Title
Behavioral Health Intervention Effectiveness and Multiple Testing

Author
Bufford, Teresa Dianne

Publication Date
2022

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Behavioral Health Intervention Effectiveness and Multiple Testing

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Biostatistics

by

Teresa Dianne Bufford

2022

ABSTRACT OF THE DISSERTATION

Behavioral Health Intervention Effectiveness and Multiple Testing

by

Teresa Dianne Bufford
Doctor of Philosophy in Biostatistics
University of California, Los Angeles, 2022
Professor Hilary Aralis, Chair

Behavioral health interventions (BHI) have unique features that pose statistical challenges, both in the controlled trial and implementation stages. Family and school-based preventive BHI involve skill-building modules delivered by trained individuals which aim to improve well-being by promoting resiliency, empathy, communication, emotional regulation, and other related skills. Outcomes are measured through validated questionnaires given before and after the intervention. When establishing and evidence-base, researchers often conduct a randomized controlled trial of the BHI, through which they measure many, potentially correlated, outcomes. If the investigator hypothesizes that the intervention impacts each outcome measured, one must consider the problem of multiple testing when determining the overall efficacy of the intervention. It would be remiss to treat these tests as independent and existing methods for dependent outcomes require specification of the unknown correlation structure. To address this situation, we propose use of a permutation method to determine statistical evidence of an overall intervention effect. Two possible versions of a permutation test are presented, one that focuses on the number of significant individual hypothesis tests needed to indicate overall efficacy, and one that uses the magnitudes of the p-values for the individual tests to calculate an overall p-value for intervention efficacy.

Once efficacy has been demonstrated in an initial randomized trial, BHIs are often broadly implemented in real-world settings where adaptations to intervention protocol naturally arise.

Prevention scientists have recognized the need for ongoing evaluation of intervention adaptations. Again, we must consider the problem of multiple testing because the total number of hypothesis tests is unknown (and potentially unlimited) as data is continually collected. Existing statistical methods fall short when using continuously-generated real-world evidence to compare concurrent intervention versions. We propose combining methods used for observational data with methods for adaptive platform clinical trials. Since the data are observational, we use a pre-processing step to account for differences in covariate distributions among intervention groups. This allows us to more accurately estimate intervention effectiveness and make comparisons. We have developed a Bayesian analysis framework for interim decision making throughout the platform trial which allows us to determine the superiority or futility of concurrent intervention versions when compared to the current best version. Performance of the analysis framework is examined using simulations. Since type I error rate and power are not well defined in this context, we develop new metrics with which to evaluate the method. We demonstrate the potential utility of the combined framework using BHI data collected from a classroom-based resilience curriculum administered to Los Angeles Unified School District (LAUSD) high school students.

The dissertation of Teresa Dianne Bufford is approved.

Catherine M Crespi

Sung-Jae Lee

Donatello Telesca

Hilary Aralis, Committee Chair

University of California, Los Angeles

2022

Table of Contents

List of Figures

List of Tables

Acknowledgments

Material from Chapter 4 has been submitted for publication to the Journal of Educational and Behavioral Statistics (JEBS) (Bufford et al., 2022a). Material from Chapter 5 has been submitted for publication to Prevention Science (Bufford et al., 2022b). All work in this dissertation was made possible by support and funding from the Division of Population Behavioral Health within the UCLA Jane and Terry Semel Institute for Neuroscience and Human Behavior.

I would first like to thank my advisor Dr. Hilary Aralis, who has spent many hours discussing these projects and guiding me through the research process. I have learned so much from you over the years about how to approach complex statistical problems, think critically, and clearly communicate results. You have been an incredible mentor and an inspiration. I would also like to thank Maegan Sinclair Cortez, Dr. Sung-Jae Lee, and the entire Research and Evaluation Team for giving me the opportunity to work with behavioral health data, supporting me, and making me feel part of the team. Finally, I would also like to thank Drs. Catherine Crespi and Donatello Telesca for serving on my doctoral committee.

A special thank you to my parents, sisters, and brothers-in-law who have all continued to encourage me along the way. Thank you to my friends and classmates for the teamwork, collaboration, advice and moral support. And finally thank you to Phillip Sundin, without whom I likely would not have completed the dissertation. You have been there for me though every step.

Curriculum Vitae

2018–2022	Graduate Student Researcher, Department of Population Behavioral Health, Jane and Terry Semel Institute for Neuroscience and Human Behavior, UCLA.
2019	M.S. Biostatistics, UCLA, Los Angeles, California.
2017–2018	Teaching Assistant, Biostatistics Department, UCLA.
2015–2017	Boeing (IT Department) St. Louis, Missouri - Programmer Analyst
2014–2015	Anatomage Inc. San Jose, California - Application Specialist
2014	B.S. Applied Mathematics with Specialization in Computing UCLA, Los Angeles, California.

Publications

Bufford, T., Aralis, H., Crespi, C. M., Ijadi-Maghsoodi, R., Kataoka, S., and Lavelle, C. (2022a). Assessing intervention adaptations using real-world evidence and ongoing analysis. Submitted to the Journal of Educational and Behavioral Statistics.

Bufford, T., Aralis, H., Kataoka, S., Lee, S.-J., Lavelle, C., and Lester, P. (2022b). Creating a statistical analysis plan to continually evaluate intervention adaptations that arise in real-world implementation. Submitted to Prevention Science.

Xiong, D., Zhang, L., Watson, G. L., Sundin, P., Bufford, T., Zoller, J. A., Shamshoian,

J., Suchard, M. A., Ramirez, C. M. (2020). Pseudo-likelihood based logistic regression for estimating COVID-19 infection and case fatality rates by gender, race, and age in California. *Epidemics*, 33. https://doi.org/10.1016/j.epidem.2020.100418.

Watson, G. L., Xiong, D., Zhang, L., Zoller, J. A., Shamshoian, J., Sundin, P., Bufford, T., Rimoin, A. W., Suchard, M. A., Ramirez, C. M. (2021) Pandemic velocity: Forecasting COVID-19 in the US with a machine learning & Bayesian time series compartmental model. *PLoS Comput Biol*. 17(3): e1008837. https://doi.org/10.1371/journal.pcbi.1008837

CHAPTER 1

Introduction

1.1 Assessing Intervention Efficacy in the Trial Stage with Correlated Outcomes

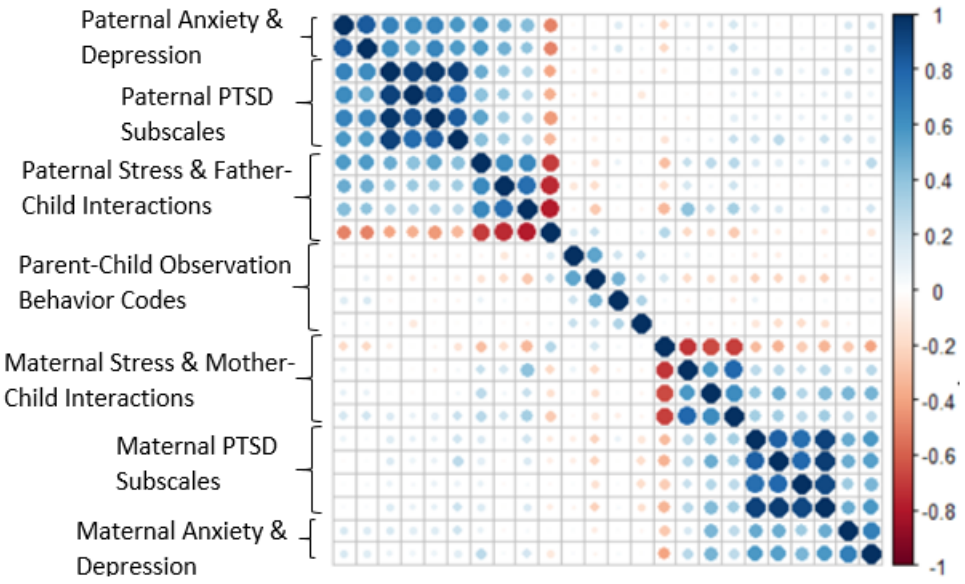
A behavioral health intervention (BHI) is a program that aims to improve well-being by promoting skills such as problem solving, resiliency, empathy, communication, and emotional regulation. These programs often take place over multiple sessions, and may be delivered in group settings. Outcomes are typically measured through validated questionnaires given before and after the intervention. Statistical challenges arise both in the trial stage, when the efficacy of a new intervention is tested in a randomized, controlled setting, and in the implementation stage, when an intervention is delivered on a large scale to various groups of people in real world situations.

In the trial stage, the main goal is to make a decision whether the intervention, overall, has a statistically significant benefit to the well-being of the participants. However there are usually a host of outcomes that are of interest to the researchers, rather than a single primary outcome. This is especially true in the case of family based BHI which often include measurements for more than one family member. For instance, researchers may want to track changes in mother's mental health, father's mental health, measures of dyadic relationships and family functioning, child developmental measures, and child behavioral measures.

An example of one such intervention is Families Overcoming Under Stress – Early Childhood (FOCUS-EC). FOCUS-EC is a multi-session trauma-informed and family-centered preventive intervention designed for military-connected families with young children and the

aim is to promote family resilience thereby reducing stress and adversity (Hajal et al., 2020; Mogil et al., 2010, 2015). With each family member completing a questionnaire consisting of multiple validated measures, each measure consisting of multiple sub-scales, we end up with 24 primary outcomes that we wish to analyze when assessing intervention impact. These outcomes include measures for PTSD symptoms, depression, anxiety, parental stress, parent-child interactions, and child behavior. Moreover, it is very reasonable to conclude that these measures are correlated with one another, but likely have a complex correlation structure. Figure 1.1 shows an empirically-estimated correlation matrix for the 24 FOCUS-EC outcomes. We note that this data appears to have smaller groups of measures that are highly correlated, and lower levels of correlation outside of these "clusters."

Figure 1.1: FOCUS EC Data Correlation Matrix.



The FOCUS-EC intervention is not unique in this predicament of wishing to assess multiple correlated outcomes, with a complex correlation structure, and draw a conclusion about overall intervention efficacy. Examples of similar BHI include The Special Education for Early Childhood Success - Reflective Parenting Program (SEEDS-RPP) which had 16 measures, Cultivating Awareness and Resilience in Education (CARE) which had 24 measures,

a resilience-oriented treatment for post-traumatic stress disorder which had 14 measures, interventions to reduce behavioral problems in children with cerebral palsy which had 11 measures, Strengthening Family Coping Resources (SFCR) which had 20 measures, a novel early intervention for preschool depression which had 18 measures, and PLAY Project Home Consultation intervention program for young children with autism spectrum disorders which had 20 measures (Jennings and et al, 2013; Kent and et al, 2011; Whittingham and et al, 2014; Kiser and et al, 2015; Luby et al., 2012; Solomon and et al, 2014).

In these and many other published studies, researchers are typically interested in comparing the changes from baseline to follow-up between groups, whether it is treatment vs. control or comparing specific sub-groups, and using the results of these comparisons to assess intervention efficacy. Efficacy can theoretically be assessed at the individual measure-level or overall with either approach necessitating the consideration of statistical issues arising from multiple testing and correlation among measures. While methodologies abound for addressing multiple testing when assessing the significance of individual measures, many of these approaches do not take into account clustered correlation and methods for determining overall efficacy, which often depend on the results of individual measure-level comparisons, are similarly limited.

To address the issue of multiple testing with many correlated outcomes, we propose a permutation test that provides a simple way of determining the number of significant results needed to provide sufficient evidence of an overall intervention effect. We then extend this permutation method to provide an single overall-pvalue for intervention efficacy that takes into account the magnitude of the p-values associated with each outcome. We posit that these methods are both valid for data with any underlying correlation structure and that the structure need not be known a priori.

Furthermore the proposed methods can help diminish publication bias among the body of work related to BHI. In the current state, researchers can claim significance of an intervention based only on a few significant findings, without accounting for multiple testing, or sometimes without fully acknowledging the large number of non-significant outcomes. We have no

measure of whether these claims are firmly supported or whether they are statistical artifacts resulting from the improper treatment of multiple testing. Adopting a convenient and widely applicable standard for assessing multiple testing among correlated outcomes would help prevent bias in this field of research. Our method could provide that standard.

1.2 Assessing Relative Effectiveness of Intervention Adaptations using Real-World Data

Once an intervention has been successfully vetted through a controlled randomized trial, service providers endeavor to deliver this evidence-based intervention to members of the intended population through widespread implementation. The intervention is delivered to groups with varying characteristics, and for practical reasons or due to perceived benefit, changes are often made to the intervention during the implementation process.

This is precisely what has taken place within Los Angeles Unified School District (LAUSD) with the intervention Families OverComing Under Stress (FOCUS). The FOCUS intervention was originally developed for military service members and their families but was then adapted for public school students to be administered in a classroom setting. Figure1.2 shows how the FOCUS intervention was adapted over the years. Our analysis in the coming chapters focuses on the implementation of the FOCUS Resilience Curriculum (FRC) at high schools. Psychiatric social workers employed by LAUSD were trained on the FOCUS intervention, referred to as the FRC, but were also given immense flexibility in implementing the program in order to meet the diverse needs of the students. FRC facilitators also varied greatly in their experience and available resources, leading to markedly different choices in implementation strategy.

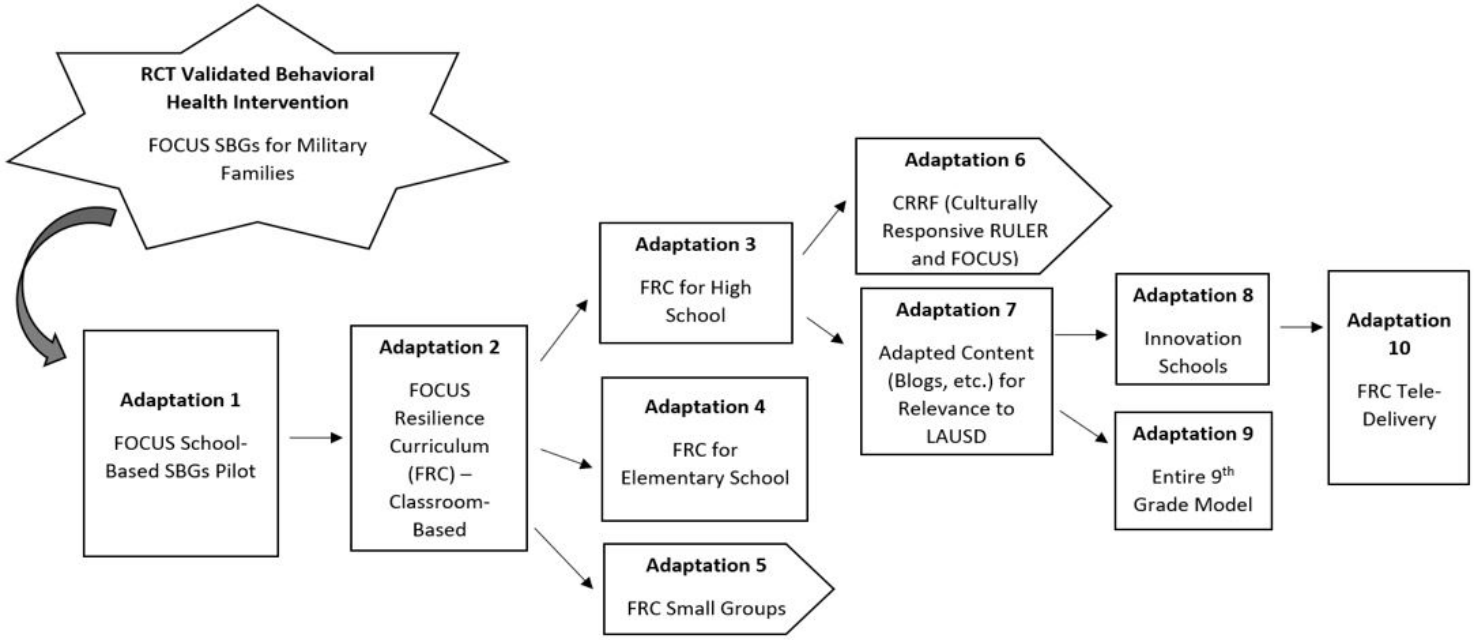


Figure 1.2: Evolution of the FOCUS Behavioral Health Intervention. The military family model was adapted for an urban school setting, piloted using UCLA staff as facilitators, adapted and rolled out for widespread use with PSW administrators (adaptation 2), adapted to be tailored to specific grades (high school, elementary school), adapted to better accommodate small groups (non-classroom based), adapted to include components from another evidence-based intervention (RULER), adapted to be administered to the entire population of 9th grade students at each school, adapted to serve specific schools in LAUSD (Innovation Schools) and, lastly, adapted to be delivered remotely during the pandemic.

