{K} b_{q3}^2 \frac{MSE_{K+1}}{(N-1)s_{m_q}^2(1-R_{(m_q)}^2)} \\
&\quad +2\sum_{q\neq q'}^{K} b_{q3}b_{q'3}\frac{MSE_{K+1}\rho_{qq'}}{(N-1)s_{m_q}s_{m_{q'}}\sqrt{(1-R_{(m_q)}^2)}\sqrt{(1-R_{(m_{q'})}^2)}} \\
&= \frac{MSE_{K+1}}{N-1}\left[\sum_{q=1}^{K} b_{q3}^2 \frac{1}{s_{m_q}^2(1-R_{(m_q)}^2)}\right. \\
&\quad \left. +2\sum_{q\neq q'}^{K} b_{q3}b_{q'3}\frac{\rho_{qq'}}{s_{m_q}s_{m_{q'}}\sqrt{(1-R_{(m_q)}^2)}\sqrt{(1-R_{(m_{q'})}^2)}}\right],
\end{aligned} \tag{4.50}$$

where MSE_{K+1} is the estimate of the conditional variance of y, $R_{(m_q)}^2$ is the multiple correlation squared when the mediator m_q is regressed on τ, s, τs, and the remaining $K-1$ mediators, and $\rho_{qq'}$ is the correlation between a_{q2} and $a_{q'2}$.

The power function for the d-test is

$$\text{power} = P\left(|t^*| > t_{n-4-K),1-\alpha/2}\|H_1\right) = P\left(|t_{n-4-K,\psi}| > t_{n-4-K,1-\alpha/2}\right), \tag{4.51}$$

where $t_{n-4-K,\psi}$ is a non-central t distribution with non-centrality parameter $\psi = \dfrac{\Delta}{\sqrt{\sigma^2}}$ for the alternative value $\Delta = \beta_{03} - \alpha_{03}$ and above conditional variance $\hat{\sigma^2}$.

4.2.3.5 Limitations

As with many significance tests, there are some limitations of extending the MMM tests to the K mediator case. With regard to the Olkin-Finn test, the standard error of the test statistic involves a vector of partial derivatives of (4.42) with respect to the outcome, the site, and each of the K mediator variables. It's evident that even with a modest number of mediators, (4.42) as a function of zero-order correlations can get very complicated.

Another limitation involves the covariance parameters in the PSC and d tests. It's reasonable to assume that they are non-zero since the K mediators are correlated, but it is not clear just how correlated they are. This will be investigated along with other future work.

4.2.4 Illustration on TORDIA data

Each of the significance tests described so far all deal with the situation where the outcome variable is continuous. In the TORDIA clinical trial, the outcome was dichotomous where clinical response was defined as the combination of the Clinical Global Impressions (CGI) score ≤ 2 and a change in the Children's Depression Rating Scale-Revised (CDRS-R) of $\geq 50\%$. For the purposes of illustrating our six tests on this data, the change in the CDRS-R was used as the outcome variable. In addition, since TORDIA involved six sites, the three sites that were most similar in terms of their CBT-MED effect were combined into one. As can be seen in Table 5, sites 1, 3, and 4 were combined into one site due to the similar direction of their respective treatment effects. The three regression equations involved in mediated

Table 4: Treatment effect across sites for TORDIA data

	Sites						
TX	1	2	3	4	5	6	ALL
MED	0.503	0.446	0.465	0.511	0.342	0.312	0.435
CBT-MED	0.351	0.600	0.368	0.485	0.520	0.480	0.485
Δ	0.152	-0.154	0.097	0.026	-0.178	-0.168	-0.050

moderation ((4.7), (4.8), and (4.9)) were each conducted with 20 variables measured at baseline as potential mediators. First, there was overall site moderation of treatment effect in (4.7) ($b_{03} = -6.1478; p = 0.001$) at the $\alpha = 0.10$ significance level. Second, the conflict behavior questionnaire score (CBQA) was the only variable in which there was a significant treatment-by-site interaction in (4.8) ($b_{13} = 0.6651; p = 0.053$). Finally, overall moderation was reduced after adjusting for CBQA ($a_{03} = -5.6542; p = 0.002$). Therefore, according to Muller et al., the criteria for mediated moderation have been met.

We then applied the three difference-in-coefficients and three product-of-coefficients tests to this data to see if the treatment-by-site interaction was significantly reduced. In the first group, both the Freedman-Schatzkin and d significance tests concluded significant decreases ($p_{FS} = 0.0108; p_d = 0.0095$), while the Olkin-Finn test did not ($p = 0.4886$). In the second group, both MacKinnon's product of standardized coefficients and the Combo test rejected the null hypothesis ($p_{PSC} = 0.0249; p_C = 0.0036$), while the significance test based on the variance stabilizing transformation did not ($p = 0.9763$).

One thing to note from this illustration is that while overall moderation was significantly reduced after accounting for the CBQA mediator, it was not eliminated. So, CBQA explains only part of the treatment-by-site interaction, and it could be the case that there are other mediator variables involved in the explanation. Another issue in this illustration is that while there were 20 variables checked for significant treatment-by-site interaction, there was no adjustment for multiple comparisons. This is clearly a limitation to this procedure.

4.3 MMM IN THE J-SITE CASE

4.3.1 Significance Testing with 1 Mediator

The extension of the two-site mediated moderation model to the J-site case is straightforward. The three necessary equations are shown below:

$$y_{ijl} = \mu_{0\ldots} + \tau_{0i} + s_{0j} + \gamma_{0ij} + \epsilon_{0ijl} \tag{4.52}$$

$$m_{ijl} = \mu_{1...} + \tau_{1i} + s_{1j} + \gamma_{1ij} + \epsilon_{1ijl} \tag{4.53}$$

$$y_{ijl} = \mu_{2...} + \tau_{2i} + s_{2j} + \gamma_{2ij} + \alpha_1 m_{ijl} + \epsilon_{2ijl} \tag{4.54}$$

where τ, s, γ, and m are the effect of treatment, site, their interaction, and the mediator variable, respectively. As was the case in the two-site case, the equations above are based on the joint distribution of y and m as well as the conditional distribution of y given m, τ, and s.

$$\begin{pmatrix} y_{ijl} \\ m_{ijl} \end{pmatrix} \sim \mathrm{N}\left\{ \begin{pmatrix} \mu_{0...} + \tau_{0i} + s_{0j} + \gamma_{0ij} + \epsilon_{0ijl} \\ \mu_{1...} + \tau_{1i} + s_{1j} + \gamma_{1ij} + \epsilon_{1ijl} \end{pmatrix}, \begin{bmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{01} & \sigma_1^2 \end{bmatrix} \right\},$$

and

$$(y_{ijl}|m_{ijl}, \tau, s) \sim \mathrm{N}\left(\mu_{0...} + \tau_{0i} + s_{0j} + \gamma_{0ij} + \frac{\sigma_{01}}{\sigma_1^2}\left(m_{ijl} - \mu_{1...} - \tau_{1i} - s_{1j} - \gamma_{1ij}\right), \sigma_0^2 - \frac{\sigma_{01}^2}{\sigma_1^2}\right),$$

and the conditional mean can be rewritten as

$$\begin{aligned} \mathrm{E}(y_{ijl}|m_{ijl}, \tau, s) &= \left(\mu_{0...} - \frac{\sigma_{01}}{\sigma_1^2}\mu_{1...}\right) + \left(\tau_{0i} - \frac{\sigma_{01}}{\sigma_1^2}\tau_{1i}\right) + \left(s_{0j} - \frac{\sigma_{01}}{\sigma_1^2}s_{1j}\right) + \left(\gamma_{0ij} - \frac{\sigma_{01}}{\sigma_1^2}\gamma_{1ij}\right) \\ &\quad + \frac{\sigma_{01}}{\sigma_1^2}m_{ijl} \\ &= \mu_{2...} + \tau_{2i} + s_{2j} + \gamma_{2ij} + \alpha_1 m_{ijl}, \end{aligned} \tag{4.55}$$

where $i = 1, ..., I$, $j = 1, ...J$, $l = 1, ..., n_{ij}$, and $N = \sum_{i=1}^{2}\sum_{j=1}^{J} n_{ij}$. Equation (4.12) can easily be extended to the GLM case such that

$$\gamma_{0ij} - \gamma_{2ij} = \frac{\sigma_{01}}{\sigma_1^2}\gamma_{1ij} = \alpha_1\gamma_{1ij}. \tag{4.56}$$

For (4.11), reparametrizing the marginal means of y_{ijl} and m_{ijl} so that $\mu_{gi} = \mu_{g...} + \tau_{gi}$ and $s_{gij} = s_{gj} + \gamma_{gij}$ for $g = 0, 1, 2$ gives the following equality:

$$s_{0ij} - s_{2ij} = \frac{\sigma_{01}}{\sigma_1^2}s_{1ij} = \alpha_1 s_{1ij}. \tag{4.57}$$

The remainder of this work will revolve around the equality in (4.56) rather than (4.57) because the primary goal is to investigate the change in site moderation rather than the effect of site as a function of treatment.

4.3.1.1 *d* **Test** The d significance test can be extended to the GLM framework as follows: $H_0 : d_{ij} = \gamma_{0ij} - \gamma_{2ij} = 0$ for $i = T, C$ with test statistics $t^*_{ij} = \dfrac{\hat{d}_{ij}}{\hat{\sigma}_{d_{ij}}}$. Because there are $(I-1)(J-1)$ interaction parameters to estimate, one could imagine how monotonous the union-intersection test could get when there are a modest number of sites. On the other hand, the estimates of the interaction coefficients are correlated and have a unique covariance structure. Consider a 2-treatment, J-site ANOVA model with the following parameter constraint: $\sum_{i=1}^{2} \tau_i = \sum_{j=1}^{J} s_j = \sum_{i=1}^{2} \gamma_{ij} = \sum_{j=1}^{J} \gamma_{ij} = 0$. Then conditional on m, τ and s, $\hat{d}_{ij} = \hat{\gamma}_{0ij} - \hat{\gamma}_{2ij} = a_1 \hat{\gamma}_{1ij}$ is normally distributed with the following conditional mean and variance:

$$\mathrm{E}(\hat{d}_{ij}|m, \tau, s) = \mathrm{E}(a_1 \hat{\gamma}_{1ij}) = \hat{\gamma}_{1ij} \alpha_1 \tag{4.58}$$

$$\mathrm{Var}(\hat{d}_{ij}|m, \tau, s) = \mathrm{Var}(a_1 \hat{\gamma}_{1ij}) = \hat{\gamma}^2_{1ij} \frac{\sigma_2^2}{(N-1) s_m^2 (1 - R^2_{(m)})}. \tag{4.59}$$

In addition, the conditional covariance between any two $\hat{d}_{ij}$'s from differing sites, j and j', is

$$\mathrm{Cov}(\hat{d}_{ij}, \hat{d}_{ij'}|m, \tau, s) = \mathrm{Cov}\left(a_1 \hat{\gamma}_{1ij}, a_1 \hat{\gamma}_{1ij'}\right) = \hat{\gamma}_{1ij} \hat{\gamma}_{1ij'} \frac{\sigma_2^2}{(N-1) s_m^2 (1 - R^2_{(m)})}. \tag{4.60}$$

In matrix form, this is

$$\begin{pmatrix} \hat{d}_{11} \\ \hat{d}_{12} \\ \vdots \\ \hat{d}_{1J-1} \end{pmatrix} \sim N \left\{ \begin{pmatrix} \mu_{\hat{d}_{11}} \\ \mu_{\hat{d}_{12}} \\ \vdots \\ \mu_{\hat{d}_{1J-1}} \end{pmatrix}, \begin{bmatrix} \sigma^2_{\hat{d}_{11}} & \sigma_{\hat{d}_{11}\hat{d}_{12}} & \cdots & \sigma_{\hat{d}_{11}\hat{d}_{1J-1}} \\ \sigma_{\hat{d}_{11}\hat{d}_{12}} & \sigma^2_{\hat{d}_{12}} & \cdots & \sigma_{\hat{d}_{12}\hat{d}_{1J-1}} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{\hat{d}_{11}\hat{d}_{1J-1}} & \sigma_{\hat{d}_{12}\hat{d}_{1J-1}} & \cdots & \sigma^2_{\hat{d}_{1J-1}} \end{bmatrix} \right\},$$

where $\mu_{\hat{d}_{ij}}$, $\sigma^2_{\hat{d}_{ij}}$, and $\sigma_{\hat{d}_{ij}\hat{d}_{ij'}}$ represent the conditional means, variances, and covariances from above. Under the assumption that $\boldsymbol{\mu}_{\hat{d}} = \mathbf{0}$, $\hat{\mathbf{d}} \sim N_{J-1}(\mathbf{0}, \boldsymbol{\Sigma})$.