Widespread implementation of evidence-based interventions has historically prioritized fidelity to the original intervention when analyzing intervention outcomes (Carroll et al., 2007; Ibrahim and Sidani, 2015). Adaptations to interventions naturally arise through the practical necessities of widespread implementation. However, any adaptation to the original intervention is typically viewed as inferior, having lower fidelity, unless a separate randomized controlled trial has been conducted to establish the efficacy of the new adaptation. This ignores the fact that some intervention adaptations may actually be beneficial, thereby demonstrating decreased fidelity but increased effectiveness.

A wealth of real-world evidence is generated through the process of intervention implementation, and it ought to be utilized to determine whether advantageous adaptations have arisen. Chambers and Norton describe the current lack of methodological development in this area as a major hindrance to attainment of population-level benefits (Chambers and Norton, 2016). Successful identification of the most beneficial intervention adaptations along with discontinuation of less advantageous adaptations can lead to improved outcomes in the long run. For this reason, we want to be able to continuously analyze outcomes of different intervention adaptations as data is collected over time. Once again we must address the issue of multiple testing as we repeatedly compare intervention versions after more data has been collected and when new adaptations arise.

We also have the added challenge that the data are observational, and issues of covariate imbalances among intervention groups must be addressed. Continuously comparing outcomes using real-world data requires the development of a new statistical framework. To address this challenge we combine an existing covariate balancing method called entropy balancing with a Bayesian platform clinical trial framework. A platform clinical trial is one in which multiple treatments are assessed simultaneously, and treatments can be added or dropped for futility at interim analyses. In combination with entropy balancing, the platform trial framework provides a way to compare intervention versions over time as data is continuously collected, while accounting for covariate imbalances.

This method can potentially be applied, not only to BHI, but more broadly to real-world

healthcare data. The need for continued learning from real-world data has been recognized in many fields, such as oncology treatment (Anatchkova et al., 2018), diabetes treatment (Schneeweiss and Patorno, 2021), generally in the world of biopharmaceutical development (Corrigan-Curay et al., 2018; Wang et al., 2021; Wise et al., 2018), and even for the purposes of regulatory decision-making (Franklin et al., 2020).

This dissertation is organized as follows. In Chapter 2 we introduce existing methods for multiple comparisons and uses of permutation tests. Then we define the first proposed permutation test, demonstrate its properties, and discuss it's uses and limitations. In Chapter 3 we introduce an extension to the permutation test described in the previous chapter, assess methodological properties of the test through simulation, and apply the method to the FOCUS EC data. In Chapter 4 we describe the development of a Bayesian statistical framework for ongoing comparisons with real-world data. We call it the 'Adaptome' framework, following the terminology used by Chambers and Norton. We explore ways to measure its performance, and demonstrate feasibility using simulations. Chapter 5 gives more information on the FOCUS Resilience Curriculum implemented at schools within LAUSD and illustrate how the Adaptome framework can potentially improve student outcomes. Further simulations give an example of how to use the existing school data for future implementation planning. Finally in Chapter6 we discuss limitations and possible extensions of the methods described in the previous chapters.

CHAPTER 2

Correlated Outcomes: Permutation Test

2.1 Introduction

2.1.1 Existing Methods for Controlling Type I Error

Conducting many statistical tests to determine the result of a single experiment warrants an adjustment or accommodation for multiple testing to control the experiment-wise Type I error rate. This is necessary when looking at the efficacy of the intervention as a whole. The Type I error rate, often denoted as α, is defined as the probability of wrongly concluding that there is an association or an effect, when in fact the null hypothesis of no effect is true. Experiment-wise error is defined as the overall probability of at least one Type I error for all of the statistical tests performed in evaluating an experiment or an intervention. Historically, 0.05 has been the accepted threshold for Type I error. Many forms of multiple testing adjustments, such as the popular Bonferroni correction or False Discovery Rate method, exist for independent tests (Aickin and Gensler, 1996; Benjamini and Hochberg, 1995).

For our purposes, these methods fall short in two ways. One, they assume each hypothesis test is independent. This would mean each outcome measured by the intervention would need to be independent of the others. It is important to account for the correlation in the data because, when compared to a series of independent tests, higher correlation will tend to increase the probability of a Type I error when conducting multiple tests (Harwood et al., 2017). Two, the methods focus on creating a new threshold for determining whether each individual test should be considered statistically significant, rather than making a conclusion about overall efficacy based on the information from multiple tests.

There has also been development of methods that adjust for multiple testing in the case of correlated hypothesis tests, especially in the statistical genetics literature since this is a common concern in genetic association studies. An example of one such method is the estimation of the effective number of independent tests, M_{eff}, developed by several researchers (Cheverud, 2001; Nyholt, 2004; Li and Ji, 2005; Gao et al., 2008). While these approaches accurately account for correlation in the data and are useful in the field of genetics, they still focus on drawing inference for individual hypothesis tests. We, instead, wish to use the information from all the hypothesis tests to draw inference at the experiment level, and determine if there was an overall intervention effect. While our problem falls under the umbrella term of "multiple comparisons" the underlying question that we wish to address is distinct from that of most existing multiple comparison methods.

Harwood et al begin to address the issue of drawing an overall conclusion from multiple correlated outcomes through a simulation study. For a fixed number of outcomes, Harwood et al simulated test statistics from a multivariate normal distribution with varying levels of correlation, from 0 to 0.9 with the assumed level of correlation being equal between all outcomes (Harwood et al., 2017). For a fixed number of outcomes and assumed correlation, they determine the minimum number (cut-point) of statistically significant one-sided tests in favor of the intervention that are needed in order to conclude there is an overall intervention effect while maintaining Type I error rate, α ¡ 0.05. A table is provided where one could potentially look up the cut-point needed, depending on the number of outcomes and the level of correlation. The limitation here is that in order to determine the correct cut-point for the number of significant tests, one would need to know, or at least be able to estimate the level of correlation among the test statistics. Additionally, one would also need to assume that the test statistics are all equally correlated. In practice, information about the correlation among test statistics for a given population and set of outcomes is unavailable and though the correlations can be estimated in many circumstances, the estimation can have high uncertainty. The assumption of equal correlation is likely a considerable oversimplification, particularly in contexts such as family-centered interventions and other BHI.

Thus, our objective is to develop and evaluate a statistical method allowing for control of Type I error when assessing multiple correlated outcomes where the level of correlation is unknown and may vary between pairs of outcomes. We also desire a method that is also non-parametric, and does not require the measured outcomes to have a Gaussian or other specified distribution. In order to have maximal benefit for determining efficacy of BHI, we require a test that can be applied to sets of outcomes that have various outcome models or modes of comparison. This additional flexibility is necessary for typical BHI outcomes.

In this chapter we describe the proposed permutation test and demonstrate a simulation-based approach for determining the power one can expect under different correlation scenarios. This can inform researchers who are designing a study to assess intervention effects using multiple correlated outcomes where they may have a range of plausible values for the correlation. This can potentially allow researchers to evaluate if there is an advantage to including an additional outcome measure based on its possible correlation with other measures.

2.1.2 Historical Use of Permutation Tests

A permutation test is a method for determining the sampling distribution of a test statistic by repeatedly permuting the covariate of interest and calculating the test statistic for each permutation, thereby creating a random sample of test statistics. Permutation tests have been used for decades as a way of simulating the null hypothesis in order to draw inference (Oden and Wedel, 1975; Berry and Mielke, 1985; Ludbrook and Dudley, 1998; Anderson, 2001; Hothorn et al., 2008). These methods can be computationally intensive due to the large number of permutations required to describe an entire distribution, and are typically used in situations when concise mathematical formulas for the desired statistical calculation do not exist. In that regard, permutation tests have a flavor similar to bootstrap methods. In the past, researchers have avoided permutation tests when possible, despite their robust nature, because conducting the test required significant computing resources and would often take inordinate amounts of time (Gao et al., 2008). With today's technology and availability of computing power, permutation tests are easier and faster to implement, giving the method

a renewed appeal (Pesarin and Salmaso, 2012).

Gao et al (2008) cite the use of a permutation test as a robust and well-established way to correct for multiple testing when dealing with many correlated outcomes, and use the permutation test as a standard of comparison for their proposed method. In their work, however, the interest is in developing a threshold for significance for individual hypothesis tests, similar to the Bonferroni and False Discovery Rate tests mentioned previously. Therefore the steps of the permutation method Gao et al describe are distinct from the method we propose because they do not assess overall efficacy by considering results across all outcomes.

2.2 Proposed Permutation Test

We propose a permutation test as a method for determining a threshold for the number of statistically significant hypothesis tests needed to give enough evidence to infer an overall intervention effect when the intervention has many correlated outcomes. We define S as the random variable that represents the number of significant findings that occur due to random chance, given we have performed M correlated hypothesis tests and there is no difference between intervention and control groups. We assume the empirical distribution of the observed data can be used to accurately approximate the true underlying distribution of S among the population of data we could have sampled. This is a standard assumption for all permutation tests. Based on the empirical distribution of S found through repeated sampling, we compute C, the estimated cut-point for number of significant findings needed. Let s_0 be the true number of significant findings, then our null hypothesis is that there is no underlying intervention effect ($s_0 < C$) and the alternative hypotheses is that the intervention caused differences in outcomes ($s_0 \geq C$).

The permutation test finds a unique cut-point for a given data set, without assuming a correlation structure but allowing the correlation present in the data to drive the result. To perform the permutation test, we randomly shuffle, or permute the intervention and control group assignments. For this new set of intervention and control groups, we calculate the

test statistics $T_1, ..., T_M$ for each of the M outcome measures being considered. For each permutation, the methods of data analysis and corresponding statistical hypothesis tests implemented for each outcome should be equivalent to the methods which will be used to analyze and draw inference using the original (un-permuted) data. These methods can be as simple as a t-test, as seen in the simulation study, or as complex as a longitudinal regression analysis, as seen in the example. The analysis model may even vary among different outcomes for the same intervention, lending great flexibility to the proposed permutation test..

Once the new set of test statistics, $T_1, ..., T_M$ have been calculated, we compare the associated p-values to the desired significance threshold. For our purposes, we do not want to allow the possibility for a significant effect in the opposite direction of what we would expect (i.e. the control group having a better outcome than the intervention group) to count positively towards an overall intervention effect. For this reason, we limit ourselves to one-sided hypothesis tests when implementing the permutation method. This leads us to define a statistically significant test as one that yields a one-sided p-value of $p < 0.025$. A one-sided p-value that meets the criteria above would yield the same critical value for the test statistic as a two-sided hypothesis test that controls the Type I error rate at $\alpha = 0.05$. If we define this rejection region corresponding to α for the m^{th} hypothesis test as $R_{m,\alpha}$, then the total number of significant tests is

$$s^{(i)} = \sum_{m=1}^{M} \mathbb{1}_{R_{m,\alpha}}\left(T_m^{(i)}\right)$$

Where $\mathbb{1}_A(x)$ is the standard notation for an indicator function, defined as:

$$\mathbb{1}_A(x) = \begin{cases} 1 \text{ if } x \in A \\ 0 \text{ if } x \notin A) \end{cases}$$

We have used the superscript (i) to denote results from the i^{th} permutation. This procedure is repeated many times to create a distribution for S, the number of statistically significant findings among M correlated outcomes under the null hypothesis. Each permutation has provided one sample from the distribution.

In order to find our cut-point for the number of significant findings needed to demonstrate an overall intervention effect, we find the 95^{th} percentile of this simulated distribution of S. By increasing the number of permutations, we increase the precision of the estimated percentile. Manly (1997) recommends using at least 1,000 permutations for tests that require 95% confidence and at least 5,000 permutations for tests that require 99% confidence. For G permutations, the 95^{th} percentile, P_{95}, is calculated as

$$P_{95} = (1 - \gamma)s_{(j)} + \gamma s_{(j+1)}$$

Where

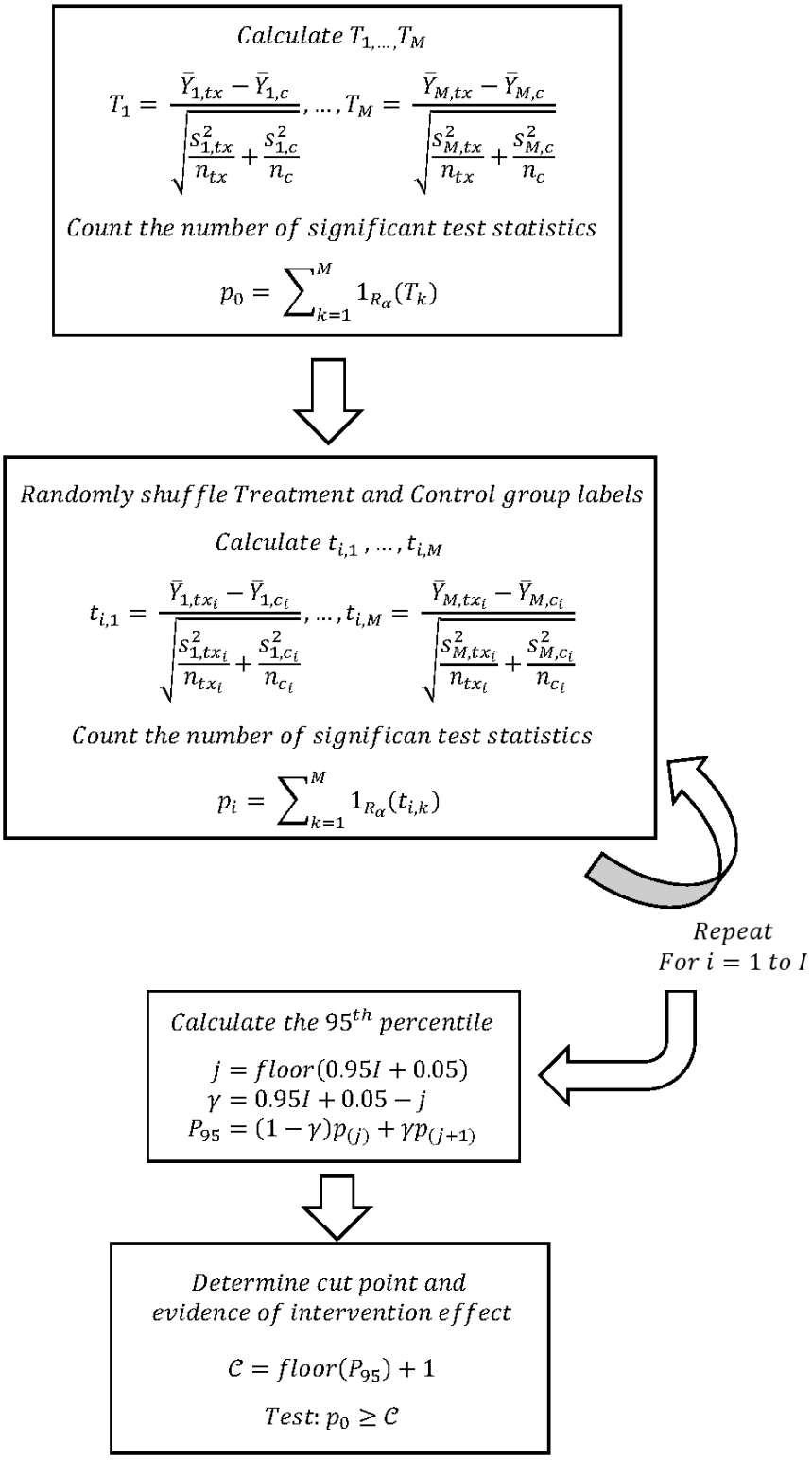
$$j = \text{floor}(0.95G + 0.05)$$
$$\gamma = 0.95G + 0.05 - j$$

Essentially, P_{95} is a weighted average of two consecutive order statistics and we note that this can be either an integer or a decimal value (R Core Team, 2020; Hyndman and Fan, 1996). Finally we define our cut-point, C, as the smallest integer that is greater than or equal to P_{95}.

$$C = \text{floor}(P_{95}) + 1$$

As an example of determining the cut-point, suppose we have 30 outcomes measured and we perform a permutation test for overall intervention efficacy with 5,000 permutations. From looking at the distribution of $s^{(1)}, ..., s^{(5,000)}$ we find that $P_{95} = 4.25$. In this case we require 5 or more of the outcomes to show statistically significant results in favor of the intervention ($s_0 \geq C = 5$) in order to conclude there was an overall intervention effect. Specifically, if $s_0 = 5$, we conclude that the intervention had a positive effect on those 5 outcomes, but naturally we do not conclude that the intervention has affected the remaining 25 outcomes, and we are at least 95% confident that these significant findings were not a result of random chance due to multiple testing. A flow chart with the steps for the permutation test are shown in Figure 2.1.