A simplification of the distribution of $\hat{\mathbf{d}}$ is to use the equality in (4.56). So,

$$\begin{aligned} \mathrm{E}(\hat{\mathbf{d}}|m, \tau, s) &= \mathrm{E}(\hat{\boldsymbol{\gamma}}_1 a_1 | m, \tau, s) \\ &= \hat{\boldsymbol{\gamma}}_1 \alpha_1 \end{aligned} \tag{4.61}$$

and

$$\begin{aligned}\text{Var}(\hat{\mathbf{d}}|m,\tau,s) &= \text{Var}(\hat{\boldsymbol{\gamma}}_1 a_1|m,\tau,s) \\ &= \hat{\boldsymbol{\gamma}}_1 \frac{\sigma_2^2}{(N-1)s_m^2(1-R_{(m)}^2)}\hat{\boldsymbol{\gamma}}_1{}' \\ &= \frac{\sigma_2^2}{(N-1)s_m^2(1-R_{(m)}^2)}\hat{\boldsymbol{\gamma}}_1\hat{\boldsymbol{\gamma}}_1{}', \end{aligned} \tag{4.62}$$

where $R_{(m)}^2$ is the multiple correlation of the mediator variables regressed on τ, s, and γ.

The estimate of the above variance is $\frac{1}{(N-1)s_m^2(1-R_{(m)}^2)}\frac{SSE_2}{N-2J-1}\hat{\boldsymbol{\gamma}}_1\hat{\boldsymbol{\gamma}}_1{}'$, which is a scaled chi-squared random variable divided by degrees of freedom $N-2J-1$ and multiplied by a vector of constants [24]. The null hypothesis can now be written as
$H_0 : \mathbf{d} = \mathbf{0}$ with test statistic

$$\begin{aligned} T^2 &= \hat{\mathbf{d}}'\hat{\boldsymbol{\Sigma}}^{-1}\hat{\mathbf{d}} \\ &= (\hat{\boldsymbol{\gamma}}_1 a_1)' \left(\frac{MSE_2}{(N-1)s_m^2(1-R_{(m)}^2)}\hat{\boldsymbol{\gamma}}_1\hat{\boldsymbol{\gamma}}_1{}'\right)^{-1}(\hat{\boldsymbol{\gamma}}_1 a_1) \\ &= a_1^2\left(\frac{MSE_2}{(N-1)s_m^2(1-R_{(m)}^2)}\right)^{-1}\hat{\boldsymbol{\gamma}}_1{}'(\hat{\boldsymbol{\gamma}}_1\hat{\boldsymbol{\gamma}}_1{}')^{-}\hat{\boldsymbol{\gamma}}_1 \\ &= a_1^2\left(\frac{MSE_2}{(N-1)s_m^2(1-R_{(m)}^2)}\right)^{-1}, \end{aligned} \tag{4.63}$$

where $()^-$ denotes the generalized inverse of the singular matrix $\hat{\boldsymbol{\gamma}}_1\hat{\boldsymbol{\gamma}}_1{}'$. It is easy to show that $\hat{\boldsymbol{\gamma}}_1{}'(\hat{\boldsymbol{\gamma}}_1\hat{\boldsymbol{\gamma}}_1{}')^{-}\hat{\boldsymbol{\gamma}}_1$ reduces to one (see Appendix E). The test statistic is distributed as $F_{1,N-2J-1}$. The critical values are chosen such that

$$\alpha = P_{H_0}\left(T^2 > F_{1,N-2J-1,1-\alpha}\right). \tag{4.64}$$

Under the alternative hypothesis, T^2 is distributed as a non-central F random variable with non-centrality parameter, $\psi = \frac{\Delta^2}{\sigma^2}$ where $\Delta = \alpha_1$ and $\sigma^2 = \text{Var}(a_1)$. Therefore, the power is calculated as follows:

$$\text{power} = P\left(F_{1,N-2J-1,\psi} > F_{1,N-2J-1,1-\alpha}\right). \tag{4.65}$$

4.3.1.2 Product of Standardized Coefficients Test In the GLM framework, MacKinnon's PSC test can be extended to the multivariate case, but it is difficult because a multivariate Bessel distribution has not been studied yet. Since there are $J-1$ estimates of the treatment-by-site interactions, there will be just as many product-of-coefficients estimates. Marginally, each of the estimates is the product of normal random variables each with unit variance:

$$\begin{pmatrix} \frac{a_1}{\sigma_a}\frac{\hat{\gamma}_{111}}{\sigma_{\hat{\gamma}}} \\ \frac{a_1}{\sigma_a}\frac{\hat{\gamma}_{112}}{\sigma_{\hat{\gamma}}} \\ \vdots \\ \frac{a_1}{\sigma_a}\frac{\hat{\gamma}_{11J-1}}{\sigma_{\hat{\gamma}}} \end{pmatrix} \sim \left\{ \begin{pmatrix} \frac{\alpha_1}{\sigma_\alpha}\frac{\gamma_{111}}{\sigma_\gamma} \\ \frac{\alpha_1}{\sigma_\alpha}\frac{\gamma_{112}}{\sigma_\gamma} \\ \vdots \\ \frac{\alpha_1}{\sigma_\alpha}\frac{\gamma_{11J-1}}{\sigma_\gamma} \end{pmatrix}, \begin{bmatrix} \Sigma_{\alpha\gamma} \end{bmatrix} \right\},$$

where $\Sigma_{\alpha\gamma}$ is the variance-covariance matrix with

$$\text{Var}\left(\frac{a_1}{\sigma_a}\frac{\hat{\gamma}_{11j}}{\sigma_{\hat{\gamma}}}\right) = \left(\frac{\gamma_{11j}}{\sigma_\gamma}\right)^2 + \left(\frac{\alpha_1}{\sigma_\alpha}\right)^2 + 1 \tag{4.66}$$

on the diagonals and

$$\begin{aligned} \text{Cov}\left(\frac{a_1}{\sigma_a}\frac{\hat{\gamma}_{11j}}{\sigma_{\hat{\gamma}}}, \frac{a_1}{\sigma_a}\frac{\hat{\gamma}_{11j'}}{\sigma_{\hat{\gamma}}}\right) &= (MSE_1)\left(\frac{\hat{\sigma}_{01}}{\hat{\sigma}_1^2}\right)^2 \left[\frac{1}{4J^2}\sum_{i=1}^{2}\sum_{j=1}^{J}\left(\frac{1}{n_{ij}}\right) - \frac{1}{4J}\sum_{i=1}^{2}\left(\frac{1}{n_{ij}} + \frac{1}{n_{ij'}}\right)\right] \\ &\quad + \left(\frac{\gamma_{11j}}{\sigma_\gamma}\frac{\gamma_{11j'}}{\sigma_\gamma}\right)\left(1 - \left(\frac{\alpha_1}{\sigma_\alpha}\right)^2\right) \end{aligned} \tag{4.67}$$

on the off-diagonals. While the variance was derived by Craig [12], the derivation of covariance of any two estimates can be shown in the following example.

Assume that X_1, X_j, and $X_{j'}$ are all random variables with unit variance and respective means μ_1, μ_j, and $\mu_{j'}$. Also, $X_1 \perp X_j$, $X_1 \perp X_{j'}$, and $\text{Cov}(X_j, X_{j'}) = \sigma_{jj'}$. Then

$$\begin{aligned}
\text{Cov}(X_1X_j, X_1X_{j'}) &= \text{E}(X_1^2X_jX_{j'}) - \text{E}(X_1X_j)\text{E}(X_1X_{j'}) \\
&= \text{E}\left[\text{E}(X_1^2X_jX_{j'}|X_1)\right] - \mu_1^2\mu_j\mu_{j'} \\
&= \text{E}\left[X_1^2\text{E}(X_jX_{j'})\right] - \mu_1^2\mu_j\mu_{j'} \\
&= \text{E}\left[X_1^2(\sigma_{jj'} + \mu_j\mu_{j'})\right] - \mu_1^2\mu_j\mu_{j'} \\
&= (\sigma_{jj'} + \mu_j\mu_{j'}) - \mu_1^2\mu_j\mu_{j'} = \sigma_{jj'} + \mu_j\mu_{j'}(1 - \mu_1^2).
\end{aligned} \tag{4.68}$$

The difficulty arises in the multivariate distribution of the above vector of estimates. To simplify this problem, let us go back to the two-site case. In this case, $\frac{a_1}{\sigma_a}\frac{\hat{\gamma}_{111}}{\sigma_{\hat{\gamma}}}$ is distributed as the product of two normal random variables each with unit variance. Under the null hypothesis assumed by MacKinnon ($\alpha_1 = \boldsymbol{\gamma} = \mathbf{0}$), this is distributed as the product of two standard normal variables, denoted by $Y = Z_0Z_1$, which has the density function $\pi^{-1}K_0(|Y|)$. The following theorem shows that a vector of correlated Bessel functions, $\mathbf{Y}$, can instead be denoted as a multivariate normal scale mixture.

Theorem 2. *Let* $\mathbf{Y} = Z_0\mathbf{Z}$ *and* $\mathbf{W} = |Z_0|\mathbf{Z}$*, where* Z_0 *is a standard normal random variable independent of* $\mathbf{Z}$*, which is a p-dimensional vector coming from a multivariate normal distribution with zero mean vector and covariance matrix* Σ*. Then* $\mathbf{Y}$ *is equivalent in distribution to* $\mathbf{W}$*. In this case,* $\mathbf{W}$ *is distributed as a scale mixture of a multivariate normal distribution with a chi-square mixing density.*

Proof. The simplest way to prove this is by showing the equivalence of the characteristic functions of $\mathbf{Y}$ and $\mathbf{W}$. So,

$$\begin{aligned}
\phi_{\mathbf{Y}}(\mathbf{t}) &= \text{E}\left(\exp\{i\mathbf{t}'Z_0\mathbf{Z}\}\right) = \text{E}\left[\text{E}\left(\exp\{i\mathbf{t}'Z_0\mathbf{Z}\}|Z_0\right)\right] \\
&= \text{E}\left(\exp\left\{\frac{-Z_0^2\mathbf{t}'\Sigma\mathbf{t}}{2}\right\}\right) \\
&= (1 + \mathbf{t}'\Sigma\mathbf{t})^{-1/2} \\
&= \phi_{\mathbf{W}}(\mathbf{t}).
\end{aligned} \tag{4.69}$$

□

Theorem 3. *Let* $\mathbf{Y} = (Z_0 + \mu_0)(\mathbf{Z} + \boldsymbol{\mu})$ *and* $\mathbf{W} = |Z_0|\mathbf{Z} + \mathbf{X}$ *where* $\mathbf{X} \sim MVN(\mu_0\boldsymbol{\mu}, \mu_0^2\Sigma + \boldsymbol{\mu}\boldsymbol{\mu}')$*, and* Z_0 *and* $\mathbf{Z}$ *are independent of each other and defined in the previous theorem. Then* $\mathbf{Y}$ *and* $\mathbf{W}$ *are equal in distribution.*

Proof. We can rewrite $\mathbf{Y}$ so that it is the sum of independent components:

$$\mathbf{Y} = Z_0\mathbf{Z} + \mu_0\mathbf{Z} + \boldsymbol{\mu}Z_0 + \mu_0\boldsymbol{\mu}. \tag{4.70}$$

If we denote the sum of the last three terms as $\mathbf{X}$, it's easy to show that this is the sum of two multivariate normal random variables and a vector of scalars. So,

$$(\mu_0\mathbf{Z}) \sim \text{MVN}(\mathbf{0}, \mu_0^2\Sigma)$$

and

$$(\boldsymbol{\mu}Z_0) \sim \text{MVN}(\mathbf{0}, \boldsymbol{\mu}\boldsymbol{\mu}')$$

and are independent of each other. Therefore, $\mathbf{X} \sim \text{MVN}(\mu_0\boldsymbol{\mu}, \mu_0^2\Sigma + \boldsymbol{\mu}\boldsymbol{\mu}')$. The characteristic function is as follows:

$$\phi_{\mathbf{Y}}(\mathbf{t}) = \phi_{Z_0\mathbf{Z}}(\mathbf{t})\phi_{\mathbf{X}}(\mathbf{t}) = \phi_{|Z_0|\mathbf{Z}}(\mathbf{t})\phi_{\mathbf{X}}(\mathbf{t}), \tag{4.71}$$

where the second equality comes from Theorem 2.