Figure 2.1: Permutation Test for Correlated Outcomes - Flow Chart



2.3 Simulation Study

2.3.1 Simulation Design

A simulation study was designed to evaluate the properties of this permutation method when used with correlated outcomes. Data was generated from a multivariate normal distribution, using 20 different correlation matrices, $\Sigma_1, ..., \Sigma_{20}$. We will describe these chosen correlation structures in further detail in the following paragraph. The simulated scenarios were labeled 1-20 based on the correlation matrix used. For each correlation structure, the required cut-points were evaluated for numbers of outcomes, M, such that $M \in 10, 15, 20, 25, 30, 35, 40$. Varying M allowed us to assess how the permutation method performed as the number of outcomes from the experiment grew. For each correlation structure and each possible value of M, we simulated data under both H_0, no difference between intervention (tx) and control (c) ($\boldsymbol{\mu}_{tx} = \boldsymbol{\mu}_c = 0$), and H_1, an intervention effect with Cohen's d of 0.3 ($\boldsymbol{\mu}_c = 0, \boldsymbol{\mu}_{tx} = 0.3$). This is a viewed as a moderate, but reasonable effect size that is within the realm of possibility for the effect size a clinician may expect to see. We assumed that researchers are not measuring outcomes in which they do not expect to see any intervention effect, so we simulated an effect across all outcome measures. For simplicity, we kept the effect size uniform across measures, however this assumption will be relaxed in the following chapter. We used a sample size of 200 subjects with equal proportions assigned to intervention and control groups, such that $n_{tx} = n_c = 100$. Each individual represents an independent sample drawn from a multivariate normal distribution with M correlated outcomes. Thus the data generation can be summarized by

$$\text{Control: } \boldsymbol{Y}_c \sim MVN(0, \Sigma_i)$$

$$\text{intervention: } \boldsymbol{Y}_{tx} \sim \begin{cases} MVN(0, \Sigma_i) \text{ under } H_0 \\ MVN(0.3, \Sigma_i) \text{ under } H_1 \end{cases}$$

Simulations under H_0 allowed us to estimate the probability of an experiment-wise Type

I error, α, and evaluate the permutation method as a way of controlling Type I error. Since we knew the true distribution of the data, simulations under H_1 allowed us to estimate β, the probability of a Type II error. A Type II error is made when one wrongly concludes there is no difference between the two groups, when in fact there is a difference. The power of a study is defined as $1-\beta$. Using the simulations, we estimated the power of the permutation method in testing the hypothesis of an overall intervention effect. Comparisons of power across scenarios can be helpful at the study design phase where researchers may have an idea of the correlation among potential outcome measures.

The 20 correlation matrices used in the simulations were separated into two general categories, equicorrelation, and what we call "clustered correlation." The clustered correlation structure had groups, or clusters, of measures with high intra-cluster correlation, and a lower level of inter-cluster correlation. While the simulations with a constant correlation matrix were useful for comparisons to existing results from Harwood et al, the simulations with a clustered correlation matrix were the main interest of this study because we believe this to be a realistic representation of the correlation between measures often observed in BHI, particularly when the intervention is family-based. One could imagine that several measures related to the mother's mental health may be highly correlated, and that measures related to communication and family functioning may also be highly correlated, forming two "clusters" of measures. There may also be a lower level of correlation between these two clusters because the mother's mental health could impact her manner of communication with family members. Since there are many possibilities for the exact correlation structure of data that falls into the "clustered correlation" category, 10 different versions of a correlation matrix with correlated clusters were simulated. The specific clustered correlation structures were chosen, not as an exhaustive set, but to give a variety of possibilities with which to demonstrate the general performance of the permutation method. Details of the simulated data structures for constant correlation are given in Table 2.1 and for clustered correlation in Table 2.2. A sample clustered correlation matrix is given in Figure 2.2.

For each correlation scenario described in Tables 1 and 2 and each value of M outcome

Table 2.1: Simulated data structures for constant correlation among outcomes

Σ	**Correlation** (ρ)
Σ_1	0
Σ_2	0.1
Σ_3	0.2
Σ_4	0.3
Σ_5	0.4
Σ_6	0.5
Σ_7	0.6
Σ_8	0.7
Σ_9	0.8
Σ_10	0.9

Figure 2.2: Example Clustered Correlation Matrix (Σ_{18}).

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10
V1	1	0.6	0.7	0.2	0.2	0.2	0.2	0.2	0.2	0.2
V2	0.6	1	0.5	0.2	0.2	0.2	0.2	0.2	0.2	0.2
V3	0.7	0.5	1	0.2	0.2	0.2	0.2	0.2	0.2	0.2
V4	0.2	0.2	0.2	1	0.6	0.6	0.5	0.6	0.2	0.2
V5	0.2	0.2	0.2	0.6	1	0.4	0.6	0.4	0.2	0.2
V6	0.2	0.2	0.2	0.6	0.4	1	0.5	0.5	0.2	0.2
V7	0.2	0.2	0.2	0.5	0.6	0.5	1	0.4	0.2	0.2
V8	0.2	0.2	0.2	0.6	0.4	0.5	0.4	1	0.2	0.2
V9	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	1	0.8
V10	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.8	1

measures, 500 datasets were simulated with no intervention effect and 500 more datasets were simulated with an intervention effect present. For each simulated dataset, a t-test was used to compare intervention and control groups for each of the M outcomes. We emphasize that the permutation method can be used with any test statistic, however the t-test provided a simple example and was beneficial for computing speed. For each simulated dataset the permutation method described in Figure 2.1 was implemented and the results recorded. Results consisted of the true number of significant outcomes, the estimated cut-point, C, and an indicator of whether a Type I or Type II error was made. From the 500 simulations, an average cut-point and either α or $1 - \beta$, respectively, are estimated.

Table 2.2: Simulated data structures for clustered correlation among outcomes.

Σ	ρ **within Cluster**	**Structure within cluster**	ρ **between Clusters**	**Custer Size**	**Outcomes per Cluster**
$\Sigma_1 1$	0.4-0.8	Constant	0.2	Equal	5
$\Sigma_1 2$	0.3-0.6	Constant	0.1	Equal	5
$\Sigma_1 3$	0.4-0.5	Varying	0.2	Equal	5
$\Sigma_1 4$	0.5-0.6	Varying	0.2	Equal	5
$\Sigma_1 5$	0.6-0.7	Varying	0.2	Equal	5
$\Sigma_1 6$	0.7-0.8	Varying	0.2	Equal	5
$\Sigma_1 7$	0.4-0.8	Varying	0.2	Equal	5
$\Sigma_1 8$	0.4-0.8	Varying*	0.2	Varying	2-8
$\Sigma_1 9$	0.3-0.6	Varying	0.1	Varying	2-8
$\Sigma_2 0$	0.6-0.8	Varying	0.2	Varying	2-8

*Higher correlation for smaller clusters and lower correlation for larger clusters

2.3.2 Simulation Results

The average cut-point for each scenario is shown in Figure 2.3. As expected, the cut-points for Scenarios 1-10 with constant correlation follow the general pattern shown by Harwood et al. The number of significant tests needed increases with number of outcomes, and also increases as correlation increases until we reach correlation of about 0.6 or 0.7, after which we see a drop off in the cut-point. This likely occurs because when the correlation gets too high, we are not gaining very much new information from the additional measures, so it's analogous to having fewer outcome measures with lower correlation. For the clustered correlation, we see that the average cut-point increases as the number of outcomes increases, and the trend is roughly linear.

The estimated values of $\hat{\alpha}$ and $1 - \hat{\beta}$ from the simulation study are found in Figure 2.4. We expect the permutation method to hold the Type I error rate for the hypothesis of an overall intervention effect at $\alpha = 0.05$. However each value of $\hat{\alpha}$ plotted is an estimate based on 500 simulated datasets, and when estimating a proportion of 0.05 using a sample of 500, there will be some variation in the exact estimate. If the true proportion is 0.05, then 95%

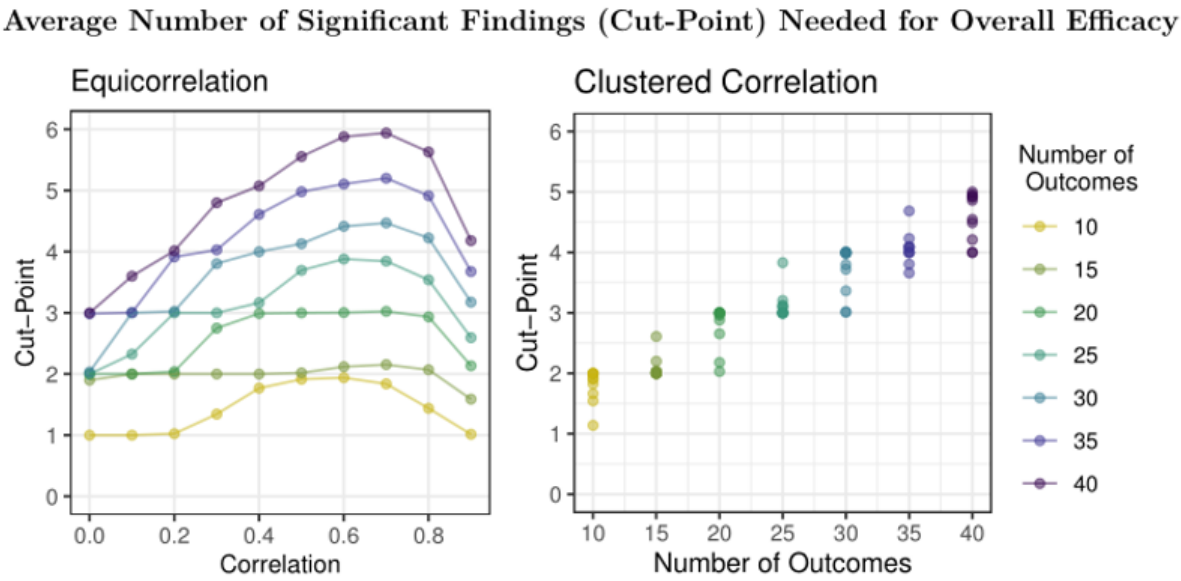


Figure 2.3: Each point represents the estimated number of significant findings needed to provide evidence for an overall intervention effect (ie. the cut-point), averaged over 1,000 simulations using the permutation method. For equi-correlated data, the correlation structures have a natural ordering, so we provide a line graph to display the pattern as correlation increases. Since the various clustered correlation structures do not have a defined order, we provide a scatterplot distribution of point estimates, one point estimate for each simulated data structure.

of estimates will fall between 0.033 and 0.073. Looking at the simulation results, we see that no estimates of α exceed this upper bound, but several estimates fall below the lower bound. For scenarios with constant correlation, 10 out of 70 estimates of α are above 0.05, with the maximum of 0.066 and minimum of 0.014. The average of these 70 estimated values using datasets with constant correlation is $\bar{\hat{\alpha}} = 0.038$. For scenarios with clustered correlation, 8 out of 70 estimates of α are above 0.05, with a maximum of 0.066 and a minimum of 0.012. The average of the 70 estimated values using datasets with clustered correlation is $\bar{\hat{\alpha}} = 0.039$. This suggests that the permutation method is generally on the conservative side, controlling α to be slightly lower than 0.05. The conservative nature of the permutation method comes from always rounding up from the 95^{th} percentile to get the cut-point. Having to select an integer value by rounding is less impactful when the number of outcomes is higher compared to when the number of outcomes is lower. For lower numbers of outcomes, the permutation

test errs on the conservative side to a greater extent.

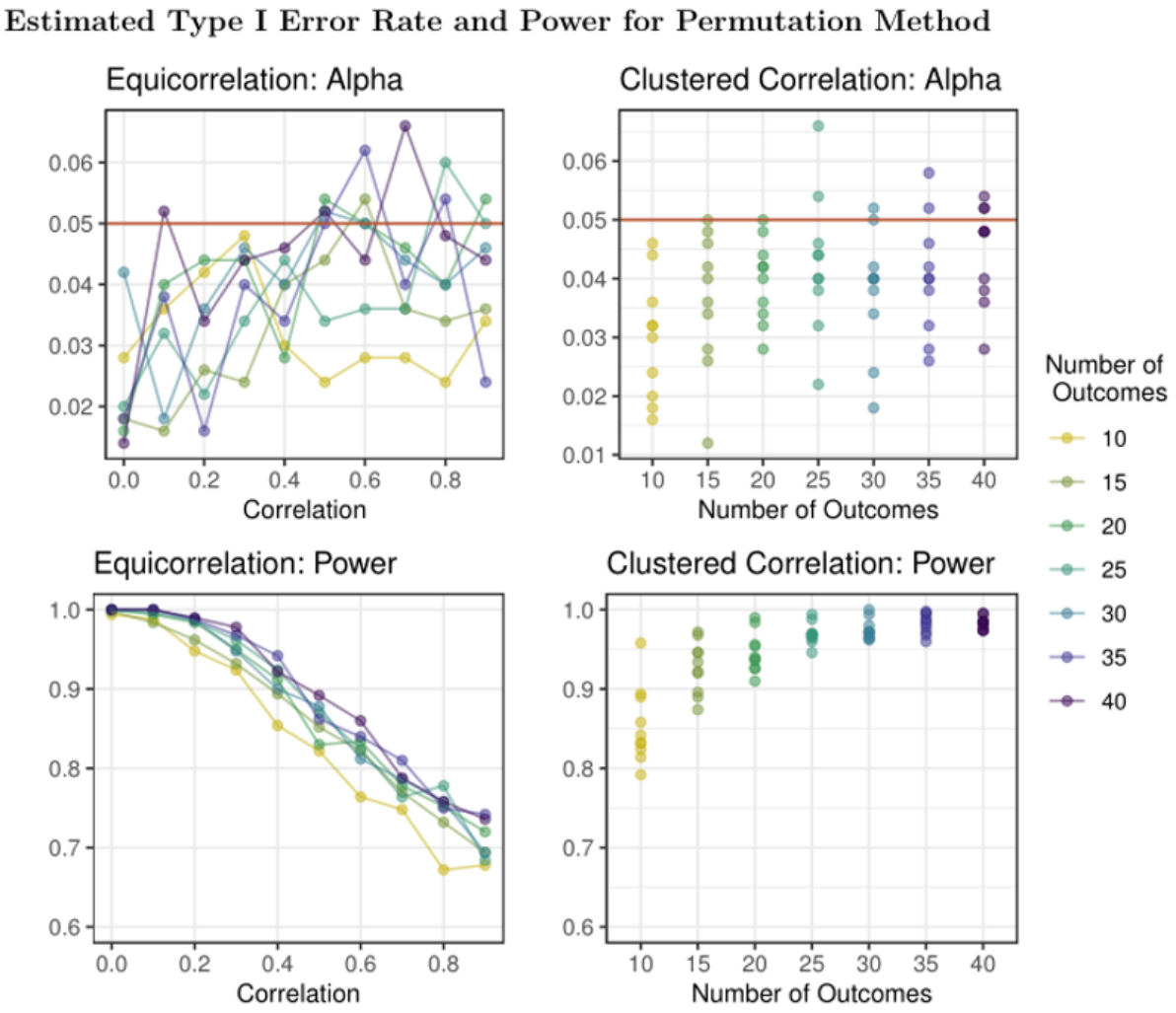


Figure 2.4: For each simulation, a cut point for the number of significant p-values is estimated, and a conclusion is drawn about whether or not the intervention was effective. We used 500 repetitions of each simulated scenario to estimate the type I error rate (alpha) for the permutation method when H0 is true and the power for the permutation method when H1 is true. For equi-correlated data, the correlation structures have a natural ordering, so we provide a line graph to display the pattern as correlation increases. Since the various clustered correlation structures do not have a defined order, we provide a scatterplot distribution of point estimates, one point estimate for each simulated data structure.

In the case of constant correlation, we see a slight increasing trend in the estimated Type I error rate as the level of correlation increases. This unanticipated result is explained by the

shape of the estimated distribution of S, the number of significant tests, under low versus high correlation and the fact that this distribution is discrete, taking only integer values. As seen in Figure 2.5, when correlation is zero, the distribution of S has a shorter right tail. Since the value of S can only be integers, this leads to multiple percentiles of the distribution having the same estimated value. For example, in one simulation with 25 outcomes and correlation of 0.0, the 87^{th} percentile and the 96^{th} percentile, and all percentiles in between, were all estimated to be 2. In this case, controlling the Type I error rate at 0.05 by using the 95^{th} percentile to find the cut-point is equivalent to using the 87^{th} percentile. This leads to lower estimated α for scenarios with low equi-correlation. Contrastingly, the distribution of S under high correlation has a longer, thinner right tail. This leads to distinct estimates for each percentile in the upper end of the distribution, allowing us to more precisely control the Type I error rate at 0.05.

In studies with multiple endpoints, it is typical to calculate the power for the study based on a chosen single endpoint. If we were to do this with our simulation, where δ is the difference in means, the power calculation would be as follows:

$$\begin{aligned}
\text{Power} &= P(\text{reject } H_0 | H_1 \text{ true}) \\
&= P(\text{Test Statistic} > 1.96 | \delta = 0.3) \\
&= P\left(\frac{\bar{y}_T - \bar{y}_c}{\hat{\sigma}/\sqrt{n}} > 1.96 | \mu_T - \mu_C = 0.3\right) \\
&\vdots \\
&= 0.56
\end{aligned}$$

Using this as a point of comparison, we can easily see that considering multiple outcomes and then using the permutation method to adjust for multiple comparisons increases the overall power of the study. In the simulated scenarios of clustered correlation, the permutation method maintains high power, above 80% for every estimate, while simultaneously controlling the Type I error rate at or below 0.05. It is important for a study to have high power so that a clinician can accurately detect any real effect that the intervention or intervention may

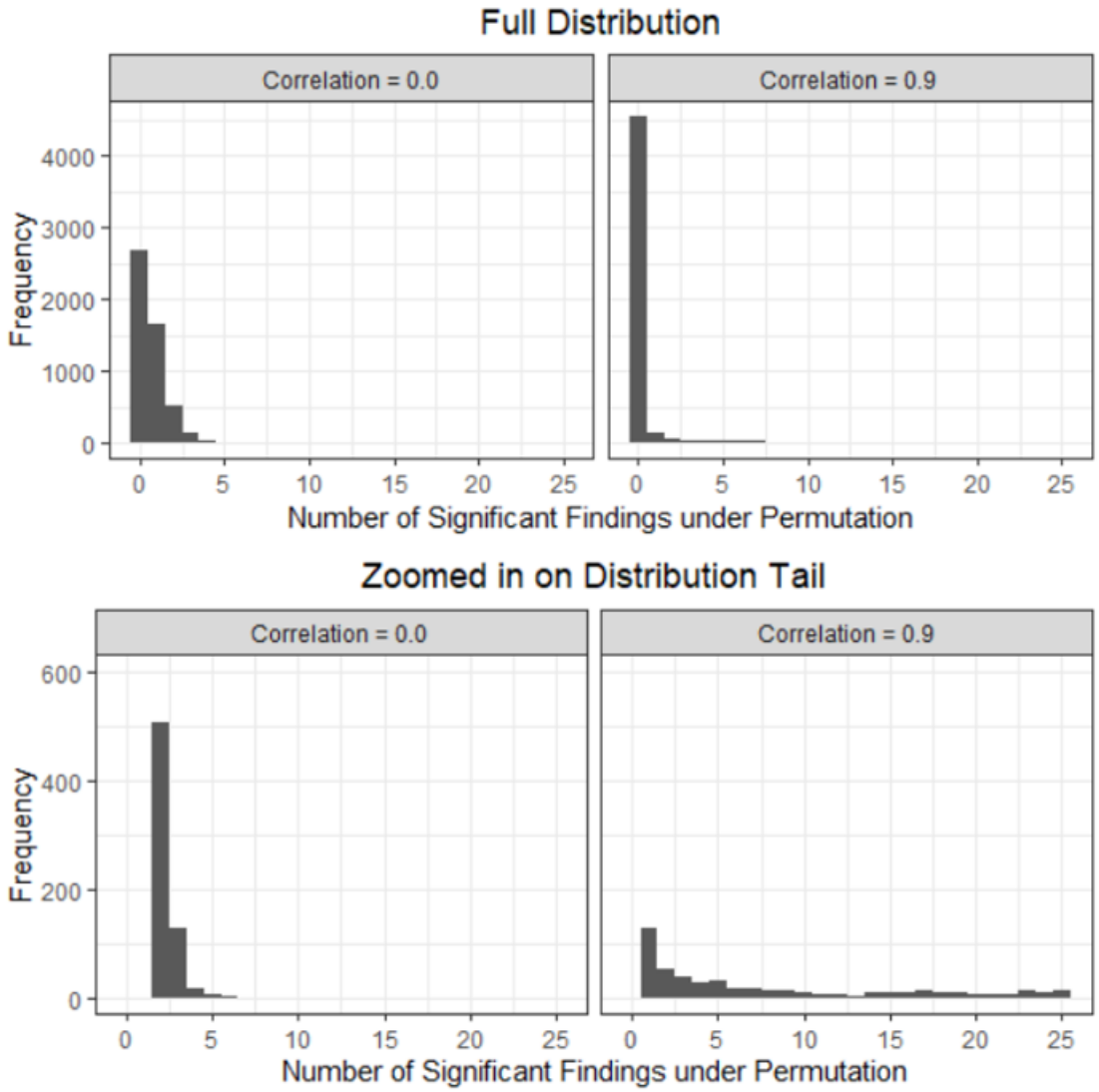


Figure 2.5: Distribution of the number of significant findings under permutation with 25 outcomes with no correlation vs. high correlation

have on the target population. Now that we have a method for assessing whether there is an overall intervention effect, researchers may be able to avoid choosing a single outcome as the primary endpoint, and opt for a study design that has multiple outcomes and therefore higher power, while also controlling Type I error rate.

When designing future studies, it is useful to know which scenarios have higher or lower power in relation to each other. The estimated power is higher overall in the clustered correlation simulations than in the equi-correlation simulations. The correlation structure

defined by scenario 16 in Table 2.2 had the lowest power and correlation structure defined by scenario 19 had the highest power. This is consistent with our finding that correlation level among outcomes and overall study power are inversely related.

2.4 Discussion

A major benefit of the proposed permutation method is that there is no distributional assumption made for the number of significant findings or the level of correlation of the test statistics. The method is non-parametric, utilizing an empirical distribution estimated through repeated sampling from the permutation distribution. We posit that this method is valid for data with any underlying correlation structure and have conducted simulations that support this conclusion.

This method also differs from conventional multiple testing methods because it focuses on the question of whether there is an underlying intervention effect that causes the observed differences between intervention and control groups for various outcome measures. In contrast, conventional multiple comparison methods focus on drawing inference for individual measures. The permutation method allows multiple findings that meet the standard significance threshold to increase our confidence in the efficacy of the intervention, rather than having to adjust the threshold for each p-value. In behavioral health we may have a relatively small sample size and many measured outcomes which show results that favored the intervention but are just barely below the traditional 0.05 threshold. If we were to use a multiple comparison method that adjusts the necessary threshold for each p-value, we may be left with no statistically significant findings, leading to the conclusion of no notable intervention effect. More realistically, we would want to recognize that the p-values are tied to sample size and impacted by correlation among outcome measures, therefore the number of findings that all show improvement as a result of the intervention can collectively give evidence of an intervention effect.

In addition, the application of this method is not limited to t-tests. While t-tests were

used in the simulation study as a simple illustration and were the basis of the previous work conducted by Harwood et al, the proposed permutation method hinges on counting the number of significant p-values at each iteration, which can easily be extended to any type of hypothesis test. It can accommodate not only normally distributed outcomes, but also binary outcomes, count data, survival data, or a mixture of these. We could perform linear regression, ANOVA, logistic regression, or use any combination of analysis models, perform hypothesis tests on the model coefficients, and then re-compute these test statistics under permutation. The extension of permutation methods, in general, to ANOVA, regression, and other common models is described in detail by Anderson (Anderson, 2001). One practical application of the permutation method as a test for overall efficacy in the presence of a mixture of outcomes, is in a drug trial. As stated by Davidson et al, it may be useful in early stages of a clinical trial to measure a variety of outcomes related to the new drug in question, and look for a minimum number of positive results in order to justify continuing the drug development and testing (Davidson and et al, 2011).

Finally we wish to address the issue of dichotomizing p-values. This method relies on the ability to determine whether or not the differences between two groups are statistically significant based on a p-value threshold. Recently there has been talk among the statistical community that recommends against dichotomizing p-values based on the traditional 0.05 threshold, and instead supports greater use of confidence intervals (Wasserstein and Lazar, 2020). Proponents of this idea typically argue that a p-value alone does not determine whether or not a finding is scientifically meaningful, but rather effect size must also be considered. Studies with large sample sizes are more likely to result in many statistically significant findings, even if the differences between groups are small, thus it will be prudent in these cases to look at both effect sizes and p-values. The above permutation method can easily be extended to include a check of effect size for each test statistic before determining whether a finding is significant. The researchers would simply specify in advance a minimum effect size needed for each hypothesis test.

The following chapter describes an alternative permutation test that similarly allows the

researcher to make a decision about overall intervention efficacy, but instead incorporates the magnitude of the p-values for each outcome into the determination of efficacy. A main difference in these two approaches is the null hypothesis used. In this chapter we explore a way to quantify the number of significant individual hypotheses needed to give evidence of an overall intervention effect, and in the following chapter we assume a strict null hypothesis such that any individual test can give evidence of an intervention effect. This distinction is necessary in order to compare performance of our proposed method to existing methods for controlling Type I error rates.

CHAPTER 3

Correlated Outcomes Extension: Permutation-Rank-Sum Test

3.1 Introduction

The permutation test described in the previous chapter provides researchers with a simple and valid method for determining whether there is an overall intervention effect when numerous correlated outcomes are measured and comparisons between groups are conducted using any of a flexible range of different statistical hypothesis testing procedures/models. However, the results of the previous chapter gave rise to several questions which we propose to address in this chapter with an extension in the form of a rank-sum-style permutation test. These questions include:

- Can we increase power by considering the magnitude of the p-values rather than dichotomizing them (significant/non-significant)?
- Can we determine when it is advantageous to use the proposed permutation test relative to other potential testing procedures?

Using the actual p-values may be advantageous relative to the method presented in the previous chapter. Specifically, we may be able to glean more power by using more information while still having the benefits described with the prior test: minimal assumptions, flexible for use with any test, test based on the null hypothesis of exchangeability. To answer the second question, we need to define our alternative hypothesis more explicitly than was done in the previous chapter. To make our method more comparable to alternative tests an analyst

might select, such as Hotelling's T2 or Bonferroni, we adopt an alternative hypothesis of at least one outcome for which the two groups are not equivalent.

As in the previous chapter, we wish to conduct overall inference on intervention effectiveness when the intervention has potentially many primary outcomes which are correlated. Going forward we define a strict null hypothesis of: $H_0 : \boldsymbol{\mu} = \mathbf{0}$ for all M outcomes. Thus our alternative hypothesis is: $H_A : \boldsymbol{\mu} \neq \mathbf{0}$. Therefore a departure from mean zero for any one of the M outcomes results in rejecting the null hypothesis. An example of a test that uses this type of null and alternative hypothesis is Hotelling's T^2 test, which is a multivariate extension of a simple t-test for a difference in means between two samples. We will use Hotelling's T^2 test as a point of comparison for our proposed method. A notable difference between the two methods is that Hotelling's T^2 is a parametric test, which assumes the data are multivariate normal and that a covariance matrix can be estimated, while the proposed method is non-parametric, drawing on existing non-parametric methodologies. Specifically, we combine the idea of a permutation test with that of a rank-sum test, where the observations being ranked are the p-values from the many permuted datasets. In this manner, by ranking the independently generated p-values, we can compare the magnitude of the p-value for each outcome in the original data with the distribution of p-values generated through permutation.

3.2 Method

Suppose we have M outcomes. Let P denote the set of M p-values $P = \{p_1, ..., p_M\}$ from the hypothesis test associated with each respective outcome. Then suppose we have K random permutations of intervention group assignment and perform the hypothesis tests for each outcome for each permuted data set. Including the set of p-values from the outcome analysis with the allocation intervention groups (the unpermuted data), we together have $K + 1$ sets of M p-values. Let P_k denote the set of p-values corresponding to the kth permutation for $k = 1, ..., K$ and P_0 denote the set of p-values from the unpermuted (original) data. Without loss of generality, let $P_0, P_1, ..., P_K$ be row vectors, that can be stacked to

form a $(K+1)xM$ matrix. Thus column m of this matrix contains the p-value for outcome m derived from the original data along with K p-values for outcome m derived from data with permuted intervention group assignments.

Rank the p-values associated with each outcome, i.e. the values in each column, such that we create a corresponding $(K+1)xM$ matrix with each column containing ranks 1 through $(K+1)$. Assume the smallest p-value receives rank 1 and the largest p-value receives rank $K+1$. For each permutation, and for the original data, sum the ranks across the M outcomes. Let R_0 denote the sum of the ranks of P_0, and let R_k denote the sum of the ranks of P_k. The overall p-value for the permutation-rank-sum test can then be calculated analytically as:

$$p = 1 - \frac{\sum_{k=1}^{K} \mathbb{1}(R_0 < R_k)}{K}.$$

The overall p-value thus reflects the proportion of instances in which the p-value ranks from the unpermuted data summed across outcomes is less than the equivalent rank-sum statistic calculated under a random permutation of the group labels.

3.3 Simulation Design

A simulation study was designed to evaluate the properties of this permutation-rank-sum test when used with correlated outcomes. Data was generated from a multivariate normal distribution, using 6 different correlation matrices, a subset of those used in the previous chapter. Going forward we re-label the correlation matrices $\Sigma_1, ..., \Sigma_6$, as defined in Table 3.1. The correlation matrices were separated into two general categories, constant correlation, and what we call "clustered correlation." The clustered correlation structure is characterized by groups, or clusters, of outcomes with high intra-cluster correlation, and a lower level of inter-cluster correlation. The simulations which use a clustered correlation matrix are of particular interest because this correlation structure is a realistic representation of the correlation between measures in a family based BHI.

For each correlation structure, the data was simulated using several values of M ($M \in$

Table 3.1: Correlation matrices for simulations.

Σ	Category	Description
Σ_1	No correlation	Identity matrix
Σ_2	Constant correlation	All outcomes have correlation 0.3
Σ_3	Constant correlation	All outcomes have correlation 0.6
Σ_4	Clustered correlation	• clusters of varying size • intra-cluster correlation 0.3-0.6 (varies within cluster) • inter-cluster correlation 0.1
Σ_5	Clustered correlation	• clusters of varying size • intra-cluster correlation 0.4-0.8 (varies within cluster) • inter-cluster correlation 0.2 • smaller clusters have higher correlation
Σ_6	Clustered correlation	• clusters of varying size • intra-cluster correlation 0.6-0.8 (varies within cluster) • inter-cluster correlation 0.2

$10, 20, 30$) and several alternative hypotheses (Table 3.2). The alternative hypotheses varied in both magnitude of the difference in means between intervention and control groups, and the proportion of the M outcomes that had a mean difference. Data was also simulated under the null hypothesis to evaluate the Type I error rate. The number of observations in the intervention and control groups was kept constant at 100 per group. In addition to the simulations with multivariate normal data, simulations were also run with non-normal data. A skew of 1.5 and kurtosis of 4 for was induced for all outcomes. The simulations were limited to $M = 10$ outcomes and correlation matrices $\Sigma_1, \Sigma_2, \Sigma_4$, and Σ_5 because it is computationally nontrivial to generate non-normal data with high levels of specified correlation and large numbers of outcomes.

For each combination of correlation matrix, hypothesis and data shape, 200 datasets were generated and assessed. When performing the permutation-rank-sum test, the intervention and control groups were compared using a t-test, and thus p-values were calculated for each of the M outcomes. The number of permutations used was 5,000. As a mode of comparison, each dataset was also assessed using a Bonferroni adjustment for each of the p-values, using Hotelling's T^2 test and using a sign test. The sign test was implemented using the set of

Table 3.2: Set of hypotheses for simulations.

Hypothesis	**Proportion of outcomes with a mean difference**	**Magnitude of mean difference**
Alternative 1	1.0	0.2
Alternative 2	1.0	0.225
Alternative 3	1.0	0.25
Alternative 4	1.0	0.275
Alternative 5	1.0	0.3
Alternative 6	0.75	0.3
Alternative 7	0.5	0.3
Alternative 8	0.25	0.3
Null	0	0

M estimated mean differences testing whether the set of mean differences had a median greater than zero. This is consistent with the common practice described by Harwood et. al. When using Bonferroni, an overall conclusion about the efficacy of the intervention was determined by assuming that if one or more of the individual tests showed significance after adjusting the significance threshold, then the intervention as a whole was deemed effective. This is a rather generous assumption and the conclusions about statistical significance of individual hypothesis tests do not truly answer our question of overall intervention efficacy. Nevertheless, we use this as a comparison because it is an approach often implemented in BHI research.

3.4 Simulation Results

3.4.1 Multivariate Normal Data

Estimated type I error rates for each method are shown in Table 3.3. The sign test does not adequately control type I error, when using correlated test statistics to perform the sign test. In instances when the data has constant correlation of 0.3, the type one error rate using the sign test peaks at 0.41. For this correlation structure and others, the sign test gives type I error rates well above the standard 0.05 threshold. Contrastingly, the Bonferroni

and Hotelling's T^2 methods both successfully control type I error rates strictly below the common 0.05 threshold for all correlation structures assessed. Hotelling's T^2 test appears to be the most conservative, with consistently low type I error rates. Estimates for type I error rates using the permutation-rank-sum test do not remain strictly below 0.05, but generally hover near the threshold.