$$\begin{aligned}
\phi_{\mathbf{Y}}(\mathbf{t}) &= \phi_{|Z_0|\mathbf{Z}}(\mathbf{t})\phi_{\mathbf{X}}(\mathbf{t}) \\
&= (1 + \mathbf{t}'\Sigma\mathbf{t})^{-1/2}\text{exp}\left\{i\mathbf{t}'\mu_0\boldsymbol{\mu} - \frac{1}{2}\mathbf{t}'\left(\mu_0^2\Sigma + \boldsymbol{\mu}\boldsymbol{\mu}'\right)\mathbf{t}\right\} \\
&= \phi_{\mathbf{W}}(\mathbf{t}).
\end{aligned} \tag{4.72}$$

Since both $\mathbf{Y}$ and $\mathbf{W}$ have the same characteristic functions, they're equal in distribution. □

Now that an appropriate multivariate distribution has been identified, we can now continue with the details of the PSC significance test. $H_0 : \alpha_1\boldsymbol{\gamma}_1 = \mathbf{0}$ where $\boldsymbol{\gamma}$ is a $J-1$ dimensional vector of the interaction parameters. Following MacKinnon et al., the test statistic is $\mathbf{W} = \frac{a_1}{\sigma_a}\frac{\hat{\boldsymbol{\gamma}}}{\sigma_{\hat{\boldsymbol{\gamma}}}}$ and is distributed under the alternative hypothesis as a scale mixture of multivariate normal random variables with the following conditional and marginal densities:

$$(\mathbf{W}|z_0) \sim \text{MVN}\left(\frac{\alpha_1}{\sigma_\alpha}\frac{\boldsymbol{\gamma}}{\sigma_\gamma}, \left(z_0^2 + \left(\frac{\alpha_1}{\sigma_\alpha}\right)^2\right)\Sigma_\gamma + \frac{\boldsymbol{\gamma}}{\sigma_\gamma}\frac{\boldsymbol{\gamma}\,'}{\sigma_\gamma}\right)$$

where

$$z_0^2 \sim \chi_1^2$$

and Σ_γ is the variance-covariance matrix of the standardized interaction estimates (see Appendix D). A Hotelling's T^2 test can be conducted with test statistic
$T^2 = \mathbf{W}'(z_0^2\hat{\Sigma}_\gamma)^{-1}\mathbf{W}$. Under the null hypothesis, $(T^2|z_0^2) \sim \frac{(N-2J)(J-1)}{N-3J+2}F_{J-1,N-3J+2}$, so the critical values are chosen such that

$$\begin{aligned}\alpha &= P_{H_0}\left(T^2 > \frac{(N-2J)(J-1)}{N-3J+2}F_{J-1,N-3J+2,1-\alpha}\right) \\ &= P_{H_0}\left(\frac{N-3J+2}{(N-2J)(J-1)}T^2 > F_{J-1,N-3J+2,1-\alpha}\right).\end{aligned} \tag{4.73}$$

Under the alternative hypothesis, $\frac{N-3J+2}{(N-2J)(J-1)}T^2$ is distributed as a non-central F random variable with non-centrality parameter,
$\psi = \left(\frac{\alpha_1}{\sigma_\alpha}\frac{\boldsymbol{\gamma}}{\sigma_\gamma}\right)'\left(\left(z_0^2 + \left(\frac{\alpha_1}{\sigma_\alpha}\right)^2\right)\Sigma_\gamma + \frac{\boldsymbol{\gamma}}{\sigma_\gamma}\frac{\boldsymbol{\gamma}\,'}{\sigma_\gamma}\right)^{-1}\left(\frac{\alpha_1}{\sigma_\alpha}\frac{\boldsymbol{\gamma}}{\sigma_\gamma}\right)$ [18]. Therefore, the power is calculated as follows:

$$\text{power} = P\left(F_{J-1,N-3J+2,\psi} > F_{J-1,N-3J+2,1-\alpha}\right). \tag{4.74}$$

As in the two-site case, the null hypothesis described by MacKinnon can only arise when both parameters, α_1 and $\boldsymbol{\gamma}$, are zero. The only requirement for the general null hypothesis is that either of the two parameters is zero, so the inflated type I error issues described in the two-site case apply here as well.

4.3.2 Power Analysis with 1 Mediator

Power analyses of each of the above significance tests were conducted using their explicit power functions and the following assumptions. First, the overall design was assumed to have two treatments at each of J sites as well as as equal sample sizes at each treatment-site combination. Second, σ_0^2 and σ_1^2 were assumed to be 1. Finally, each of the $\hat{\gamma}_{1ij}$, the interactive effects of treatment and site on the mediator variable, were identically chosen to be 1. For each of the power analyses, effect sizes were chosen to correspond to small (.20), medium (.50), and large (.80) according to Cohen [10]. The sample sizes were chosen to be 100, 200, 400, 500, and 1000. In addition, the number of sites were chosen to be 5, 10, and 20. The overall type I error rate was chosen to be 0.05 for all hypothesis tests.

For the PSC test, the non-centrality parameter involves $\sigma_\alpha = \dfrac{1}{\sqrt{1 - R^2_{(m)}}}$, where $R^2_{(m)}$ is the multiple correlation of the mediator variables regressed on τ, s, and γ. Just as in the two-site, two-treatment case, $R^2_{(m)}$ was chosen to be 0.51. Also, the test statistic T^2 contains a chi-squared random variable, z_0^2. Therefore, the first, second, and third quartiles of χ_1^2 were chosen for values of z_0^2.

4.3.2.1 Results The results of the d and PSC tests are displayed in the following three tables. When the values of z_0^2 were varied, the differences in the resulting power were negligible. Therefore, only the first quartile of χ_1^2 was used. Unlike in the two-site case, the PSC test is much more powerful than the d test regardless of the number of sites. Also, the power of the d test did not decrease with the number of sites as rapidly as the PSC test did. As in the two-site case, the issue of inflated type I error in the PSC test may occur. The degree to which this effects the significance test will be investigated in future work. Similar to the variance stabilizing transformation used previously, a logarithmic transformation could be used on the test statistic, $\mathbf{W}$, as a remedy for the inflated type I error.

4.3.3 Illustration on TORDIA data

As in the two-site case, each of the significance tests was conducted on the TORDIA clinical trial dataset with the outcome being the change in the CDRS-R variable. Since this is a

Table 5: Type I error & power for MMM with 1 mediator and $j = 5$ sites

		Sample Size				
ES	Method	100	200	400	500	1000
0	d	0.050	0.050	0.050	0.050	0.050
	Prod Stand Coef	0.050	0.050	0.050	0.050	0.050
0.20	d	0.395	0.675	0.929	0.970	>0.999
	Prod Stand Coef	0.898	0.998	>0.999	>0.999	>0.999
0.50	d	0.989	0.999	>0.999	>0.999	>0.999
	Prod Stand Coef	>0.999	>0.999	>0.999	>0.999	>0.999
0.80	d	>0.999	>0.999	>0.999	>0.999	>0.999
	Prod Stand Coef	>0.999	>0.999	>0.999	>0.999	>0.999

Table 6: Type I error & power for MMM with 1 mediator and $j = 10$ sites

		Sample Size				
ES	Method	100	200	400	500	1000
0	d	0.050	0.050	0.050	0.050	0.050
	Prod Stand Coef	0.050	0.050	0.050	0.050	0.050
0.20	d	0.394	0.675	0.929	0.970	>0.999
	Prod Stand Coef	0.785	0.992	>0.999	>0.999	>0.999
0.50	d	0.988	>0.999	>0.999	>0.999	>0.999
	Prod Stand Coef	0.998	>0.999	>0.999	>0.999	>0.999
0.80	d	>0.999	>0.999	>0.999	>0.999	>0.999
	Prod Stand Coef	>0.999	>0.999	>0.999	>0.999	>0.999

Table 7: Type I error & power for MMM with 1 mediator and $j = 20$ sites

		Sample Size				
ES	Method	100	200	400	500	1000
0	d	0.050	0.050	0.050	0.050	0.050
	Prod Stand Coef	0.050	0.050	0.050	0.050	0.050
0.20	d	0.391	0.674	0.929	0.970	>0.999
	Prod Stand Coef	0.526	0.956	>0.999	>0.999	>0.999
0.50	d	0.988	>0.999	>0.999	>0.999	>0.999
	Prod Stand Coef	0.958	>0.999	>0.999	>0.999	>0.999
0.80	d	>0.999	>0.999	>0.999	>0.999	>0.999
	Prod Stand Coef	0.999	>0.999	>0.999	>0.999	>0.999

multisite example, the original six sites and their respective data were preserved.

The three models in the multisite mediated moderation case ((4.52), (4.53), and (4.54)) were conducted with 20 variables measured at baseline as potential mediators. There was overall site moderation of treatment effect in (4.52) ($p = 0.025$) at the $\alpha = 0.05$ significance level. There were two variables in which significant treatment-by-site interaction persisted: 1) the CBQA variable mentioned in the previous section ($p = 0.044$), and 2) the TOTALD variable ($p = 0.004$), which is a measure of the Drug Use Screening Inventory (DUSI).

After adjusting for the CBQA score, the magnitude of site moderation was reduced ($p = 0.066$), which is a criterion for mediated moderation. Moreover, the interaction was non-significant after the adjustment. Both tests were applied to the data to see if the reductions in magnitudes were significant. Since the PSC test involves a χ^2_1 random variable, the expected value of 1 was chosen. The d test concluded that subject's CBQA score significantly explained the differing effect of CBT-MED therapy across the six sites ($p = 0.0153$). The Product of Standardized Coefficients test concluded that the magnitude of interaction was significantly reduced ($p < 0.001$). This agrees with the conclusions of the paper by Spirito et al.

As was the case with the CBQA variable, the magnitude of site moderation due to the adjustment for the TOTALD variable was reduced ($p = 0.041$). Both of the aforementioned tests were applied to the data and only the PSC test concluded a significant reduction ($p_d = 0.193; p_{PSC} < 0.001$). In other words, there were conflicting results as to whether the TOTALD score alone explained the differing effects of CBT-MED therapy across sites.

5.0 MMM IN GENERALIZED LINEAR MODELS

5.1 LOGISTIC REGRESSION

There has not been much in the literature with regards to mediation in logistic regression models. Huang et al. used structural equation models (SEM) to give a causal interpretation of the meditational effect in the logistic model setting [22]. More recently, MacKinnon et al. presented a comprehensive overview of the difference-in-coefficients, product-of-coefficients, and the proportion mediated effect methods in the case of a binary independent variable [34]. As in the regression and ANOVA cases, no one to this author's knowledge has extended logistic mediation models to the multisite clinical trial setting.

In the logistic regression case, the three necessary equations described in the previous chapter can be extended without difficulty. Instead of being written in regression form, they are presented in terms of the expected values of the outcome y and mediator m, respectively.

$$\text{logit}[\text{E}(y_{ijl})] = \mu_{0\ldots} + \tau_{0i} + s_{0j} + \gamma_{0ij} \tag{5.1}$$

$$\text{E}(m_{ijl}) = \mu_{1\ldots} + \tau_{1i} + s_{1j} + \gamma_{1ij} \tag{5.2}$$

$$\text{logit}[\text{E}(y_{ijl})] = \mu_{2\ldots} + \tau_{2i} + s_{2j} + \gamma_{2ij} + \alpha_1 m_{ijl} \tag{5.3}$$

where τ, s, γ, and m are the effect of treatment, site, their interaction, and the mediator variable, respectively. The subscript ijl represents the l-th person in the j-th site taking the i-th treatment. Unlike in the regression and ANOVA cases, the necessary equations in the logistic case cannot be derived from the bivariate normal distribution. Instead, the equations

are derived from the marginal distribution of m and the conditional distribution of y:

$$(m_{ijl}|\tau, s) \sim N\left(\mu_m, \sigma_1^2\right)$$

where μ_m is (5.2) and

$$(y_{ijl}|m, \tau, s) \sim \text{Bernoulli}(p_{ij})$$

where $p_{ij} = \dfrac{1}{1 + e^{-\mu_{2...} - \tau_{2i} - s_{2j} - \gamma_{2ij} - \alpha_1 m_{ijl}}}$. Therefore the marginal distribution of y is

$$\begin{aligned} f(y_{ijl}|\tau, s) &= \int_{-\infty}^{\infty} f(y|m_{ijl}, \tau, s) f(m_{ijl}|\tau, s) dm \\ &= \int_{-\infty}^{\infty} p_{ij}^{y_{ijl}} (1 - p_{ij})^{1 - y_{ijl}} \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left(\frac{-(m_{ijl} - \mu_m)^2}{2\sigma_1^2}\right) dm, \end{aligned} \tag{5.4}$$

and the marginal expectation of y is

$$\begin{aligned} \mathrm{E}(y_{ijl}|\tau, s) &= \mathrm{E}\left[\mathrm{E}(y_{ijl}|m_{ijl}, \tau, s)\right] = \mathrm{E}(p_{ij}|\tau, s) \\ &= \int_{-\infty}^{\infty} \frac{1}{1 + e^{-w}} \frac{1}{\sqrt{2\pi\sigma_w^2}} \exp\left(\frac{-(w - \mu_w)^2}{2\sigma_w^2}\right) dw, \end{aligned} \tag{5.5}$$

where $w = \mu_{2...} + \tau_{2i} + s_{2j} + \gamma_{2ij} + \alpha_1 m_{ijl}$ is a normal random variable with mean and variance:

$$\begin{aligned} \mu_w &= \mu_{2...} + \tau_{2i} + s_{2j} + \gamma_{2ij} + \alpha_1\mu_{1...} + \alpha_1\tau_{1i} + \alpha_1 s_{1j} + \alpha_1\gamma_{1ij} \\ &= (\mu_{2...} + \alpha_1\mu_{1...}) + (\tau_{2i} + \alpha_1\tau_{1i}) + (s_{2j} + \alpha_1 s_{1j}) + (\gamma_{2ij} + \alpha_1\gamma_{1ij}) \\ &= \mu_{0...} + \tau_{0i} + s_{0j} + \gamma_{0ij} \end{aligned}$$

$$\sigma_w^2 = \alpha_1^2\sigma_1^2.$$

Frederic & Lad [15] showed that p comes from a logitnormal distribution with density

$$f(p_{ij}|\tau, s) = \frac{1}{\sqrt{2\pi\sigma_w^2} p_{ij}(1 - p_{ij})} \exp\left(\frac{-\left(\text{logit}(p_{ij}) - \mu_w\right)^2}{2\sigma_w^2}\right), \tag{5.6}$$

where the logit function is defined as $\text{logit}(x) = \log\left(\frac{x}{1-x}\right)$. With regard to the moments of a logitnormal random variable, unless μ_w is 0, the expected value and variance cannot be obtained in closed form [15].

In order to obtain (5.1), the logit function of (5.5) needs to be linear with respect to the marginal expectation of the mediator variable, μ_m. To investigate the above assertion, we conducted a simulation study that computed the integral in (5.5), using Gauss-Legendre quadrature with 100 knots [1], for varying values of $\mu_w \in [-10, 10]$ and $\sigma_w \in [1, 3, 5]$. For each combination of μ_w and σ_w parameter values, $\mathrm{E}(p)$ and $\mathrm{logit}[\mathrm{E}(p)]$ were computed. Below are separate plots of the expectation and logit versus μ_w.

It's evident from Figure 6 that $\mathrm{E}(p)$, and hence the marginal expectation of y, is sigmoidal regardless of the value of σ_w. Also, all three curves intersect at $\mu_w = 0$ which represents the only closed form for $\mathrm{E}(p)$, which is 0.5. Taking the logit of each of the curves helps to linearize them with the most pronounced effect when $\sigma_w = 1$.

This simulation study shows that the logit of the marginal expectation of y is approximately linear with respect to the marginal expectation of m. In other words:

$$\mathrm{logit}[\mathrm{E}(y_{ijk}|\tau, s)] \approx \mu_{0..} + \tau_{0i} + s_{0j} + \gamma_{0ij}.$$

Similar to the MMM case with a continuous outcome, there is a useful equality that arises although it is now an approximation:

$$\boldsymbol{\gamma}_0 - \boldsymbol{\gamma}_2 \approx \boldsymbol{\gamma}_2 + \alpha_1\boldsymbol{\gamma}_1 - \boldsymbol{\gamma}_2 = \alpha_1\boldsymbol{\gamma}_1. \tag{5.7}$$

An estimate of $\mathbf{d} = \boldsymbol{\gamma}_0 - \boldsymbol{\gamma}_2$ is $\hat{\mathbf{d}} = \hat{\boldsymbol{\gamma}}_0 - \hat{\boldsymbol{\gamma}}_2 \approx a_1\hat{\boldsymbol{\gamma}}_1$ with the following mean and variance:

$$\mathrm{E}(\hat{\mathbf{d}}|\tau, s, m) = \hat{\boldsymbol{\gamma}}_1\alpha_1 \tag{5.8}$$

$$\mathrm{Var}(\hat{\mathbf{d}}|\tau, s, m) = \hat{\boldsymbol{\gamma}}_1\mathrm{Var}(a_1|\tau, s, m)\hat{\boldsymbol{\gamma}}_1'. \tag{5.9}$$

To get the conditional variance of the estimate, a_1, one has to go back to parameter estimation in generalized linear models. A standard result is:

$$\hat{\boldsymbol{\theta}} \sim AN\left(\boldsymbol{\theta}, \mathbf{I}^{-1}(\boldsymbol{\theta})\right),$$

where $\hat{\boldsymbol{\theta}}$ is the vector of parameter estimates and $\mathbf{I}(\boldsymbol{\theta})$ is the respective Fisher Information [36]. In the logistic regression case,

$$\begin{aligned}\mathbf{I}(\boldsymbol{\theta}) &= -\mathrm{E}\left(\frac{\partial^2 \mathrm{logL}}{\partial\theta^2}\right) \\ &= \mathbf{X}'\mathbf{W}\mathbf{X},\end{aligned}$$

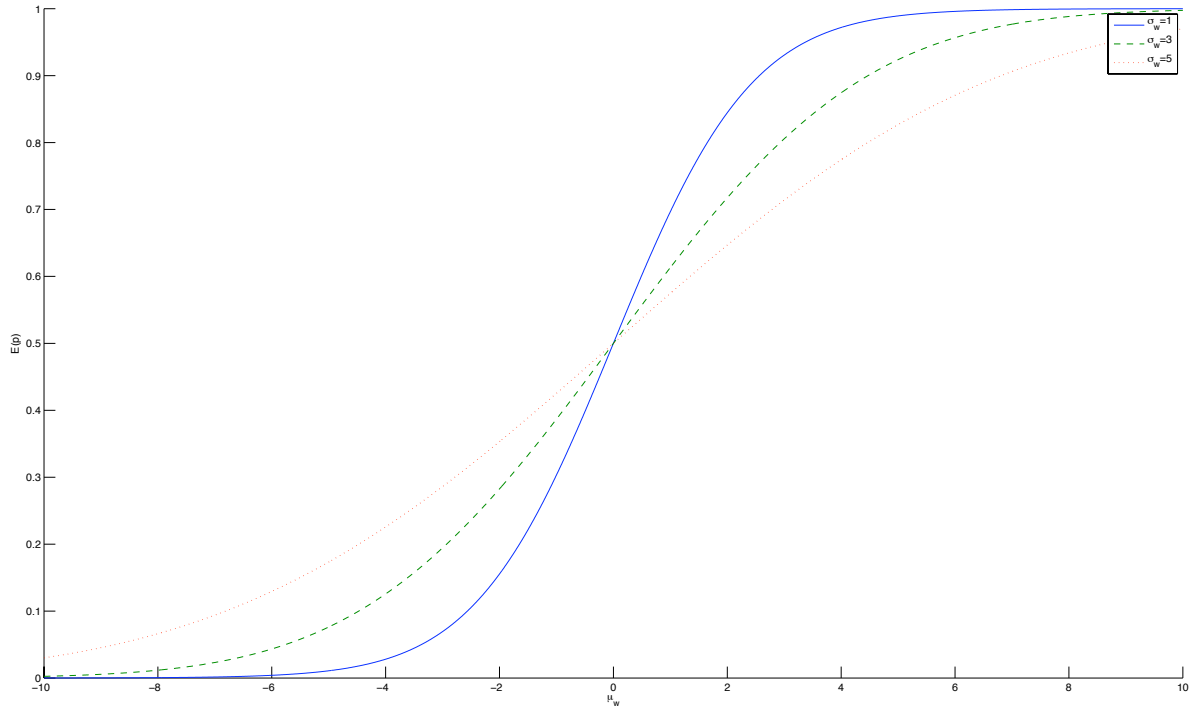

Figure 6: E(p) versus μ_w

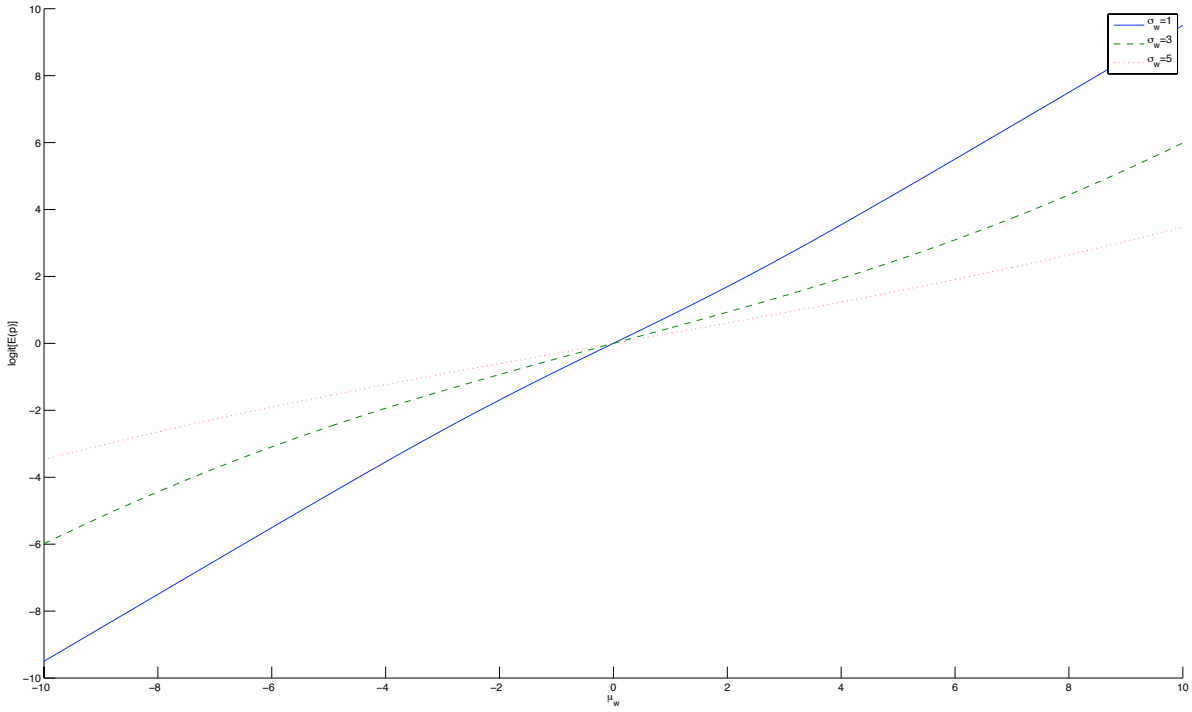

Figure 7: logit[E(p)] versus μ_w

where $\mathbf{W}$ is a diagonal matrix with $p_{ij}(1-p_{ij})$ as the entries. Since $\hat{\boldsymbol{\theta}}' = \begin{bmatrix} \hat{\mu}_{\ldots} & \hat{\tau}_1 & \hat{s}_1 & \ldots & \hat{s}_{J-1} & \hat{\gamma}_{11} & \ldots & \hat{\gamma}_{1J-1} \end{bmatrix}$, then the large sample multivariate distribution of the parameter estimates is asymptotically

$$\hat{\boldsymbol{\theta}} \sim \mathrm{N}\left(\boldsymbol{\theta}, (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\right).$$

5.1.1 Significance Testing with 1 Mediator

We now extend the d-test and PSC test to the logistic regression case. The first test is as follows:

5.1.1.1 d Test

$H_0 : \mathbf{d} = \boldsymbol{\gamma}_0 - \boldsymbol{\gamma}_2 = \mathbf{0}$ with test statistic

$$\begin{aligned} T^2 &= (\hat{\boldsymbol{\gamma}}_1 a_1)' \left[\hat{\boldsymbol{\gamma}}_1 (\mathbf{X}'\mathbf{W}\mathbf{X})_a^{-1} \hat{\boldsymbol{\gamma}}_1{}'\right]^{-1} (\hat{\boldsymbol{\gamma}}_1 a_1) \\ &= a_1^2 \left[(\mathbf{X}'\mathbf{W}\mathbf{X})_a^{-1}\right]^{-1} \left[\hat{\boldsymbol{\gamma}}_1{}'(\hat{\boldsymbol{\gamma}}_1 \hat{\boldsymbol{\gamma}}_1{}')^{-} \hat{\boldsymbol{\gamma}}_1\right] \\ &= \frac{a_1^2}{(\mathbf{X}'\mathbf{W}\mathbf{X})_a^{-1}}, \end{aligned} \tag{5.10}$$

where subscript a denotes the matrix entry corresponding to the parameter estimate of the mediator. Under the null hypothesis, T^2 is distributed as a χ_1^2, so the critical values are chosen such that

$$\alpha = P_{H_0}\left(T^2 > \chi^2_{1,1-\alpha}\right). \tag{5.11}$$

This is also known as a Wald test [36].