Table 3.3: Type I error rates estimated from 200 simulations

			Proportion of datasets where null hypothesis is rejected			
	Covariance	M	Bonferroni	Sign Test	Hotelling's T^2	Permutation-Rank-Sum
Multivariate Normal Data	Σ_1	10	0.02	0.05	0.015	0.055
		20	0.04	0.065	0.005	0.035
		30	0.045	0.065	0.02	0.04
	Σ_2	10	0.02	0.21	0.03	0.03
		20	0.035	0.255	0.04	0.03
		30	0.04	0.295	0.02	0.07
	Σ_3	10	0.045	0.31	0.01	0.06
		20	0.035	0.41	0.02	0.06
		30	0.01	0.38	0.02	0.04
	Σ_4	10	0.04	0.165	0.025	0.06
		20	0.03	0.22	0.02	0.04
		30	0.04	0.27	0.01	0.045
	Σ_5	10	0.045	0.23	0.03	0.055
		20	0.035	0.28	0.015	0.06
		30	0.04	0.365	0.035	0.045
	Σ_6	10	0.025	0.235	0.035	0.04
		20	0.015	0.255	0.01	0.035
		30	0.02	0.325	0.01	0.04
Skewed Data	Σ_1	10	0.045	0.045	0.01	0.03
	Σ_2	10	0.04	0.165	0.03	0.05
	Σ_4	10	0.05	0.135	0.025	0.07
	Σ_5	10	0.035	0.215	0.02	0.06

When data is simulated under one of the alternative hypotheses, we use the proportion of datasets that successfully identify an overall intervention effect as an estimate of power. As expected, power increases as the proportion of outcomes that have a true mean differ-

ence increases (Figure 3.1). This is true for the permutation-rank-sum test, the Bonferroni method, and the sign test. However Hotelling's T^2 shows an unexpected drop in power when all of the outcomes have a true mean difference between intervention and control groups, and the outcomes are correlated. When 25% to 75% of the outcomes have a true difference in means, Hotelling's T^2 and Bonferroni have higher power than the permutation-rank-sum test. The difference between the power curves of the various tests increases as the level of correlation between outcomes increases. When more than 75% of the outcomes have a true mean difference, the power curves cross, and the permutation-rank-sum test becomes more powerful. This occurs for all correlation structures assessed when correlation among outcomes is indeed present.

In instances where all outcomes in the data are simulated with a true mean difference between intervention and control groups, the permutation-rank-sum test consistently has higher power than either the Bonferroni method or Hotelling's T^2 test (Figure 3.2). This holds for all assessed magnitudes of the mean difference, from 0.2 to 0.3. The sign test has even higher power than the permutation-rank-sum test in these scenarios, however recall that the sign test does not adequately control type I error. When the data have clustered correlation, these differences in power between tests are greater when the mean differences are smaller. As is anticipated, power increases as the magnitude of the difference in means increases. We also note that the choice of statistical test has a greater impact on power than the number of outcomes in the data (M).

3.4.2 Non-normal (Skewed) Data

The patterns in type I error rates discussed above hold true for the data simulated with induced skew and kurtosis (Table 3.3). The differences in power between the permutation-rank-sum test and both Hotelling's T^2 and Bonferroni are attenuated when the data are not normally distributed, particularly when 25% to 75% of the outcomes have a mean difference, as shown in Figure 3.3. Hotelling's ^{2} test performs more similarly to the Bonferroni method when the normality assumption is violated. One may infer that further departures from the

Figure 3.1: Estimated power for multivariate normal data with a mean difference of 0.3 across varying proportions of outcomes.

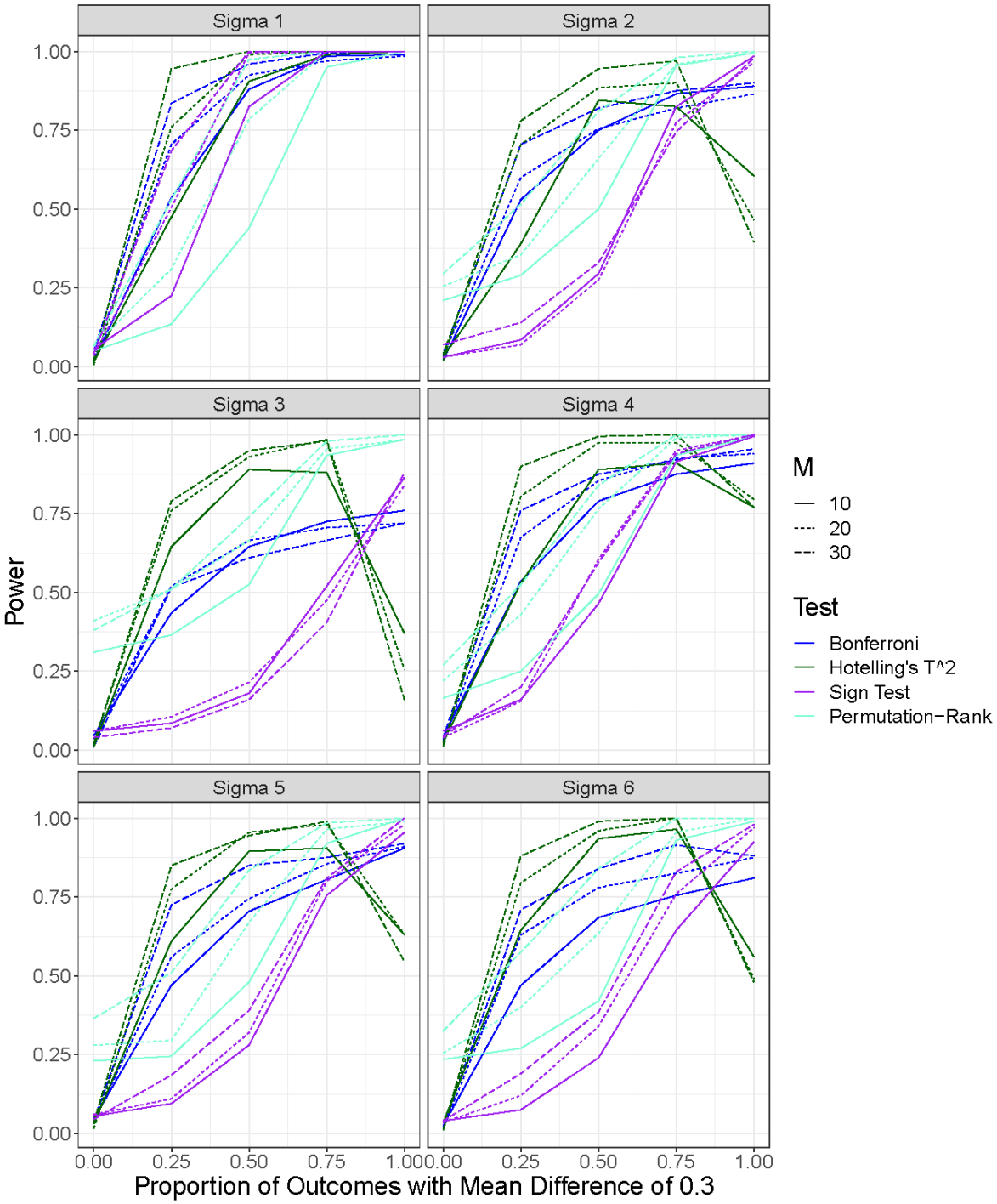


Figure 3.2: Estimated power for multivariate normal data with varying magnitudes of mean differences across all outcomes.

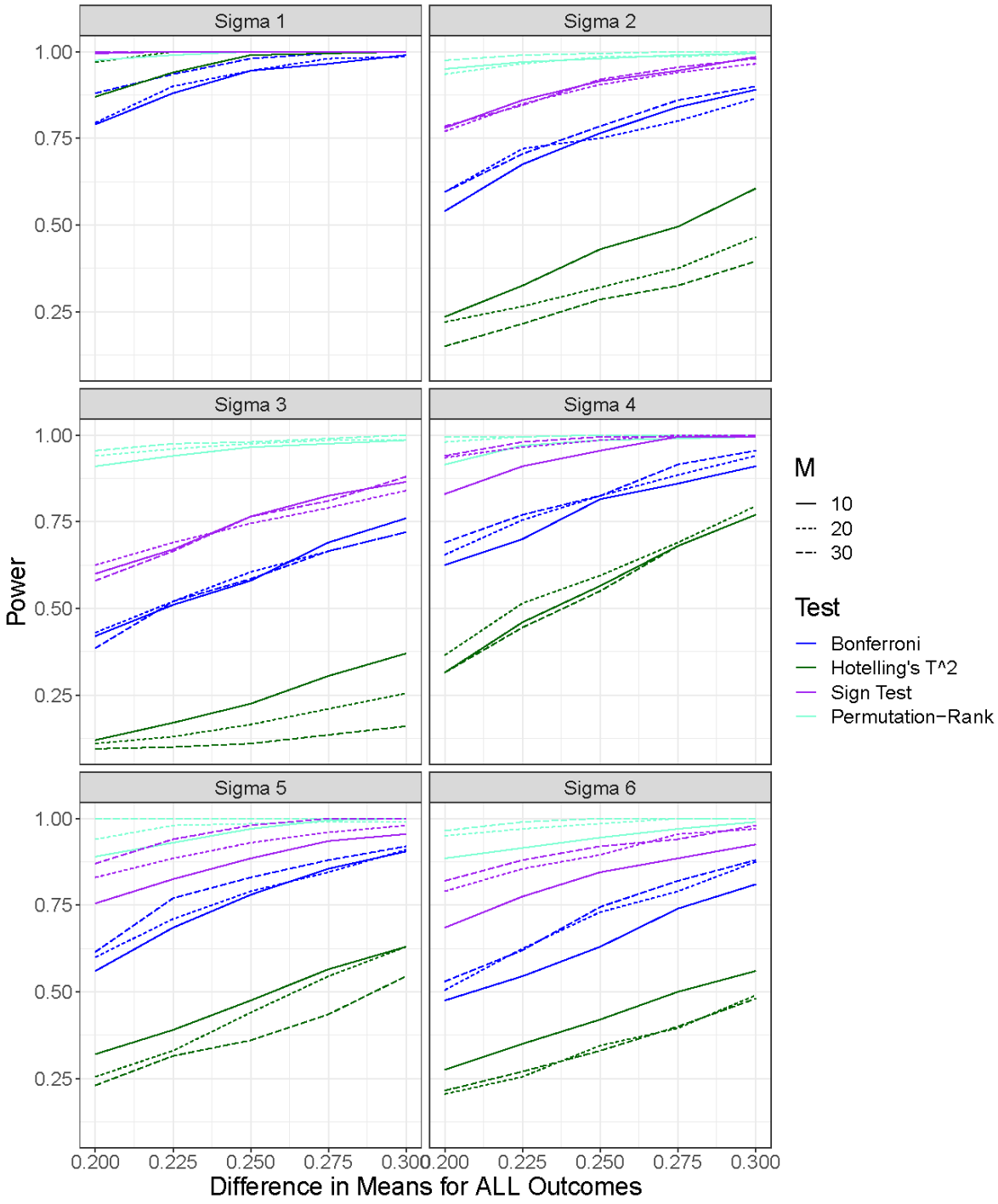
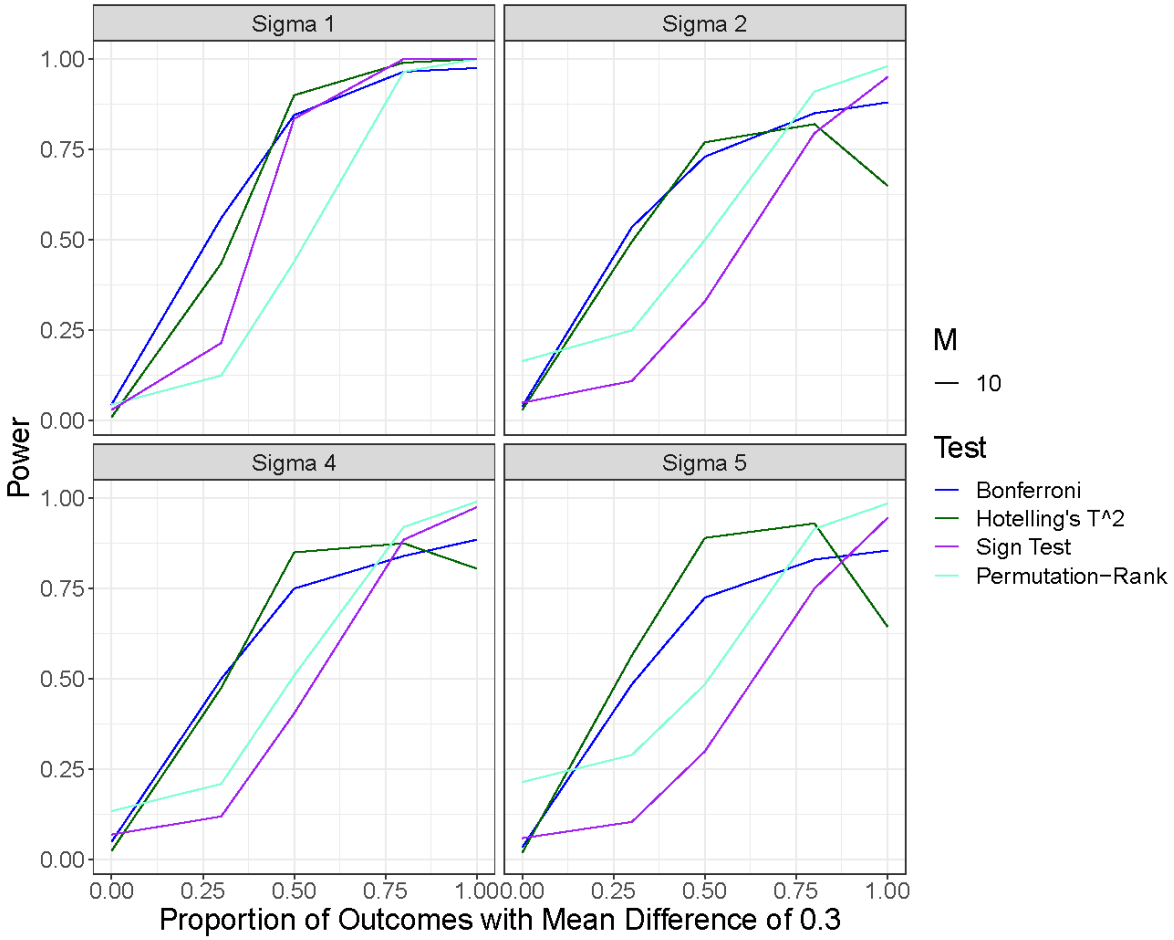


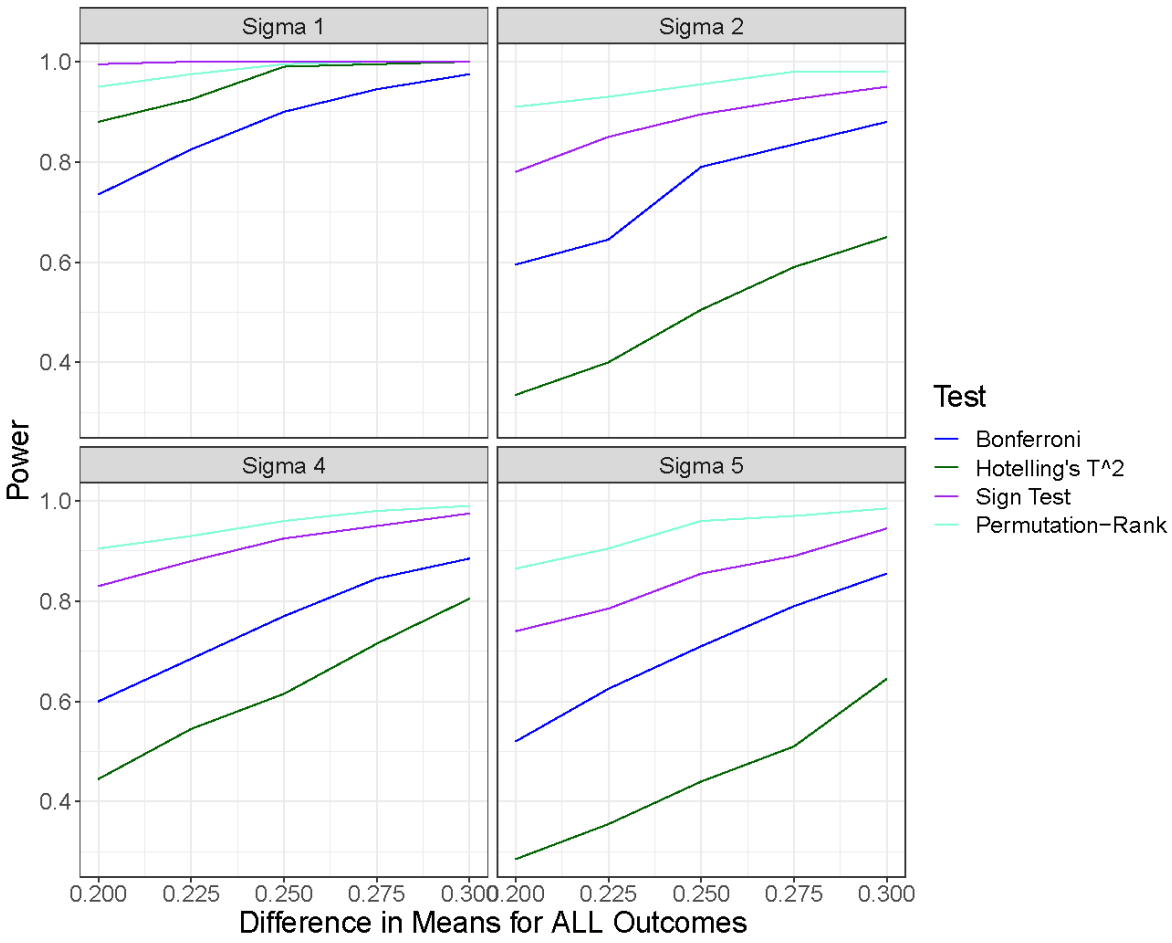
Figure 3.3: Estimated power for skewed data with a mean difference of 0.3 across varying proportions of outcomes.



normal distribution, such as a bimodal distribution, may further shift the power curves in the observed direction. When all outcomes have a mean shift, we again see a larger difference in power between tests when the mean difference is 0.2 than when the mean difference is 0.3 (Figure 3.4). And as seen with the multivariate normal data, when all of the outcomes are simulated with a mean difference the non-normal data show that the permutation-rank-sum test retains higher power than either Hotelling's T^2 or Bonferroni. In fact, the permutation-rank-sum test has higher power with non-normal data than with normally distributed data in this instance. Even with an effect size (cohen's d) as small as 0.2, the permutation-rank-sum test has approximately 80% chance of successfully identifying the intervention effect.

Overall, the permutation-rank-sum test controlled type 1 error and was reasonably pow-

Figure 3.4: Estimated power for skewed data with varying magnitudes of mean differences across all outcomes.



erful across a range of different outcome numbers and correlation structures. Importantly, the permutation-rank-sum test was more powerful than alternative tests like Hotelling's T^2 and Bonferroni when a relatively high proportion of outcomes had relatively smaller effect sizes.

3.5 Application to FOCUS-EC data

To further demonstrate the usefulness of the permutation method, we apply it to data from a behavioral health intervention called FOCUS-EC, described previously. A randomized controlled trial was conducted to ascertain the benefits of the intervention, and 12 distinct measures were included in the primary outcome analysis (Hajal et al., 2020). Each of the 12 measures of parental wellbeing were assessed for both the mother and father, leading to 24 total outcomes. The measures consisted of 2 subscales from the Brief Symptom Inventory which measure depression and anxiety, 3 subscales from the Parental Stress Index, 4 subscales from the Posttraumatic Diagnostic Scale, 1 scale measuring sensitivity of parenting from the Parental Behavior with Preschooler Q-Sort, and 2 codes from an observed parent-child interaction. There were 4 time points at which surveys could be administered; baseline, 3-, 6-, and 12-month follow up, however some measures were only assessed at 2 or 3 of the time points. The outcomes were assessed using longitudinal models to detect differences over time. Of main interest was the difference between baseline and last available follow up.

Following the model structure as described by Mogil et al, we implemented linear mixed effects models in R to estimate intervention effects. The models included time point as a repeated, within-subject factor, as well as intervention group and child gender as between subject factors, and an interaction between intervention group and time point. A compound symmetry covariance structure was used. We calculated one-sided p-values for each outcome using the test statistic and degrees of freedom of the estimated model coefficients. This ensures that intervention and control group differences with directionality opposite of expected, i.e. showing greater improvement in the control group than the intervention group, do not contribute towards our overall determination of a successful intervention effect.

As seen in Table 3.4, which details the results of the FOCUS EC intervention including the one-sided p-values estimated from the longitudinal models, 11 of the 24 estimates are statistically significant at the 0.05 α level. If we use the Bonferroni threshold of $0.05/24 \approx 0.0021$, then only 2 estimates retain statistical significance, namely the decrease in PTSD total score reported by mothers and the increase in the mothers' observed affective behavior. Using the Bonferroni method to adjust for multiple testing still leaves us questioning whether we can claim overall efficacy of the FOCUS EC intervention based on changes observed for 2 of the 24 measured outcomes. Furthermore, we do not have a good way of estimating the correlation between the test statistics from the various longitudinal models in order to utilize Hotelling's T^2 test. And as discussed previously, conducting a sign test using the model estimates to make an overall conclusion about intervention efficacy does not control the type I error rate.

To determine overall intervention efficacy, we apply the permutation-rank-sum test to the FOCUS-EC data, using the longitudinal models described above to estimate intervention effects for each of 5,000 permuted datasets. The overall p-value calculated using this method was 0.002, leading us to conclude that the intervention did improve parental psychological health, parent-child relationships, and child behavior when compared to the control group.

3.6 Discussion

As demonstrated with the FOCUS EC example, the permutation-rank-sum test is flexible enough to use with any type of outcome model, including the ability to use differing models for the various outcomes of interest. We also do not need to determine the structure of the correlation among the test statistics, as is required for Hotelling's T^2 test. Estimating the correlation of the test statistics may be unreliable, especially when fitting various outcome models. This flexibility arises from the permutation-rank-sum's reliance on ranking sets of p-values with indifference to the method of p-value generation. We recommend using one-sided p-values when performing the permutation-rank-sum test, so that one can ensure the directionality of the outcome estimates are in line the expected directionality of the

intervention effect. If this is not done, a small p-value that gives evidence of the intervention adversely affecting an individual outcome could misleadingly add to the collective evidence of an overall, positive intervention effect when looking at the collection of outcomes. In general one should inspect the directionality of the intervention effects for all outcomes and the anticipated direction corresponding to a positive intervention effect should be determined a priori during study planning.

The permutation-rank-sum test is somewhat unique in the way evidence towards rejecting the null hypothesis is accumulated. In multiple testing methods such as Bonferroni and False Discovery Rate, adjustment for multiple hypothesis tests is made by requiring stronger evidence for each individual hypothesis test in order to demonstrate significance. The permutation-rank-sum test however allows a small amount of evidence of a mean difference for many outcomes to collectively suggest an overall intervention effect. That is to say, the larger the proportion of outcomes that show a difference between treatment and control group, the more confident we are that the intervention is the cause of the difference in means. We believe that this aligns with our intuition as researchers when examining results from a study assessing BHI effectiveness. Experts in this area have often questioned whether the conclusion of a non-significant overall intervention effect is actually warranted despite observing a relatively high proportion of outcomes in the expected direction - higher than might reasonably be anticipated by chance alone. The permutation-rank-sum test gives us a valid method for operationalizing this intuition in the form of a statistical hypothesis test.

When designing BHI, researchers will choose outcomes measures that they expect the intervention to affect, even if the effect size may be small relative to the within-subject variability of the measure. Furthermore it is advisable when designing BHI not to measure additional outcomes that are unlikely to be affected by the intervention, as this is similar to a "fishing" expedition. This set up is in contrast to genetic studies, which also have many correlated outcomes, but only a few of the outcomes are expected to show a significant result. The permutation-rank-sum test is most useful in detecting an intervention effect when there are many outcomes which display differences between intervention and control groups, as is

the case with a well-designed BHI. In fact, these are the exact instances in which many BHI researchers would mistakenly conclude no intervention effect. This also has repercussions for how many subjects are needed to power your study. One could propose a powerful study with fewer subjects if it is suspected that the intervention has a small effect on many outcomes that can be measured and the proposed permutation-rank-sum test were used to determine overall intervention efficacy.

Table 3.4: FOCUS EC Outcomes estimated using linear mixed effect (longitudinal) model for Fathers (F) and Mothers (M) comparing the difference in baseline measure and last available follow up between intervention and control groups.

Outcome	**Expected sign**[1]	**Estimate**	**One-sided p-value**	**Signif at 0.05**	**Signif using BF**
Anxiety (F)	(-)	0.08	0.810		
Depression (F)	(-)	0.14	0.905		
Parental Distress (F)	(-)	-1.39	0.187		
Parent-Child Dysfunctional Interaction (F)	(-)	-0.29	0.405		
Difficult Child (F)	(-)	-0.60	0.316		
Sensitive Parenting (F)	(+)	0.88	0.227		
Anxiety (M)	(-)	-0.06	0.204		
Depression (M)	(-)	-0.09	0.139		
Parental Distress (M)	(-)	-1.55	0.090		
Parent-Child Dysfunctional Interaction (M)	(-)	-1.81	0.022	*	
Difficult Child (M)	(-)	-2.09	0.013	*	
Sensitive Parenting (M)	(+)	1.61	0.018	*	
PTSD Total Score (F)	(-)	-0.75	0.358		
PTSD: Re-Experiencing (F)	(-)	-0.15	0.404		
PTSD: Avoidance (F)	(-)	-0.38	0.339		
PTSD: Arousal (F)	(-)	-0.52	0.249		
PTSD Total Score (M)	(-)	-4.40	<0.001	*	*
PTSD: Re-Experiencing (M)	(-)	-1.30	0.004	*	
PTSD: Avoidance (M)	(-)	-1.17	0.023	*	
PTSD: Arousal (M)	(-)	-1.41	0.009	*	
Observed Child Affective Behavior (F)	(+)	0.29	0.041	*	
Observed Parent Affective Behavior (F)	(+)	0.38	0.014	*	
Observed Child Affective Behavior (M)	(+)	0.37	0.004	*	
Observed Parent Affective Behavior (M)	(+)	0.37	0.001	*	*

[1] Expected sign refers to the sign of the estimated coefficient that would indicate improvement in the outcome as a result of the intervention

CHAPTER 4

Adaptome Framework: Development and Performance

4.1 Introduction

Our goal is to develop a statistical framework to provide feedback and guide decisions for ongoing intervention implementation, therefore improving population-level outcomes. Chambers and Norton call this hypothetical framework the "adaptome" (pronounced "adapt-ohm"), and here we adopt his phraseology. Statistical advancements are needed in the utilization of real-world data. Not only must we take into account the covariate imbalances among intervention groups that arise from non-randomized intervention participation, we must also consider statistical error rates when conducting multiple comparisons across intervention versions and repeatedly comparing intervention versions as data accumulates over time.

In creating an adaptome framework for statistical analysis, we consider a few key features. A Bayesian approach is chosen as an easy way to seamlessly incorporate historical data, using it to inform prior distributions for estimated parameters. Importantly, we want to be able to identify which version of the treatment/intervention produces best results for the population targeted by the intervention. We also want to avoid bias in estimated outcomes that are caused by confounders and maintain low overall statistical error rates. To address this challenge we propose combining an existing method for covariate balancing, called entropy balancing, with a Bayesian adaptive platform clinical trial framework.

Covariate balancing is an umbrella term that includes a variety of statistical methods used when analyzing real-world data to account for group differences that may lead to biased estimates when comparing outcomes among groups. Common covariate balancing methods

include matching, propensity scores, and more recently, entropy balancing (Stuart, 2010; Austin, 2011; Hainmueller, 2012).

Historically, matching and propensity score methods have been widely used, however they come with a few drawbacks. Matching methods often result in omitting data from the analysis when individuals in the intervention group do not have an exact covariate match in the control group (Stuart, 2010). Propensity score methods involve estimating an individual's probability of receiving an intervention based on the measured covariates, and subsequently conditioning on these probabilities (Austin, 2011). However, propensity score methods do not directly match covariate distributions between groups, and require the analyst to manually check covariate distributions and then iteratively update the propensity score model accordingly(Hainmueller, 2012). Furthermore, there is no consensus in the literature on how to validate propensity score model specification (Austin, 2011; Lee et al., 2010) and a misspecified propensity score model can lead to biased outcomes(Drake, 1993). For these reasons, we have chosen entropy balancing as our preferred method of covariate balancing, although alternative methods could be considered with minimal modifications to the broader framework being proposed.

The previously described methods for observational data typically focus on a single comparison. Extensions exist for comparing multiple groups at once, but only at a single time point (Feng et al., 2012; McCaffrey et al., 2013). We seek to combine methods for balancing covariates in real-world data with methods for repeatedly comparing multiple outcomes across an indefinite amount of time.

In the clinical trial world, methods exist for controlling type I errors when making repeated comparisons over time. These are called group sequential methods and were first introduced by Pocock in 1977, then expanded upon by O'Brien and Fleming in 1979 (Pocock, 1977; O'Brien and Fleming, 1979). Over the following decades, extensive literature on the subject has been developed including contributions by many others (Müller and Schäfer, 2001). These methods are instrumental in reducing the overall cost of clinical trials by optimizing power while reducing sample size. A more recent extension of these group sequential

methods, which has been gaining popularity, is the adaptive platform trial.

An adaptive platform trial is one in which multiple treatments are assessed simultaneously, and treatments can be added or dropped for futility at interim analyses. A more extensive introduction to adaptive platform trials is given by Angus et al (Angus et al., 2019). Multiple authors have established the benefits of using adaptive platform trials in the clinical setting (Lin and Bunn, 2017; Madariaga et al., 2021; Saville and Berry, 2016). The process of adding and dropping treatments over time and comparing multiple treatments at a given point in time, as is done in the platform trial, resembles the natural way in which data collection and analysis would occur during real-world implementation of behavioral health interventions. Instead of comparing different treatments, we compare adaptations of the intervention to determine which adaptation(s) might be most advantageous.

By combining the statistical methods used in an adaptive platform trial with covariate balancing methods, we can extend the analysis framework to accommodate real-world data. Another adjustment to the platform trial framework that we must make in order to use it for our real-world data is to consider a platform trial with no defined stopping point. Data must be allowed to accumulate in perpetuity. However, group sequential methods rely on having a defined trial stopping point and a pre-set number of interim analyses at defined time points. Using this information, the critical boundaries for hypothesis testing are adjusted accordingly. Without a defined stopping point, and with an unknown number of ongoing interim analyses, a new approach is needed to ensure we are making good decisions along the way, and not being misguided by frequent type I statistical errors. Instead of focusing on type I error, we consider alternative measures of performance including the conditional probability that an action taken throughout the course of the platform trial is beneficial.

Using these extensions to the platform trial analysis framework, we create what we call an adaptome framework, which provides a way to compare intervention adaptations over time as data are continuously collected, while accounting for group differences that are inherent in the use of real-world data. Extensive simulations are often used to assess the properties and performance of platform trial designs prior to implementation (Hummel et al., 2015;

Saville et al., 2014). Similarly, we use simulations to examine our proposed methodology and advocate for use of simulations to guide practical implementation of the framework.

4.2 Methods

We consider a vector $\boldsymbol{Y}$ of outcomes y_i for a set of N individuals, $i = 1, ..., N$. This will represent the data that are collected before an interim analysis. We assume these N individuals all received some form of the intervention of interest, and that there are multiple adaptations of this intervention. Let t_{ik} indicate whether individual i received intervention k for $k = 1, ..., K$ where K denotes the number of available intervention adaptations for which there is sufficient data to be included in the analysis. Thus $\boldsymbol{t}_i$ will represent the vector of K indicators for a given individual who receives one of the possible intervention versions. We refer to these intervention versions as the 'active' versions, to distinguish them from other possible intervention adaptations for which data are *not* actively being collected. Denote the sample size of each of the mutually exclusive intervention groups as N_k, $k = 1, ..., K$ such that $N = \sum_{k=1}^{K} N_k$.

We also require information on individual characteristics or factors that may be associated with receipt of the intervention and intervention outcomes. Call these measured covariates $\boldsymbol{x}_1, ..., \boldsymbol{x}_J$ and let x_{ij} indicate the observed value of covariate j for individual i, $j = 1, ..., J$ and $i = 1, ..., N$. We assume that these J covariates will have different distributions among each intervention group since the individuals were not randomized. The proposed analysis method involves two steps implemented at each interim analysis: a covariate balancing step, and an intervention effect estimation step.