5.1.1.2 Product of Standardized Coefficients Test The PSC significance test in the logistic case is similar to that of the ANOVA case: $H_0 : \alpha_1\boldsymbol{\gamma}_1 = \mathbf{0}$, where $\boldsymbol{\gamma}$ is a $J-1$ dimensional vector of the interaction parameters. The product $\alpha_1\boldsymbol{\gamma}_1$ is estimated by $a_1\hat{\boldsymbol{\gamma}}_1$, where

$$a_1 \sim N\left(\alpha_1, (\mathbf{X}'\mathbf{W}\mathbf{X})_a^{-1}\right)$$

and

$$\hat{\boldsymbol{\gamma}}_1 \sim MVN\left(\boldsymbol{\gamma}_1, (\mathbf{X}'\mathbf{W}\mathbf{X})_\gamma^{-1}\right).$$

Following MacKinnon et al, the test statistic is $\mathbf{W} = \frac{a_1}{\sigma_a}\frac{\hat{\boldsymbol{\gamma}}}{\sigma_{\hat{\boldsymbol{\gamma}}}}$, which is distributed under the alternative hypothesis as a scale mixture of multivariate normal random variables with the following conditional and marginal densities:

$$(\mathbf{W}|z_0) \sim MVN\left(\frac{\alpha_1}{\sigma_\alpha}\frac{\boldsymbol{\gamma}}{\sigma_\gamma}, \left(z_0^2 + \left(\frac{\alpha_1}{\sigma_\alpha}\right)^2\right)\Sigma_\gamma + \frac{\boldsymbol{\gamma}}{\sigma_\gamma}\frac{\boldsymbol{\gamma}'}{\sigma_\gamma}\right)$$

where

$$z_0^2 \sim \chi_1^2$$

and Σ_γ is the variance-covariance matrix of the standardized interaction estimates with a diagonal entry of ones and off-diagonal entries:

$$(\Sigma_\gamma)_{ij} = \frac{(X'WX)_{ij}^{-1}}{\sqrt{(X'WX)_{ii}^{-1}}\sqrt{(X'WX)_{jj}^{-1}}}.$$

A Wald test can be conducted with test statistic $T^2 = \mathbf{W}'(z_0^2\hat{\Sigma}_\gamma)^{-1}\mathbf{W}$, and under the null hypothesis $(T^2|z_0^2)$ is distributed as χ_{J-1}^2. Critical values are chosen such that

$$\alpha \;=\; P_{H_0}\left(T^2 > \chi_{J-1,1-\alpha}^2\right). \tag{5.12}$$

5.1.2 Simulation Study with 1 Mediator

To estimate power, a simulation study was conducted for each of the above significance tests. First, the mediator and outcome variables were drawn according to the following distributions:

$$(m_{ijl}|\tau, s) \sim N\left(\mu_{1\ldots} + \tau_{1i} + s_{1j} + \gamma_{1ij}, \sigma_1^2\right)$$

and

$$(y_{ijl}|m, \tau, s) \sim \text{Bernoulli}(p_{ij}),$$

where $p_{ij} = (1 + \exp[-(\mu_{2\ldots} + \tau_{2i} + s_{2j} + \gamma_{2ij} + \alpha_1 m_{ijl})])^{-1}$.

We used the TORDIA clinical trial as a basis for selecting the parameter values in our simulation study. The grand means and treatment effects were $\mu_{1\ldots} = 1.48$, $\mu_{2\ldots} = 1.01$, $\tau_{11} = 0.57$, and $\tau_{21} = -1.24$, respectively. The values for the site parameters in both distributions were $\mathbf{s}_{1j} = \{0.38, -0.10, -0.04, 0.04\}$ and $\mathbf{s}_{2j} = \{-0.79, -1.06, -0.30, -1.66\}$. The interaction values for the distribution of the response variable were
$\boldsymbol{\gamma}_{21j} = \{2.47, 1.62, 1.09, 3.07, 2.95\}$.

Regarding the effect size, we used the approximation from (5.7). The coefficient of the mediator α_1 was set to equal $\sqrt{\text{es}}$, the square root of the effect size. In addition, the interaction parameter values, $\gamma_{11}, \gamma_{12}, \ldots, \gamma_{1,J-1}$, were generated from the same distribution above with the exception that the mean was $\sqrt{\text{es}}$. Effect size values of zero, small (0.20), medium (0.50), and large (0.80) were chosen according to Cohen [10]. The sample sizes were chosen to be 100, 200, 400, 500, and 1000. In addition, the number of sites were chosen to be 2 and 5. In the former case, only the first value was chosen from $\mathbf{s}_{1j}$, $\mathbf{s}_{2j}$, and $\boldsymbol{\gamma}_{21j}$, while all the values were chosen in the latter case. Also, the test statistic for the PSC test, T^2, contains a chi-squared random variable, z_0^2. Therefore, the first, and third quartiles as well as the expected value of χ_1^2 were chosen for values of z_0^2.

For each of the effect size/sample size/site number combinations, the three necessary equations ((5.1), (5.2), (5.3)) were fit and the two significance tests were conducted 10,000 times. The number of times the null hypothesis was rejected gave an estimate of the type I error and power.

5.1.2.1 Results The results of the simulation study for the d test are presented in Tables 8 & 9. With only 2 sites, both the d and PSC tests seem fairly close with regards to power, although it appears the type I error of the PSC is consistently overestimating the true value of 0.05. This phenomenon is similar to the what was happening in the regression and ANOVA cases, although the reason behind it is not the same. One of the things that MacKinnon et al showed was that the useful equality (in the logistic MMM case: $\boldsymbol{\gamma}_0 - \boldsymbol{\gamma}_2 = \alpha_1\boldsymbol{\gamma}_1$) does not hold at times [34]. When the number of sites increases to 5, this issue is even more pronounced – namely in the $n = 100$ and $n = 200$ cases.

Table 8: Type I error & power for logistic MMM with 1 mediator and $j = 2$ sites

		Sample Size				
ES	Method	100	200	400	500	1000
0	d	0.052	0.046	0.048	0.046	0.048
	PSC	0.073	0.065	0.070	0.069	0.076
0.20	d	0.481	0.680	0.928	0.970	>0.999
	PSC	0.465	0.692	0.939	0.968	0.998
0.50	d	0.768	0.930	0.997	0.999	>0.999
	PSC	0.757	0.963	0.999	0.999	>0.999
0.80	d	0.872	0.979	>0.999	>0.999	>0.999
	PSC	0.905	0.998	>0.999	>0.999	>0.999

To investigate this, I conducted a simulation study with the same setup as the preceding one, but only with $j = 2$ sites. For each of the 10,000 iterations, I computed the estimates $\hat{\gamma}_0 - \hat{\gamma}_2$ and $a_1\hat{\gamma}_1$ as well as their squared difference from the true value. The mean, standard deviation, and mean square error were then calculated for each sample size/effect size combination.

As can be seen in Table 10, the product of coefficients does a better job than the difference in coefficients in estimating the true effect. The difference is stark with smaller sample sizes. The fact that $\hat{\gamma}_0 - \hat{\gamma}_2$ severely overestimates the truth (especially in the case of a large effect size and small sample size) can lead to an exaggerated estimate of power. In the $j = 5$ case,

Table 9: Type I error & power for logistic MMM with 1 mediator and $j = 5$ sites

		Sample Size				
ES	Method	100	200	400	500	1000
0	d	0.252	0.068	0.048	0.050	0.052
	PSC	0.337	0.157	0.128	0.155	0.131
0.20	d	0.874	0.818	0.950	0.979	>0.999
	PSC	0.917	0.867	0.969	0.990	>0.999
0.50	d	0.981	0.983	0.999	>0.999	>0.999
	PSC	0.987	0.989	>0.999	>0.999	>0.999
0.80	d	0.993	0.999	>0.999	>0.999	>0.999
	PSC	0.993	0.998	>0.999	>0.999	>0.999

there is an added issue of cell size. When $n = 100$, there are only 10 samples in each of the 10 treatment-by-site cells. As a result, complete separation can occur when all of the responses in a particular cell are identical. Since the PSC test involves a χ^2 random variable, type I error and power were re-estimated with the lower and upper quartiles of χ^2_1 and χ^2_4, respectively. As can be seen in tables 11 & 12, as the χ^2 value decreases, so do the type I error and power.

Table 10: Comparison of estimators in logistic MMM with 1 mediator and $j = 2$ sites

			Sample Size				
ES	Method		100	200	400	500	1000
0	$\hat{\gamma}_0 - \hat{\gamma}_2$	Mean	-0.026	-0.013	-0.007	-0.005	-0.003
		SD	0.129	0.048	0.025	0.020	0.010
		MSE	0.017	0.003	0.001	<0.001	<0.001
	$a_1\hat{\gamma}_1$	Mean	>-0.001	>-0.001	>-0.001	>-0.001	>-0.001
		SD	0.097	0.045	0.023	0.018	0.010
		MSE	0.010	0.002	0.001	<0.001	<0.001
0.20	$\hat{\gamma}_0 - \hat{\gamma}_2$	Mean	0.336	0.151	0.123	0.117	0.102
		SD	0.728	0.340	0.110	0.088	0.059
		MSE	0.558	0.117	0.018	0.014	0.010
	$a_1\hat{\gamma}_1$	Mean	0.181	0.194	0.201	0.199	0.181
		SD	0.255	0.166	0.113	0.100	0.068
		MSE	0.065	0.028	0.013	0.010	0.005
0.50	$\hat{\gamma}_0 - \hat{\gamma}_2$	Mean	1.178	0.800	0.470	0.360	0.340
		SD	1.468	1.040	0.529	0.352	0.116
		MSE	2.772	1.178	0.282	0.136	0.035
	$a_1\hat{\gamma}_1$	Mean	0.430	0.517	0.531	0.476	0.492
		SD	0.408	0.279	0.191	0.162	0.116
		MSE	0.168	0.079	0.037	0.026	0.013
0.80	$\hat{\gamma}_0 - \hat{\gamma}_2$	Mean	2.421	1.997	1.325	0.891	0.646
		SD	2.247	1.682	1.317	0.994	0.438
		MSE	8.001	4.085	1.916	1.002	0.215
	$a_1\hat{\gamma}_1$	Mean	0.796	0.930	0.922	0.788	0.807
		SD	0.576	0.393	0.266	0.222	0.157
		MSE	0.341	0.157	0.071	0.050	0.025

Table 11: Type I error & power for logistic PSC test with varying χ^2 and $j = 2$ sites

		Sample Size				
ES	$\chi^2_{1,\alpha}$	100	200	400	500	1000
0	$\alpha = 0.25$	0.355	0.338	0.353	0.356	0.373
	$\alpha = 0.75$	0.052	0.046	0.049	0.050	0.056
0.20	$\alpha = 0.25$	0.764	0.899	0.987	0.994	>0.999
	$\alpha = 0.75$	0.419	0.645	0.928	0.960	0.998
0.50	$\alpha = 0.25$	0.919	0.992	>0.999	>0.999	>0.999
	$\alpha = 0.75$	0.720	0.950	0.999	0.999	>0.999
0.80	$\alpha = 0.25$	0.978	>0.999	>0.999	>0.999	>0.999
	$\alpha = 0.75$	0.885	0.997	>0.999	>0.999	>0.999

Table 12: Type I error & power for logistic PSC test with varying χ^2 and $j = 5$ sites

		Sample Size				
ES	$\chi^2_{4,\alpha}$	100	200	400	500	1000
0	$\alpha = 0.25$	0.691	0.593	0.578	0.603	0.582
	$\alpha = 0.75$	0.302	0.117	0.090	0.108	0.089
0.20	$\alpha = 0.25$	0.985	0.985	0.999	>0.999	>0.999
	$\alpha = 0.75$	0.900	0.829	0.952	0.982	>0.999
0.50	$\alpha = 0.25$	0.999	0.999	>0.999	>0.999	>0.999
	$\alpha = 0.75$	0.979	0.983	>0.999	>0.999	>0.999
0.80	$\alpha = 0.25$	0.999	>0.999	>0.999	>0.999	>0.999
	$\alpha = 0.75$	0.991	0.997	>0.999	>0.999	>0.999

5.1.3 Significance Testing with K Mediators

The necessary equations from (5.1), (5.2), and (5.3) can be extended to the K mediator case in the following way:

$$\begin{aligned}
\text{logit}[\text{E}(y_{ijl})] &= \mu_{0\ldots} + \tau_{0i} + s_{0j} + \gamma_{0ij} && (5.13)\\
\text{E}(m_{1ijl}) &= \mu_{1\ldots} + \tau_{1i} + s_{1j} + \gamma_{1ij} && (5.14)\\
&\vdots \\
\text{E}(m_{Kijl}) &= \mu_{K\ldots} + \tau_{Ki} + s_{Kj} + \gamma_{Kij} && (5.15)\\
\text{logit}[\text{E}(y_{ijl})] &= \mu_{K+1\ldots} + \tau_{K+1,i} + s_{K+1,j} + \gamma_{K+1,ij} + \alpha_1 m_{1ijl} + \cdots + \alpha_{K+1} m_{K+1,ijl}. && (5.16)
\end{aligned}$$

In terms of the multivariate distribution of the mediators, $m_1, \cdots, m_K$, and the conditional distribution of the outcome variable, we have

$$\begin{pmatrix} m_1 \\ \vdots \\ m_k \end{pmatrix} \sim \text{N}\left\{ \begin{pmatrix} \mu_{1\ldots} + \tau_{1i} + s_{1j} + \gamma_{1ij} \\ \vdots \\ \mu_{K\ldots} + \tau_{Ki} + s_{Kj} + \gamma_{Kij} \end{pmatrix}, \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \ldots & \sigma_{1K} \\ \sigma_{21} & \sigma_2^2 & \ldots & \sigma_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{K1} & \sigma_{K2} & \ldots & \sigma_K^2 \end{bmatrix} \right\},$$

and

$$(y_{ijl}|m_1, \cdots, m_K, \tau, s) \sim \text{Bernoulli}(p_{ij}),$$

where $p_{ij} = \dfrac{1}{1 + e^{-(\mu_{K+1\ldots} + \tau_{K+1,i} + s_{K+1,j} + \gamma_{K+1,ij} + \alpha_1 m_{1ijl} + \cdots + \alpha_{K+1} m_{K+1,ijl})}}$. As in the single mediator case, the density of the marginal distribution of y is

$$\begin{aligned}
f(y|\tau, s) &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(y|\mathbf{m}, \tau, s) f(\mathbf{m}|\tau, s) d\mathbf{m} \\
&= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} p^y (1-p)^{1-y} (2\pi)^{-K/2} |\Sigma_{\mathbf{m}}|^{-1/2} \exp\left(\frac{-1}{2} (\mathbf{m} - \boldsymbol{\mu}_{\mathbf{m}})' \Sigma_{\mathbf{m}}^{-1} (\mathbf{m} - \boldsymbol{\mu}_{\mathbf{m}}) \right) d\mathbf{m}. \qquad (5.17)
\end{aligned}$$

From above, we can obtain the marginal expectation of the outcome:

$$\begin{aligned}
\text{E}(y|\tau, s) &= \text{E}\left[\text{E}(y|\mathbf{m}, \tau, s)\right] = \text{E}(p|\tau, s) \\
&= \int_{-\infty}^{\infty} \frac{1}{1 + e^{-w}} \frac{1}{\sqrt{2\pi\sigma_w^2}} \exp\left(\frac{-(w - \mu_w)^2}{2\sigma_w^2} \right) dw, \qquad (5.18)
\end{aligned}$$

where $w = \mu_{K+1\ldots} + \tau_{K+1,i} + s_{K+1,j} + \gamma_{K+1,ij} + \sum_{q=1}^{K} \alpha_q m_{qijl}$ is a normal random variable with mean and variance:

$$\begin{aligned}
\mu_w &= \mu_{K+1\ldots} + \tau_{K+1,i} + s_{K+1,j} + \gamma_{K+1,ij} + \sum_{q=1}^{K} \alpha_q \mu_{q\ldots} + \sum_{q=1}^{K} \alpha_q \tau_{qi} + \sum_{q=1}^{K} \alpha_q s_{qj} + \sum_{q=1}^{K} \alpha_q \gamma_{qij} \\
&= (\mu_{K+1\ldots} + \sum_{q=1}^{K} \alpha_q \mu_{q\ldots}) + (\tau_{K+1,i} + \sum_{q=1}^{K} \alpha_q \tau_{qi}) + (s_{K+1,j} + \sum_{q=1}^{K} \alpha_q s_{qj}) + (\gamma_{K+1,ij} + \sum_{q=1}^{K} \alpha_q \gamma_{qij}) \\
&= \mu_{0\ldots} + \tau_{0i} + s_{0j} + \gamma_{0ij}
\end{aligned}$$

$$\sigma_w^2 = \sum_{q=1}^{K} \alpha_q^2 \sigma_q^2 + 2 \sum_{q \neq q'}^{K} \alpha_q \alpha_{q'} \sigma_{qq'}.$$

If we adhere to the linearity assumption between the marginal expectation of y and the marginal expectation of m that was shown in the single mediator case, the following approximation arises:

$$\boldsymbol{\gamma}_0 - \boldsymbol{\gamma}_{K+1} \approx \sum_{q=1}^{K} \alpha_q \boldsymbol{\gamma}_q. \tag{5.19}$$

An estimate of $\mathbf{d} = \boldsymbol{\gamma}_0 - \boldsymbol{\gamma}_{K+1}$ is $\hat{\mathbf{d}} = \hat{\boldsymbol{\gamma}}_0 - \hat{\boldsymbol{\gamma}}_{K+1} \approx \sum_{q=1}^{K} a_q \hat{\boldsymbol{\gamma}}_q$ with the following mean and variance:

$$\mathrm{E}(\hat{\mathbf{d}}|\tau, s, \mathbf{m}) = \sum_{q=1}^{K} \hat{\boldsymbol{\gamma}}_q \alpha_q \tag{5.20}$$

$$\mathrm{Var}(\hat{\mathbf{d}}|\tau, s, \mathbf{m}) = \sum_{q=1}^{K} \hat{\boldsymbol{\gamma}}_q (\mathbf{X}'\mathbf{W}\mathbf{X})_{a_q} \hat{\boldsymbol{\gamma}}_q' + 2 \sum_{q \neq q'}^{K} \hat{\boldsymbol{\gamma}}_q (\mathbf{X}'\mathbf{W}\mathbf{X})_{a_q, a_{q'}} \hat{\boldsymbol{\gamma}}_{q'}'. \tag{5.21}$$

5.1.3.1 *d* **Test** The null hypothesis is $H_0 : \mathbf{d} = \boldsymbol{\gamma}_0 - \boldsymbol{\gamma}_{K+1} = \mathbf{0}$ with test statistic

$$T^2 = \left(\sum_{q=1}^{K} a_q \hat{\boldsymbol{\gamma}}_q \right)' \left[\sum_{q=1}^{K} \hat{\boldsymbol{\gamma}}_q (\mathbf{X}'\mathbf{W}\mathbf{X})_{a_q} \hat{\boldsymbol{\gamma}}_q' + 2 \sum_{q \neq q'}^{K} \hat{\boldsymbol{\gamma}}_q (\mathbf{X}'\mathbf{W}\mathbf{X})_{a_q, a_{q'}} \hat{\boldsymbol{\gamma}}_{q'}' \right]^{-1} \left(\sum_{q=1}^{K} a_q \hat{\boldsymbol{\gamma}}_q \right), \tag{5.22}$$

where subscript a_q denotes the matrix entry corresponding to the parameter estimate of the q-th mediator, and $a_q, a_{q'}$ denotes the off-diagonal entry corresponding to the estimates of the mediators, m_q and $m_{q'}$. Under the null hypothesis, T^2 is distributed as a χ_1^2, so the critical values are chosen such that

$$\alpha = P_{H_0}\left(T^2 > \chi_{1,1-\alpha}^2\right). \tag{5.23}$$

5.1.3.2 Product of Standardized Coefficients Test

The null hypothesis for the PSC significance test is $H_0 = \sum_{q=1}^{K} \alpha_q \boldsymbol{\gamma_q} = \mathbf{0}$ with test statistic $\sum_{q=1}^{K} \mathbf{V}_q = \sum_{q=1}^{K} \left(\frac{a_q}{\sigma_{a_q}} \frac{\hat{\boldsymbol{\gamma}}_q}{\sigma_{\hat{\gamma}_q}} \right)$. Using the normal scale mixture theory from the previous chapter, we get that $\sum_{q=1}^{K} \mathbf{V}_q$ is distributed according to a multivariate normal distribution with mean and variance:

$$\mathrm{E}\left(\sum_{q=1}^{K} \mathbf{V}_q \right) = \sum_{q=1}^{K} \left(\frac{\alpha_q}{\sigma_{\alpha_q}} \frac{\boldsymbol{\gamma}_q}{\sigma_{\gamma_q}} \right) \tag{5.24}$$

$$\tag{5.25}$$

$$\begin{aligned} \mathrm{Var}\left(\sum_{q=1}^{K} \mathbf{V}_q \right) &= \sum_{q=1}^{K} \mathrm{Var}(\mathbf{V}_q) + 2 \sum_{q \neq q'}^{K} \mathrm{Cov}(\mathbf{V}_q, \mathbf{V}_{q'}) \\ &= \sum_{q=1}^{K} \left[\left(z_{a_q}^2 + \left(\frac{\alpha_q}{\sigma_{\alpha_q}} \right)^2 \right) \Sigma_{\gamma_q} + \frac{\boldsymbol{\gamma}_q}{\sigma_{\gamma_q}} \frac{\boldsymbol{\gamma}_q}{\sigma_{\gamma_q}}' \right] \\ &\quad + 2 \sum_{q \neq q'}^{K} \left[\mathrm{Cov}\left(\frac{\hat{\boldsymbol{\gamma}}_q}{\sigma_{\hat{\gamma}_q}}, \frac{\hat{\boldsymbol{\gamma}}_{q'}}{\sigma_{\hat{\gamma}_{q'}}} \right) \left(|z_{a_q} z_{a_{q'}}| + |z_{a_q}| \left(\frac{\alpha_{q'}}{\sigma_{\alpha_{q'}}} \right) + |z_{a_{q'}}| \left(\frac{\alpha_q}{\sigma_{\alpha_q}} \right) \right) \right], \end{aligned} \tag{5.26}$$

where $|z_{a_q}|$ and $|z_{a_{q'}}|$ are folded normal random variables and $z_{a_q}^2$ is a chi squared random variable.

5.1.4 Illustration on TORDIA data

The two significance tests – the d and PSC – were conducted on the TORDIA clinical trial dataset with the original binary outcome of clinical response. As mentioned before, the primary outcome was defined as the combination of the Clinical Global Impressions score ≤ 2 and a change in the Children's Depression Rating Scale-Revised of $\geq 50\%$.