4.2.1 Step 1: Entropy Balancing

In entropy balancing (EB), we estimate a set of weights for the individuals in the intervention group of interest such that the covariate moments of the sample directly match the covariate moments of the target population (Hainmueller, 2012). We define the target

population as the entire population of individuals who could receive the intervention. This could be a population of students within a school district, as in our example, or a population of patients with a particular condition. We assume information about covariate distributions within this population is available. Let μ_j be the mean of covariate j in the target population for, $j = 1, ..., J$. For this application, we match covariate means but it is possible to match means in addition to higher covariate moments, such as variance. Choice of the number of covariates and which, if any, to select for higher order covariate moments is up to the analyst and requires consideration of parsimony in light of the limitations/richness of the data available. The entropy balancing weights for intervention group k are estimated such that:

$$\begin{aligned}
\left| \frac{1}{N_k} \sum_{i=1}^{N} w_i x_{i1} \mathbb{1}(t_{ik} = 1) - \mu_1 \right| &\leq \delta \\
\left| \frac{1}{N_k} \sum_{i=1}^{N} w_i x_{i2} \mathbb{1}(t_{ik} = 1) - \mu_2 \right| &\leq \delta \\
&\vdots \\
\left| \frac{1}{N_k} \sum_{i=1}^{N} w_i x_{iJ} \mathbb{1}(t_{ik} = 1) - \mu_J \right| &\leq \delta
\end{aligned}$$

where w_i denotes the estimated weight for individual i and δ represents some small amount of tolerance for the constraint. One set of weights for the individuals in intervention group k must simultaneously satisfy all of the above conditions. The chosen amount of tolerance may depend on the application at hand, since an estimable set of weights is not guaranteed for any arbitrarily small δ. Entropy balancing is performed separately for each intervention group and a set of weights is estimated for the individuals in each group. The same set of target means, $\mu_1, ..., \mu_J$, is used for each intervention group and at each interim analysis. Once the entropy weights are estimated, the weighted observations are then used in analysis of the outcome.

4.2.2 Step 2: Intervention Effect Estimation

Estimating the relative intervention effectiveness is done through fitting a Bayesian statistical model. For demonstration, we assume the outcome is normally distributed, and use a classic Normal-Inverse Gamma conjugate model. If the outcome is binary, a Beta-Binomial conjugate model can be used in a similar fashion. The Normal-Inverse Gamma model we implement is as follows:

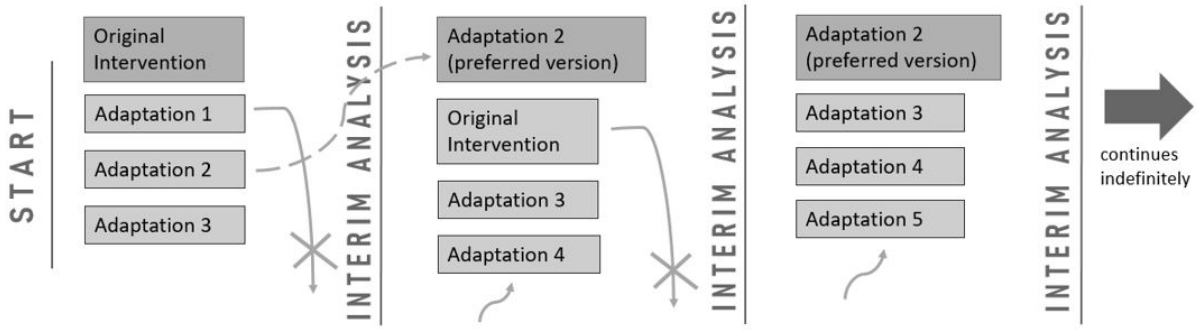
$$y_i \sim N\left(\boldsymbol{t}_i^T\boldsymbol{\theta}, \frac{1}{\tau w_i}\right)$$
$$\theta_k \sim N\left(\mu_{0,k}, \frac{1}{\sigma_{0,k}^2}\right)$$
$$\tau \sim Gamma\,(a, b)$$

Here tau is the precision, which is defined as the inverse of the variance, and is multiplied by the weight estimated through entropy balancing. The parameter $\boldsymbol{\theta}$ is a vector containing the estimated mean outcomes for each intervention group, the elements of which are denoted θ_k, for $k = 1, ..., K$ available intervention adaptations. This model assumes a constant variance for the outcomes across intervention versions.

Each element in $\boldsymbol{\theta}$ has a Normal prior distribution. For new intervention adaptations that have no prior data collected, we use a vague prior, and for others we calculate prior parameters from the estimates in the previous interim analysis. Similarly for τ, we use a vague prior for the first analysis, and subsequently use estimates of the population variance to inform priors going forward. Posterior samples for the unknown parameters $\boldsymbol{\theta}$ and τ in the above model are generated using Markov Chain Monte Carlo (MCMC) estimation. The estimated mean outcome for each intervention adaptation is simply the mean of the posterior sample for each θ_k.

At an interim analysis, the active intervention adaptations are compared using a set of pairwise comparisons. One intervention adaptation is deemed the "preferred" version and the remaining adaptations are compared to the preferred version. At the first analysis, the

Figure 4.1: Diagram of platform trial flow.



original intervention, or the adaptation that most closely resembles the original, is deemed the preferred version. Going forward, if any intervention adaptation is found to be superior to the current preferred version based on meeting an established superiority criteria, to be defined shortly, it becomes the new preferred version. If more than one intervention adaptation meet the superiority criteria at the same interim analysis, then the one with the most favorable estimated mean outcome is chosen as the new preferred version.

Intervention adaptations can also be dropped for futility at interim analyses if they meet an established inferiority criteria, as defined below. If multiple intervention adaptations simultaneously meet the inferiority criteria, all are dropped. New intervention adaptations can be immediately added to replace the ones dropped, or as they naturally arise. Figure 4.1 diagrams how the process of adding and dropping intervention adaptations and switching preferred versions can occur throughout the platform trial as a result of sequential interim analyses.

The posterior sample is used to draw inference about intervention adaptation superiority and futility. Without loss of generality, we assume going forward that higher mean outcomes are better. Define ν as the probability that θ_k is larger than θ_{pref} the mean outcome of the preferred version. Let $\hat{\theta}_{m,k}$ be the estimated mean outcome of intervention adaptation k based on the m^{th} sample from the posterior, and let M be the total number of samples from the posterior. Then the estimated probability that intervention adaptation k is superior to

the preferred version is calculated as

$$\hat{\nu} = \frac{1}{M} \sum_{m=1}^{M} \mathbb{1} \left(\hat{\theta}_{m,k} > \hat{\theta}_{m,pref} \right)$$

If $\hat{\nu}$ is higher than a specified threshold, such as 0.9, we say the superiority criteria has been met. Contrastingly, if this probability is lower than a separate specified threshold, such as 0.1, we say that the inferiority criteria has been met. Preferred version switches and futility drops are then made accordingly.

Finally the results of this interim analysis are used to inform priors for the next analysis once more data have been collected. The prior parameters for the elements of θ are:

$$\mu_{0,k} = E\left[\hat{\theta}_k\right]$$
$$\sigma^2_{0,k} = \frac{1}{N_k} Var\left[\hat{\theta}_k\right]$$

We scale the variance prior so that as we accumulate more data for a given intervention adaptation, we put more emphasis on the prior. However for practical purposes we put a cap on the value of N_k used in this equation so that the prior variance does not eventually become too close to zero as large amounts of data accumulate. To calculate prior parameters for τ we use properties of the Gamma distribution that relate its shape parameters to the mean and variance of the distribution.

$$a = \frac{E\left[\hat{\tau}\right]^2}{Var\left[\hat{\tau}\right]}$$
$$b = \frac{E\left[\hat{\tau}\right]}{Var\left[\hat{\tau}\right]}$$

This method is meant to be carried out in perpetuity. Interim analyses are conducted continuously as long as data continue to be collected.

4.3 Simulation Study Design

To evaluate this method we simulate data in which the measured covariates influence the individual's probability of receiving each intervention adaptation and also influence the outcome. After one set of data is simulated, we use the two-step framework detailed above to assess the relative effectiveness of intervention adaptations. For simplicity, we allow up to four active intervention adaptations at a time. If an intervention adaptation is found to be superior to the current preferred version, then it becomes the new preferred version. If any intervention adaptations for which data are being actively collected meet the futility criteria, they are dropped and promptly replaced with new adaptations. A subsequent set of data are then simulated. This cycle continues until either a maximum overall sample size is reached, or there have been three analyses after the last available adaptation has been added. In its entirety, we refer to this as one simulated trial. The architecture of the platform trial simulation stems from an online clinical trial resource called the Highly Efficient Clinical Trials Simulator (Thorlund et al., 2019).

The data are generated following a simulation design used by Hainmeuller, which builds upon that of Markus Frolich (Hainmueller, 2012; Folich, 2007). There are six covariates with the following distributions.

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \sim MVN \left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 2 & 1 & -1 \\ 1 & 1 & -0.5 \\ -1 & -0.5 & 1 \end{bmatrix} \right)$$

$$x_4 \sim Unif\,(-3, 3)$$

$$x_5 \sim \chi^2_1(0)$$

$$x_6 \sim Binom\,(1, 0.5)$$

Covariates x_4, x_5, and x_6 are all independent from one another and from the multivariate normal covariates. Following generation of a set of covariates, observations are binned into intervention groups using a multinomial logit model. Let R be the number of possible

intervention versions over the entire simulated trial. Note that the K active intervention versions at a given interim analysis will be a subset of the R possible versions. And let ϵ_1,..., ϵ_{R-1} be $R-1$ independent normally distributed random variables representing random noise added to the system.

$$
\begin{aligned}
log\left(\frac{Pr(t_{i1}=1)}{Pr(t_{iR}=1)}\right) &= \boldsymbol{\beta}_{.1}\boldsymbol{x}_{i.} + \epsilon_1 \\
log\left(\frac{Pr(t_{i2}=1)}{Pr(t_{iR}=1)}\right) &= \boldsymbol{\beta}_{.2}\boldsymbol{x}_{i.} + \epsilon_2 \\
&\vdots \\
log\left(\frac{Pr(t_{i(R-1)}=1)}{Pr(t_{iR}=1)}\right) &= \boldsymbol{\beta}_{.(R-1)}\boldsymbol{x}_{i.} + \epsilon_{(R-1)}
\end{aligned}
$$

where $\boldsymbol{x}_{i.}$ represents the vector of all 6 covariates for individual i and $\boldsymbol{\beta}_{.1}$ represents the first column of the matrix $\boldsymbol{\beta}$. Thus, the probability of observation i receiving each intervention adaptation is calculated as:

$$
\begin{aligned}
Pr(t_{i1}=1) &= \frac{e^{(\boldsymbol{\beta}_{.1}\boldsymbol{x}_{i.}+\epsilon_1)}}{\rho} \\
Pr(t_{i2}=1) &= \frac{e^{(\boldsymbol{\beta}_{.2}\boldsymbol{x}_{i.}+\epsilon_2)}}{\rho} \\
&\vdots \\
Pr(t_{i(R-1)}=1) &= \frac{e^{(\boldsymbol{\beta}_{.(R-1)}\boldsymbol{x}_{i.}+\epsilon_{(R-1)})}}{\rho} \\
Pr(t_{iR}=1) &= \frac{1}{\rho}
\end{aligned}
$$

where ρ is defined:

$$
\rho = 1 + e^{(\boldsymbol{\beta}_{.1}\boldsymbol{x}_{i.}+\epsilon_1)} + e^{(\boldsymbol{\beta}_{.2}\boldsymbol{x}_{i.}+\epsilon_2)} + \cdots + e^{(\boldsymbol{\beta}_{.(R-1)}\boldsymbol{x}_{i.}+\epsilon_{(R-1)})}.
$$

The β coefficients represent the direction and degree to which each covariate affects the probability of receiving a given intervention adaptation. These have been arbitrarily chosen for the simulation study but can be based on prior knowledge or preliminary data analysis

when using similar simulations for the purpose of developing an analysis plan for practical application. The important aspect is inducing correlation between the covariates and receipt of an intervention adaptation. We used the same set of β coefficients for each simulated trial, but shuffle the order of the probabilities once per simulated trial in order to yield results that are balanced, and do not depend on an arbitrary order of covariate distributions among the intervention groups. The set of coefficients and parameters for generating the random noise variables are detailed in the appendix. Note that the variance of ϵ_1,..., ϵ_{R-1} affects the amount of covariate overlap among the intervention groups, and sufficient overlap is needed in order to successfully estimate weights using entropy balancing.

Next it is necessary to induce correlation between the covariates and the intervention outcome. We use two different outcome models, based on the simulation designs in (Hainmueller, 2012).

$$\text{Linear: } z_i = x_{i1} + x_{i2} + x_{i3} - x_{i4} + x_{i5} + x_{i6} + \eta_i$$
$$\text{Non-linear: } z_i = x_{i1} + x_{i2} + 0.2x_{i3}x_{i4} - \sqrt{x_{i5}} + \eta_i$$

where η_i is another normally distributed random variable used to add random noise into the system. The intervention effects are then added to these base outcomes, z_i. Let the set of intervention effects be denoted as $\boldsymbol{\alpha}$. We consider a range of possible intervention effects spaced at regular intervals: $\boldsymbol{\alpha}$ = (0.0,0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8). Thus we have differing effects for nine possible intervention adaptations, and $y_i = z_i + \boldsymbol{\alpha t}_i$. Table 4.1 shows how the outcome is calculated as a combination of the dependency on the covariates and the intervention effect in the simulated data. It also displays how the raw, unweighted mean outcome for each group, denoted $\bar{y}$ differs from what the mean would be if the covariates were perfectly balanced, θ^*. The true mean outcome, θ^*, is calculated by adding the fixed values of 1.5 and 0.8 to the alpha component for linear and nonlinear outcome models, respectively. θ^* is the quantity we wish to estimate and draw inference on using our adaptome statistical framework.

Table 4.1: Outcome designs for simulations (increasing order of intervention effectiveness).

Outcome Design	Intervention Version	$\bar{z}^a$	Corresponding element in $\boldsymbol{\alpha}^b$	$\bar{y} = \bar{z} + \boldsymbol{\alpha t}$	θ^*
Linear	1	1.36	0.0	1.36	1.5
	2	2.00	0.1	2.10	1.6
	3	1.79	0.2	1.99	1.7
	4	1.33	0.3	1.63	1.8
	5	1.14	0.4	1.54	1.9
	6	1.50	0.5	2.00	2.0
	7	2.32	0.6	2.92	2.1
	8	0.60	0.7	1.30	2.2
	9	1.60	0.8	2.40	2.3
Non-linear	1	-1.23	0.0	-1.23	-0.8
	2	-0.49	0.1	-0.39	-0.7
	3	-1.30	0.2	-1.10	-0.6
	4	-0.27	0.3	0.03	-0.5
	5	-2.25	0.4	-1.85	-0.4
	6	-0.31	0.5	0.19	-0.3
	7	0.54	0.6	1.14	-0.2
	8	-0.49	0.7	0.21	-0.1
	9	-0.92	0.8	-0.12	0.0

[a] This shows one possible order of the covariate dependencies for different intervention versions. The order is shuffled for each simulated trial.
[b] For trials with a random order of intervention effectiveness, this column is also shuffled once per trial.

We also consider the order in which adaptations are added into the platform trial. We conduct simulated trials with intervention adaptations added in increasing order of effectiveness, and with random orders of effectiveness, where the elements in $\boldsymbol{\alpha}$ are shuffled once at the beginning of each trial. In some instances, it may be reasonable to assume that the majority of adaptations introduced will confer greater benefit relative to the previously introduced adaptations, thus our choice to simulate an increasing order of effectiveness. On the other hand, the random ordering of the intervention effects allows us to break any arbitrary relationship between the intervention order and the performance of the platform trial.

We simulate trials prospectively by starting with four active intervention groups, conducting routine interim analyses, and adding new intervention groups as others are dropped

for futility. For each interim analysis, we generate a new set of data then select only data for the active intervention groups and analyze according to the methods enumerated above. We use a constant sample size, 300, for the number of observations per intervention group, per interim analysis. For the entropy balancing step, we match the covariate means to means from a large external sample which is representative of known population-level covariate distributions.

We also compare three variations of the Bayesian model in step 2. The first is exactly the model described above, and is referred to as the model with "EB weights." The second includes the covariates in the Bayesian model as follows:

$$\text{EB + covariates: } y_i \sim N\left(\boldsymbol{t}_i^T\boldsymbol{\theta} + \boldsymbol{x}_i\boldsymbol{\gamma}, \frac{1}{\tau w_i}\right)$$

where γ denotes the estimated residual effect of the covariates on the outcome, after balancing the covariates. However θ is still the parameter of interest. The third model includes neither the covariates nor the weights estimated through entropy balancing, and is simply:

$$\text{Neither: } y_i \sim N\left(\boldsymbol{t}_i^T\boldsymbol{\theta}, \frac{1}{\tau}\right)$$

This third model gives a point of comparison to determine the benefits of adding in entropy balancing as a data pre-processing step. Meanwhile the second model allows us to determine if there are any added benefits or drawbacks of increasing the model complexity by adjusting for covariates in addition to balancing them in the pre-processing. Table 4.2 summarizes the various set ups for the simulated trials.

We ran 500 simulated trials for each possible combination of the hyper-parameters. The simulations were run in R, using JAGS for the Gibbs sampling (Plummer, 2003). Simulation code that can be used as a guide in implementing the Adaptome framework is available for interested readers at: `https://github.com/tdbufford/Adaptome-Simulations`.