The three models in the multisite mediated moderation case ((5.1), (5.2), and (5.3)) were conducted with 20 variables measured at baseline as potential mediators. There was overall site moderation of treatment effect in (4.52) ($p = 0.001$) at the $\alpha = 0.05$ significance level. Since the second necessary equation did not change going from the continuous outcome to

the binary outcome case, treatment-by-site interaction persisted in the same two variables: CBQA ($p = 0.044$) and TOTALD ($p = 0.004$).

After adjusting for the CBQA score, the magnitude of site moderation was reduced ($p = 0.006$), which is a criteria for mediated moderation. Both tests were applied to the data to see if the reductions in magnitudes were significant. Since the PSC test involves a χ^2_1 random variable, the expected value of 1 was chosen. The d test concluded that subject's CBQA score significantly explained the differing effect of CBT-MED therapy across the six sites ($p = 0.002$). The Product of Standardized Coefficients test concluded that the magnitude of interaction was significantly reduced ($p < 0.001$).

The magnitude of site moderation due to the adjustment for the TOTALD variable was reduced, but not eliminated ($p = 0.003$). Both of the aforementioned tests were applied to the data and only the PSC test concluded a significant reduction ($p_d = 0.173; p_{PSC} < 0.001$).

6.0 DISCUSSION & FUTURE WORK

6.1 DISCUSSION

In summary, this dissertation has focused on the issue of site heterogeneity in both the design and analysis stages of multisite clinical trials. While it is ideal to account for potential sources at the design stage of a trial, this is only possible if there is statistical methodology to identify them at the analysis end. The contribution of this dissertation is exactly that.

Using the concept of mediated moderation described by Muller et al. [39], three tests popular in the mediation literature were extended to the multiple regression, ANOVA, and logistic regression models used to analyze multisite clinical trials. The result is a battery of significance tests to help explain treatment-by-site interaction. Each of the test were developed in the two-treatment, two-site case and were extended to the J-site case.

The significance tests in the multiple regression and logistic regression cases were also broadened to include K meditators. While the groundwork has already been laid, some care is needed in the details of the PSC tests. This will be the focus of an upcoming paper.

Finally, once potential sources of site heterogeneity are chosen by the above tests, they can inform the design for future clinical trials. Instead of basing sample size and power off of a t-test or an ANOVA model, one can use the hierarchical linear models – (2.19) and (2.20) – described in Chapter 2 with the potential sources as site-level covariates. One of the conclusions mentioned in Spirito et al. was that multisite clinical trials should be powered to detect site differences [47]. While the author agrees with this, it should be used if site variability cannot be accounted for by methods such as HLMs mentioned above.

6.2 FUTURE WORK

The author plans to investigate the following:

1. Study the details of the PSC and d tests (multiple regression and logistic regression) with K mediators. Conduct simulation studies and apply to the TORDIA dataset.
2. A generalization of MMM to models in the generalized linear model framework.
3. Multisite clinical trials are not just limited to two treatments, so the extension of the proposed MMM to a model with three or more treatments is needed.
4. All of the difference in coefficients test that have been outlined have treated the mediator(s) variable as fixed, whereas in the clinical trial setting the mediators vary. Accounting for this added variability will provide more accurate estimates of power.
5. Extend the MMM to the case where site is a random effect; explore different possibilities for the random distribution such as the skew-t distribution as well as Bayesian nonparametric approaches.
6. As in the two-site case, investigate the type I error inflation issue in the PSC tests.

APPENDIX A

DERIVATION OF THE EQUALITY IN MMM

It is straightforward to show that (4.8) and (4.9) lead to (4.7) in the two-site, two-treatment case. This leads to the useful equality seen in mediated moderation.

$$m = \beta_{10} + \beta_{11}\tau + \beta_{12}s + \beta_{13}(\tau * s) + \epsilon_1 \tag{A.1}$$

$$\begin{aligned} y &= \alpha_{00} + \alpha_{01}\tau + \alpha_{02}s + \alpha_{03}(\tau * s) + \alpha_{12}m + \alpha_{13}(m * \tau) + \epsilon_2 \\ &= (\alpha_{00} + \alpha_{01}\tau) + (\alpha_{02} + \alpha_{03}\tau)s + (\alpha_{12} + \alpha_{13}\tau)(\beta_{10} + \beta_{11}\tau) \\ &\quad +(\alpha_{12} + \alpha_{13}\tau)(\beta_{12} + \beta_{13}\tau)s + (\alpha_{12} + \alpha_{13}\tau)\epsilon_1 + \epsilon_2 \\ &= [(\alpha_{00} + \alpha_{01}\tau) + (\alpha_{12} + \alpha_{13}\tau)(\beta_{10} + \beta_{11}\tau)] \\ &\quad + [(\alpha_{02} + \alpha_{03}\tau) + (\alpha_{12} + \alpha_{13}\tau)(\beta_{12} + \beta_{13}\tau)]\, s + [(\alpha_{12} + \alpha_{13}\tau)\epsilon_1 + \epsilon_2] \\ &= [\beta_{00} + \beta_{01}\tau] + [\beta_{02} + \beta_{03}\tau]\, s + \epsilon_0. \end{aligned} \tag{A.2}$$

From this, we get

$$\begin{aligned} (\beta_{02} + \beta_{03}\tau) &= (\alpha_{02} + \alpha_{03}\tau) + (\alpha_{12} + \alpha_{13}\tau)(\beta_{12} + \beta_{13}\tau) \\ (\beta_{02} + \beta_{03}\tau) - (\alpha_{02} + \alpha_{03}\tau) &= (\alpha_{12} + \alpha_{13}\tau)(\beta_{12} + \beta_{13}\tau) \end{aligned} \tag{A.3}$$

which implies

$$\beta_{03} - \alpha_{03} = \alpha_{13}\beta_{12} + \alpha_{12}\beta_{13}. \tag{A.4}$$

Muller et al. requires that at least one of the products on the right hand side of (A.4) be non-zero [39]. So, by allowing the partial effect to not be moderated, no assumptions of mediated moderation are violated.

APPENDIX B

GAUSS-HERMITE QUADRATURE

If $x \sim N(\mu_x, 1)$ and $y \sim N(\mu_y, 1)$, then the pdf of $p = xy$ is

$$f(p) = (2\pi)^{-1} \int_{-\infty}^{\infty} \frac{1}{y} e^{\frac{-1}{2}\left(\frac{p}{y} - \mu_x\right)^2} e^{\frac{-1}{2}(y-\mu_y)^2} dy. \tag{B.1}$$

When $\mu_x = \mu_y = 0$, the above integral is approximated by a modified Bessel function of the second kind [12]. Otherwise, Gauss-Hermite quadrature is one way of approximating the integral such that

$$\int_{-\infty}^{\infty} e^{-x^2} f(x) \approx \sum_{i=1}^{n} w(x_i) f(x_i), \tag{B.2}$$

where $w(x_i)$ is a weight function evaluated at the i-th root of a Hermite polynomial. The total number of nodes, $n = 32$, was chosen based on how well the quadrature approximated the modified Bessel function in the case where both parameters are null.

APPENDIX C

CONDITIONAL MEAN OF Y IN THE K-MEDIATOR CASE

Let the joint distribution of the outcome variable, y, and the mediator variables, m_1,...,m_k, be

$$\begin{pmatrix} y \\ m_1 \\ \vdots \\ m_k \end{pmatrix} \sim \mathrm{N}\left\{ \begin{pmatrix} \beta_{00} + \beta_{01}\tau + \beta_{02}s + \beta_{03}(\tau * s) \\ \beta_{10} + \beta_{11}\tau + \beta_{12}s + \beta_{13}(\tau * s) \\ \vdots \\ \beta_{k0} + \beta_{k1}\tau + \beta_{k2}s + \beta_{k3}(\tau * s) \end{pmatrix}, \begin{bmatrix} \sigma_0^2 & \sigma_{01} & \dots & \sigma_{0k} \\ \sigma_{10} & \sigma_1^2 & \dots & \sigma_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{k0} & \sigma_{k1} & \dots & \sigma_k^2 \end{bmatrix} \right\}.$$

If we partition y from the mediator variables such that the variance-covariance matrix is

$$\Sigma_{ym} = \begin{bmatrix} \sigma_0^2 & \Sigma_{0m} \\ \Sigma_{m0} & \Sigma_{mm} \end{bmatrix},$$

then the conditional distribution of y is

$$(y|m_1, \dots, m_k, \tau, s) \sim \mathrm{N}\left(\mu_y + \Sigma_{0m}\Sigma_{mm}^{-1}(\mathbf{m} - \boldsymbol{\mu}_m), \sigma_0^2 - \Sigma_{0m}\Sigma_{mm}^{-1}\Sigma_{m0}\right).$$

The term $\Sigma_{0m}\Sigma_{mm}^{-1}$ is referred to as the matrix of *regression coefficients* [24]. Then, we can write the conditional mean as

$$\begin{aligned}
\mu_{y|m_1,\ldots,m_k,\tau,s} &= \mu_y + \begin{bmatrix} \alpha_{12} & \alpha_{22} & \ldots & \alpha_{k2} \end{bmatrix} \begin{bmatrix} m_1 - \mu_{m_1} \\ m_2 - \mu_{m_2} \\ \vdots \\ m_k - \mu_{m_k} \end{bmatrix} \\
&= (\beta_{00} + \beta_{01}\tau + \beta_{02}s + \beta_{03}(\tau * s)) + \sum_{i=1}^{k} \alpha_{i2} \left(m_i - \beta_{i0} + \beta_{i1}\tau + \beta_{i2}s + \beta_{i3}(\tau * s)\right) \\
&= \left(\beta_{00} - \sum_{i=1}^{k} \beta_{i0}\alpha_{i2}\right) + \left(\beta_{01} - \sum_{i=1}^{k} \beta_{i1}\alpha_{i2}\right)\tau + \left(\beta_{02} - \sum_{i=1}^{k} \beta_{i2}\alpha_{i2}\right)s \\
&\quad + \left(\beta_{03} - \sum_{i=1}^{k} \beta_{i3}\alpha_{i2}\right)(\tau * s).
\end{aligned} \tag{C.1}$$

APPENDIX D

DERIVATION OF COVARIANCE OF INTERACTION EFFECT ESTIMATES

In a 2-treatment, J-site, unbalanced ANOVA design with the following model, the $J-1$ interaction effect estimates $\hat{\gamma}_{ij} = \bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...}$ are correlated with each other.

$$Y_{ijk} = \mu + \tau_i + s_j + \gamma_{ij} + \epsilon_{ijk} \tag{D.1}$$

Let $\hat{\gamma}_{ij}$ and $\hat{\gamma}_{ij'}$ be two estimates from the same treatment, but different sites. Then, the covariance is as follows:

$$\begin{aligned}
\mathrm{Cov}(\hat{\gamma}_{ij}, \hat{\gamma}_{ij'}) &= \mathrm{Cov}(\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...}, \bar{y}_{ij'.} - \bar{y}_{i..} - \bar{y}_{.j'.} + \bar{y}_{...}) \\
&= -\mathrm{Cov}(\bar{y}_{ij.}, \bar{y}_{i..}) + \mathrm{Cov}(\bar{y}_{ij.}, \bar{y}_{...}) - \mathrm{Cov}(\bar{y}_{i..}, \bar{y}_{ij'.}) + \mathrm{Cov}(\bar{y}_{i..}, \bar{y}_{i..}) + \mathrm{Cov}(\bar{y}_{i..}, \bar{y}_{.j'.}) \\
&\quad -\mathrm{Cov}(\bar{y}_{i..}, \bar{y}_{...}) + \mathrm{Cov}(\bar{y}_{.j.}, \bar{y}_{i..}) - \mathrm{Cov}(\bar{y}_{.j.}, \bar{y}_{...}) + \mathrm{Cov}(\bar{y}_{...}, \bar{y}_{ij'.}) - \mathrm{Cov}(\bar{y}_{...}, \bar{y}_{i..}) \\
&\quad -\mathrm{Cov}(\bar{y}_{...}, \bar{y}_{.j'.}) + \mathrm{Cov}(\bar{y}_{...}, \bar{y}_{...}) \\
&= -\frac{\sigma^2}{Jn_{ij}} + \frac{\sigma^2}{IJn_{ij}} - \frac{\sigma^2}{Jn_{ij'}} + \frac{\sigma^2}{J^2}\sum_{j=1}^{J}\left(\frac{1}{n_{ij}}\right) + \frac{\sigma^2}{IJn_{ij'}} - \frac{\sigma^2}{IJ^2}\sum_{j=1}^{J}\left(\frac{1}{n_{ij}}\right) \\
&\quad + \frac{\sigma^2}{IJn_{ij}} - \frac{\sigma^2}{I^2J}\sum_{i=1}^{I}\left(\frac{1}{n_{ij}}\right) + \frac{\sigma^2}{IJn_{ij'}} - \frac{\sigma^2}{IJ^2}\sum_{j=1}^{J}\left(\frac{1}{n_{ij}}\right) - \frac{\sigma^2}{I^2J}\sum_{i=1}^{I}\left(\frac{1}{n_{ij'}}\right) \\
&\quad + \frac{\sigma^2}{I^2J^2}\sum_{i=1}^{I}\sum_{j=1}^{J}\left(\frac{1}{n_{ij}}\right) \\
&= \sigma^2\left[\frac{1}{4J^2}\sum_{i=1}^{2}\sum_{j=1}^{J}\left(\frac{1}{n_{ij}}\right) - \frac{1}{4J}\sum_{i=1}^{2}\left(\frac{1}{n_{ij}} + \frac{1}{n_{ij'}}\right)\right].
\end{aligned} \tag{D.2}$$

The variance of a particular interaction effect estimate is

$$\begin{aligned}
\mathrm{Var}(\hat{\gamma}_{ij}) &= \mathrm{Var}(\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...}) \\
&= \frac{\sigma^2}{n_{ij}} + \frac{\sigma^2}{J^2}\sum_{j=1}^{J}\left(\frac{1}{n_{ij}}\right) + \frac{\sigma^2}{I^2}\sum_{i=1}^{I}\left(\frac{1}{n_{ij}}\right) + \frac{\sigma^2}{I^2J^2}\sum_{i=1}^{I}\sum_{j=1}^{J}\left(\frac{1}{n_{ij}}\right) \\
&\quad +2\left[-\frac{\sigma^2}{Jn_{ij}} - \frac{\sigma^2}{In_{ij}} + \frac{\sigma^2}{IJn_{ij}} + \frac{\sigma^2}{IJn_{ij}} - \frac{\sigma^2}{IJ^2}\sum_{j=1}^{J}\left(\frac{1}{n_{ij}}\right) - \frac{\sigma^2}{I^2J}\sum_{i=1}^{I}\left(\frac{1}{n_{ij}}\right)\right]. \\
&= \sigma^2\left[\frac{J-2}{4J}\sum_{i=1}^{2}\left(\frac{1}{n_{ij}}\right) + \frac{1}{4J^2}\sum_{i=1}^{2}\sum_{j=1}^{J}\left(\frac{1}{n_{ij}}\right)\right]. \qquad \text{(D.3)}
\end{aligned}$$

APPENDIX E

GENERALIZED INVERSE

Theorem 4. *Let* $\mathbf{A}$ *be a* p*-by-*1 *dimensional vector, and let* $(\mathbf{AA}')$ *be a square singular matrix with generalized inverse* $(\mathbf{AA}')^-$ *defined in McCulloch & Searle [36]. Then,* $\mathbf{A}'(\mathbf{AA}')^-\mathbf{A} = 1$.

Proof. We know that

$$\mathbf{A}'(\mathbf{AA}')^-\mathbf{A} = c, \tag{E.1}$$

where c is a scalar. If we pre- and post-multiply by $\mathbf{A}$ and $\mathbf{A}'$, respectively, we get

$$(\mathbf{AA}')(\mathbf{AA}')^-(\mathbf{AA}') = \mathbf{A}c\mathbf{A}'. \tag{E.2}$$

Because $(\mathbf{AA}')^-$ is a generalized inverse, then $c = 1$ must be true. □

BIBLIOGRAPHY

[1] Abramowitz, M. and Stegun, I.A. *Handbook of Mathematical Functions.* Dover, 1972.

[2] Agresti, A. and Hartzel, J. Tutorial in Biostatistics: Strategies for comparing treatments on a binary response with multi-centre data. *Statistics in Medicine*, 19:1115–1139, 2000.

[3] Aiken L. and West, S. *Multiple Regression: Testing and Interpreting Interactions.* Sage, 1991.

[4] C. A. Azzalini, A. Statistical applications of the multivariate skew-normal distribution. *Journal of the Royal Statistical Society*, 61:579–602, 1999.

[5] Azzalini, A. and Capitanio, A. Distributions generated by perturbation of symmetry with emphasis on a multivariate skew-t distribution. *Journal of the Royal Statistical Society*, 65:367–389, 2003.

[6] Baron, R. and Kenny, D. The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51(6):1173–1182, 1986.

[7] Brent, D., Emslie, G., Clarke, G., Wagner, K., Asarnow, J., Keller, M., Vitiello, B., Ritz, L., Iyengar, S., Abebe, K., Birmaher, B., Ryan, N., Kennard, B., Hughes, C., DeBar, L., McCracken, J., Strober, M., Suddath, R., Spirito, A., Leonard, H., Mehlem, N., Porta, G., Onorato, M., and Zelazny, J. Switching to venlafaxine or another SSRI with or without cognitive behavioral therapy for adolescents with SSRI-resistant depression: The TORDIA randomized control trial. *Journal of the American Medical Association*, 299(8):901–913, 2008.

[8] Bridge, J., Iyengar, S., Salary, C., Barbe, R., Birmaher, B., Pincus, H., Ren, L., and Brent, D. Clinical response and risk for reported suicidal ideation and suicide attempts in pediatric antidepressant treatment: A meta-analysis of randomized controlled trials. *Journal of the American Medical Association*, 297:1683–1696, 2007.

[9] Casella, G. and Berger, R. *Statistical Inference.* Duxbury, 2nd edition, 2002.

[10] Cohen, J. *Statistical Power Analysis for the Behavioral Sciences.* Lawrence Erlbaum Associates, 2nd edition, 1988.

[11] Cooper, H. and Hedges, L., editors. *The Handbook of Research Synthesis*. Sage, 1994.

[12] Craig, C. On the frequency function of xy. *Annals of Mathematical Statistics*, 7:1–15, 1936.

[13] Elandt, R.C. The folded normal distribution: Two methods of estimating parameters from moments. *Technometrics*, 3(4):551–562, 1961.

[14] Fleiss, J. *The Design and Analysis of Clinical Experiments*. Wiley, 1985.

[15] Frederic, P. and Lad, F. Two moments of the logitnormal distribution. *Communications in Statistics - Simulation and Computation*, 37:1263–1269, 2008.

[16] Freedman, L. and Schatzkin, A. Sample size for studying intermediate endpoints within intervention trials of observational studies. *American Journal of Epidemiology*, 136:1148–1159, 1992.

[17] Gallo, P. Center-weighting issues in multicenter clinical trials. *Journal of Biopharmaceutical Statistics*, 10(2):145–163, 2000.

[18] Gupta, S. and Perlman, M. Power of the noncentral f-test: Effect of additional variates on hotelling's t-test. *Journal of the American Statistical Association*, 69(345):174–180, 1974.

[19] Hedges, L. Issues in meta-analysis. *Review of Research in Education*, 13:353–398, 1986.

[20] Hedges, L. and Olkin, I. *Statistical Methods for Meta-Analysis*. Academic Press, 1985.

[21] Hoyle, M.H. Transformations: An introduction and a bibliography. *International Statistical Review*, 41(2):203–223, 1973.

[22] Huang, B., Sivaganesan, S., Succop, P., and Goodman, E. Statistical assessment of mediational effects for logistic mediational models. *Statistics in Medicine*, 23:2713–2728, 2004.

[23] ICH E9 Expert Working Group. Statistical principles for clinical trials: ICH Harmonised Tripartitite Guideline. *Statistics in Medicine*, 18:1905–1942, 1999.

[24] Johnson, R. and Wichern, D. *Applied Multivariate Statistical Analysis*. Prentice Hall, 5th edition, 2002.

[25] Judd, C. and Kenny, D. Process analysis: Estimating mediation in treatment evaluations. *Evaluation Review*, 5:602–619, 1981.

[26] Kraemer, H. *Evaluating medical tests: Objective and Quantitative Guidelines*. Sage, 1992.

[27] Kraemer, H. Pitfalls of multisite randomized clinical trials of efficacy and effectiveness. *Schizophrenia Bulletin*, 26(3):533–541, 2000.

[28] Kraemer, H., Frank, E., and Kupfer, D. Moderators of treatment outcomes: Clinical, research, and policy importance. *Journal of the American Medical Association*, 296(10):1286–1289, 2006.

[29] Kraemer, H. and Robinson, T. Are certian multicenter randomized clinical trial structures misleading clinical and policy decision? *Contemporary Clinical Trials*, 26:518–529, 2005.

[30] Kraemer, H., Wilson, G., Fairbun, C., and Agras, W. Mediators and moderators of treatment effects in randomized clinical trials. *Archives of General Psychiatry*, 59:877–883, 2002.

[31] Kutner, M., Nachtsheim, C., Neter, J., and Li, W. *Applied Linear Statistical Models*. McGraw-Hill, 5th edition, 2005.

[32] Lehmann, E.L. *Elements of Large-Sample Theory*. Springer, 1999.

[33] Lin, Z. An issue of statistical analysis in controlled multi-centre studies: How shall we weight the centres? *Statistics in Medicine*, 18:365–373, 1999.

[34] MacKinnon, D., Lockwood, C., Brown, C., Wang, W., and Hoffman, J. The intermediate endpoint effect in logistic and probit regression. *Clinical Trials*, 4:499–513, 2007.

[35] MacKinnon, D., Lockwood, C., Hoffman, J., West, S., and Sheets, V. A comparison of methods to test mediation and other intervening variable effects. *Psychological Methods*, 7(1):83–104, 2002.

[36] McCulloch, C. and Searle, S. *Generalized, Linear, and Mixed Models*. Wiley, 2001.

[37] McGaw, B. and Glass, G. Choice of metric for effect size in meta-analysis. *American Educational Research Journal*, 17(3):325–337, 1980.

[38] Meinert, C. *Clinical trials: design, conduct, and analysis*. Oxford University, 1986.

[39] Muller, D., Judd, C., and Yzerbyt, V. When moderation is mediated and mediation is moderated. *Journal of Personality and Social Psychology*, 89(6):852–863, 2005.

[40] Olkin, I. and Soitani, M. Asymptotic distribution of functions of a correlation matrix. In Ikeda, S., editor, *Essays in probability and statistics*, pages 235–251. Shinko Tsusho, 1976.

[41] Raudenbush, S. and Bryk, A. *Hierarchical Linear Models: Applications and Data Analysis Methods*. Sage, 2nd edition, 2002.

[42] Raudenbush, S. and Liu, X. Statistical power and optimal design for multisite randomized trials. *Psychological Methods*, 5(2):199–213, 2000.

[43] Raveh, A. On the use of the inverse of the correlation matrix in multivariate data analysis. *The American Statistician*, 39(1):39–42, 1985.

[44] Robinson, W. Ecological correlations and the behavior of individuals. *American Sociological Review*, 15:351–357, 1950.

[45] Schwemer, G. General linear models for multicenter clinical trials. *Controlled Clinical Trials*, 21:21–29, 2000.

[46] Senn, S. Some controversies in planning and analysing multi-center trials. *Statistics in Medicine*, 17:1753–1765, 1998.

[47] Spirito, A., Abebe, K., Keller, M., Iyengar, S., Vitiello, B., Clarke, G., Wagner, K., Brent, D., Asarnow, J., and Emslie, G. Sources of site differences in the efficacy of a multi-site clinical trial: The treatment of SSRI resistant depression in adolescents. *Journal of Consulting and Clinical Psychology*, 77(33):439–450, June 2009.

[48] Sun, Z. Type ii and type iii test in multi-center studies.

[49] Vierron, E. and Giraudeau, B. Sample size calculation for multicenter randomized trial: Taking the center effect into account. *Contemporary Clinical Trials*, 28:451–458, 2007.

[50] Vittinghoff, E., Glidden, D., Shiboski, S, and McCulloch, C. *Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models*. Springer, 2005.

[51] Worthington, H. Methods for pooling results from multi-center studies. *Journal of Dental Research*, 83(Special Issue C):C119–C121, 2004.