Table 4.2: Simulation set ups.

Hyper Parameter	Possible Variations
Intervention effect order	• Increasing • Random
Bayesian model	• EB weights • EB weights + covariates • Neither
Outcome design	• Linear • Non-linear
N_k	• 300
Thresholds	• Superiority = 0.9 • Futility = 0.1
Max # of active intervention versions	• 4

4.4 Measuring Performance

The typical statistical measures of Type I error rate and power are not well defined for our ongoing Bayesian platform trial. Calculation of these measures is dependent upon having an established number of hypothesis tests, which we do not have because comparison of adaptations is ongoing in response to real-world intervention implementation. To assess performance of this analysis framework, we consider a variety of other measures that describe the decision-making process throughout the trial. Many of these measures describe actions taken, meaning switching which adaptation is the preferred version or dropping an intervention adaptation for futility. For example, we define how a preferred version switch can be sub-optimal, how a futility drop can be sub-optimal, and look at proportions of actions that are sub-optimal. Table 4.3 defines these counts and proportions which help us understand the ability of the Bayesian framework to distinguish between intervention versions. Many of these are also very straightforward, but since we have multiple intervention groups evaluated simultaneously and multiple interim analyses, exact definitions are required for clarity.

Based on these definitions of suboptimal actions, we calculate a *Positive Action Probability*. This probability is the proportion of actions, both superiority switches and futility

Table 4.3: Measure definitions for Adaptome framework performance measures.

Measure	Definition
Mean Sample Size	Average total number of participants over the course of a single trial (all treatment versions combined)
Mean # Analyses	Average number of interim analyses conducted during a single trial
Total # Analyses	Total number of interim analyses conducted across all simulated trials with the same set up
Percent Top One	Percent of platform trials that successfully identified the intervention version with the highest mean outcome as the preferred version at the end of the trial
Percent Top Two	Percent of platform trials that have identified one of the top two intervention versions as the preferred version at the end of the trial
# Superiority Switches	Total number of times a new preferred version was chosen, summed across all simulated trials with the same set up
Sub-optimal Switches	Total number of superiority switches where the true mean outcome of the new preferred version was lower than that of the current preferred version, summed across all simulated trials with the same set up
% Sub-optimal Switches	Percent of superiority switches that are sub-optimal according to above criteria
# Futility Drops	Total number of times an intervention version was dropped for futility, summed across all simulated trials with the same set up
Best Dropped	Number of trials in which the intervention version with the highest mean outcome was dropped for futility
Futility Ties	Total number of times that more than one intervention version was dropped at the same time, summed across all simulated trials with the same set up
Sub-optimal Futility Drops	Total number of times an intervention version with a higher true mean outcome than the current preferred version was dropped for futility, summed across all simulated trials with the same set up
% Sub-optimal Drops	Percent of futility drops that met the above sub-optimality criteria
Sub-optimal Action Probability	Proportion of total actions taken which resulted in either a sub-optimal switch or sub-optimal futility drop.
Positive Action Probability	1−(Sub-optimal Action Probability)

drops, that are in alignment with the true mean outcomes of the intervention versions, i.e. not considered suboptimal. The denominator for this calculation is the total number of actions taken, superiority switches plus futility drops.

$$1 - \frac{\text{Sub-optimal Superiority Switches + Sub-optimal Futility Drops}}{\text{Total Superiority Switches + Total Futility Drops}}$$

In determining whether an action taken is sub-optimal we do not penalize type II statistical errors (failures to assign the optimal intervention version as the preferred version at a given interim analysis), since we allow for additional data to be collected and subsequent analyses to successfully identify true differences in adaptation effectiveness. Interim analyses in which no action is taken are not accounted for in this summary measure. On the other hand, a single interim analysis can result in more than one action taken, and each action is considered separately. In calculating the Positive Action Probability, we combine results from all simulated trial runs with identical set-up. Since the Positive Action Probability conditions on an action being taken it does not require a defined number of interim analyses or a set trial stopping point. It gives applied researchers an understanding of the probability of an error each time they make a change to intervention implementation strategy (dropping a version or making a preferred version switch). Of course, there should be an understanding among researchers using this measure that the probability of taking at least one erroneous action over any prolonged period of time is increased based as a function of the length of time. The Positive Action Probability is to be interpreted at a discrete instant in time.

We also assess the percent of simulated trials that have successfully identified the intervention version with the highest mean outcome as the preferred version by the last interim analysis (*Percent Top One*). This can be considered a secondary measures of success because it depends upon the arbitrary point at which we end the simulations. In some instances, investigators may find value in being able to state that after some definitive amount of time, the analysis plan can be expected to identify the best intervention version with some high level of confidence. While the intention is to use this framework in an ongoing fashion, this statement would provide added assurance that in the long-run preferable intervention

adaptations are being identified.

As with typical simulation studies involving covariate balancing for observational data, we also assess the bias of our estimates. The bias is defined as the difference between the true mean outcome and the estimated mean outcome for each intervention group, using the estimate with the largest sample size for a given trial.

Lastly, we consider the estimated average intervention effect received for a defined number of individuals, assuming we can drop intervention versions for futility and add new ones as data are collected. We call it the *Average Effect Received*. This number can vary depending on the order in which new adaptations are added to the trial, in addition to random variation in the data. Therefore, we look at the distribution of the Average Effect Received for a set of simulated trials where the order of adaptations is varied for each simulation while other parameters are held constant. This metric reflects the potential for short- and long-term population-level benefits and is particularly useful for guiding selection of an optimal interim analysis plans. This last metric will be assessed in the following chapter.

4.5 Simulation Results

Table 4.4 shows that in the simulated trials that include entropy balancing, whether or not we include the covariates in the Bayesian model, the Positive Action Probability is at least 0.978. This means type I errors occur less than 2.2% of the time, which is an improvement upon the standard 5% error rate. These results are especially promising considering we used 0.9 and 0.1 as our superiority and futility thresholds, respectively, which a priori could suggest the Positive Action Probability be no higher than 0.9. In this respect our Adaptome framework out-performs expectations using simulated data, suggesting that it will perform sufficiently well with real data which is likely to be more nuanced than simulated data and may have unmeasured confounders.

Contrastingly, simulations that do not include entropy balancing yield decisions to drop adaptations or switch the preferred version that are often not in alignment with the true

Table 4.4: Simulation results demonstrating performance of Adaptome analysis framework.

Linear Outcome Design						
	Increasing Version Effectiveness			Random Order Effectiveness		
	EB weights	EB + Co-variates	Neither	EB weights	EB + Co-variates	Neither
Mean sample size	11,997	7,777	6,347	7,129	4,972	5,810
Mean # Analyses	11.2	7.2	6.1	6.9	4.7	5.5
Total # Analyses	5,598	3,622	3,044	3,447	2,361	2,728
Percent Top One	**86.0%**	**94.6%**	**24.6%**	**95.6%**	**95.6%**	**25.6%**
Percent Top Two	99.6%	100%	41.6%	100%	100%	42.8%
# Superiority Switches	1,657	1,754	983	621	671	706
Sub-optimal Switches	0	8	15	3	10	260
% Sub-optimal Switches	0%	0.5%	1.6%	0.5%	1.5%	36.8%
# Futility Drops	3,837	3,949	3,946	3,936	3,978	3,965
Best Dropped	2	10	365	8	9	365
Futility Ties	1,037	1,348	1,243	1,210	1,358	1,289
Sub-optimal Drops	8	74	2,427	9	19	1,294
% Sub-optimal Drops	0.2%	1.9%	61.5%	0.2%	0.5%	32.6%
Sub-optimal Action Probability	0.001	0.014	0.500	0.003	0.006	0.333
Positive Action Probability	**0.999**	**0.986**	**0.500**	**0.997**	**0.994**	**0.667**

Non-linear Outcome Design						
	Increasing Version Effectiveness			Random Order Effectiveness		
	EB weights	EB + Co-variates	Neither	EB weights	EB + Co-variates	Neither
Mean sample size	10,522	7,989	5,032	6,419	5,191	4,815
Mean # Analyses	9.9	7.4	4.6	6.3	5.0	4.4
Total # Analyses	4,942	3,720	2,297	3,134	2,482	2,195
Percent Top One	**87.2%**	**92.4%**	**12.0%**	**92.4%**	**92.2%**	**13.6%**
Percent Top Two	98.6%	99.6%	23.0%	100%	99.6%	24.4%
# Superiority Switches	1,703	1,770	755	651	695	703
Sub-optimal Switches	0	12	1	4	32	339
% Sub-optimal Switches	0%	0.7%	0.1%	0.6%	4.6%	48.2%
# Futility Drops	3,858	3,962	3,993	3,940	3,970	3,992
Best Dropped	4	20	434	11	28	431
Futility Ties	1,168	1,334	1,291	1,248	1,361	1,355
Sub-optimal Drops	42	113	2,708	13	39	1,641
% Sub-optimal Drops	1.1%	2.9%	67.8%	0.3%	1.0%	41.1%
Sub-optimal Action Probability	0.008	0.022	0.571	0.004	0.015	0.422
Positive Action Probability	**0.992**	**0.978**	**0.429**	**0.996**	**0.985**	**0.578**

effectiveness of each intervention adaptation. The Positive Action Probability can be as low as 0.429, slightly worse than a 50-50 chance. This point of comparison tells us that we have indeed successfully designed the simulations in a way that induces correlation between covariates, intervention groups, and outcomes such that the covariate differences cannot be ignored in assessing outcomes. It also indicates that using the weights estimated through entropy balancing in our Bayesian model is an effective method of adjusting for covariates in this framework which mimics a platform clinical trial but uses real-world data.

Another indicative measure is the percent of trials that successfully identify the best adaptation by the last interim analysis. Again we see a stark contrast between simulated trials that use entropy balancing and those that do not. In simulations using the linear outcome design without entropy balancing, about 25% of the trials successfully identify the best adaptation. But with entropy balancing, the number increases to about 95%.

We notice that in the trials using just entropy balancing without covariates in the model and with intervention versions added in increasing order of effectiveness, the percent of trials that identify the best adaptation is lower, at 86%. These simulations also have a higher mean number of analyses and mean sample size, but the number of trials that drop the best adaptation for futility is low. This indicates that entropy balancing alone, without covariates in the model, requires more data to successfully distinguish between intervention adaptations with small differences in effectiveness. Interim analyses where no action is taken occur more frequently. This is due to slightly greater variability in the estimated intervention effects at interim analyses when the covariates are not included in the Bayesian model. We see similar trends for the percent of trials that identify the best adaptation when looking at simulations that used the non-linear outcome design. This suggests that adding covariates to the Bayesian model results in more precise estimates of intervention effects and increases the framework's ability to distinguish between versions using smaller amounts of data, which subsequently promotes a faster rate of action being taken.

The rate at which actions are taken is important because swifter discarding of inferior intervention adaptations results in more participants receiving intervention versions with

greater effectiveness. This subsequently results in an increase in the positive outcomes at the population-level. For example, one could look at the average effectiveness of intervention versions received by a population of individuals where different intervention adaptations were implemented over a specified period of time. In order to maximize overall outcomes, we would want the majority of the population to receive the intervention version with the greatest effectiveness, and would want relatively few individuals to receive the less effective adaptations.

On the other end of the spectrum, using solely a linear Bayesian model without entropy balancing to estimate mean outcomes when the outcome design is non-linear results in about 1/9 trials successfully identifying the best adaptation out of 9 possibilities, so the analysis framework performs no better than random chance in this case.

The final measure we wish to discuss is the bias in the outcome estimates, pictured in Figure 4.2 and Figure 4.3. The figures show that using entropy balancing to account for covariate imbalances between groups greatly decreases the bias in the estimated outcomes, whether or not we include covariates in the Bayesian model. When the outcome design is linear, including covariates in the model further shrinks the bias towards zero. However when the outcome design is non-linear, including covariates in the model creates bias in all the estimates. In these simulated trials, the Positive Action Probability and the percent of trials that identify the best adaptation remain high because the bias in the estimates is consistent across adaptations, so comparing relative effectiveness still usually results favorable outcomes even though the estimated magnitude of the intervention effect is incorrect. Researchers may want to consider the linearity of the relationship between their outcome and their measured covariates when deciding how to implement the Adaptome framework.

4.6 Discussion

These simulations demonstrate the adaptome framework's ability to successfully reduce bias by balancing confounding covariates while maintaining a high Positive Action Prob-

ability in making decisions about comparative intervention version effectiveness. In these simulations we assume, however, that all confounding covariates are known and measured, which is a strong assumption, but necessary in observational data methods.

In order to implement the described statistical analysis in a real world setting, several decisions need to be made before putting the analysis plan into action. These decisions include determination of when to conduct interim analyses and specification of optimal superiority and futility thresholds. Similar to undertaking power calculations prior to conducting a randomized study, these decisions should be made in the design-stage and driven by consideration of various possibilities and their resulting impact on established performance benchmarks. Simulations similar to the ones described in this chapter may be used in the planning of ongoing intervention implementation with interim analysis and feedback to intervention providers. An example of how study design simulations might be approached and implemented is detailed in the subsequent chapter for a specific BHI application.

Figure 4.2: Bias in estimates for linear outcome design.

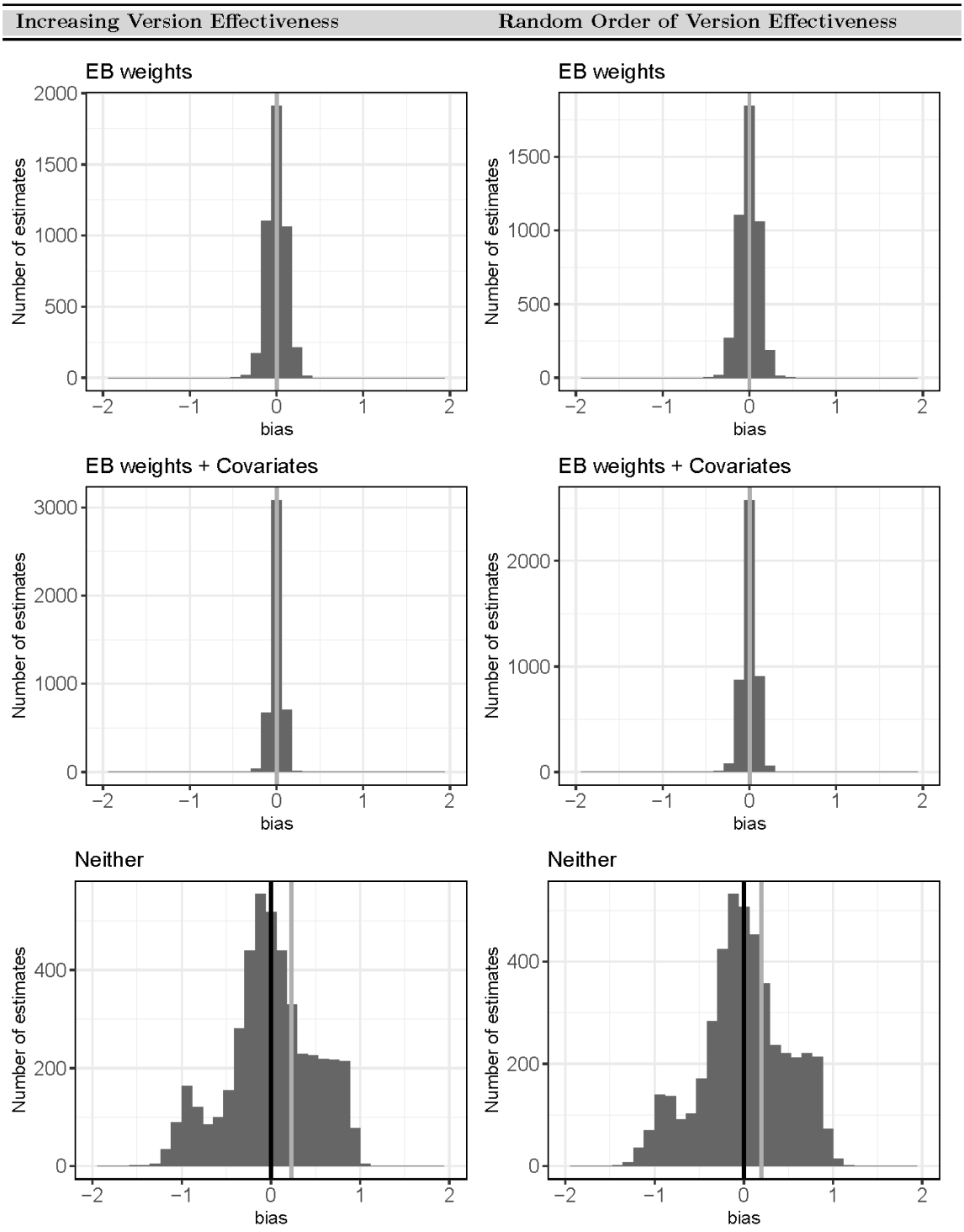
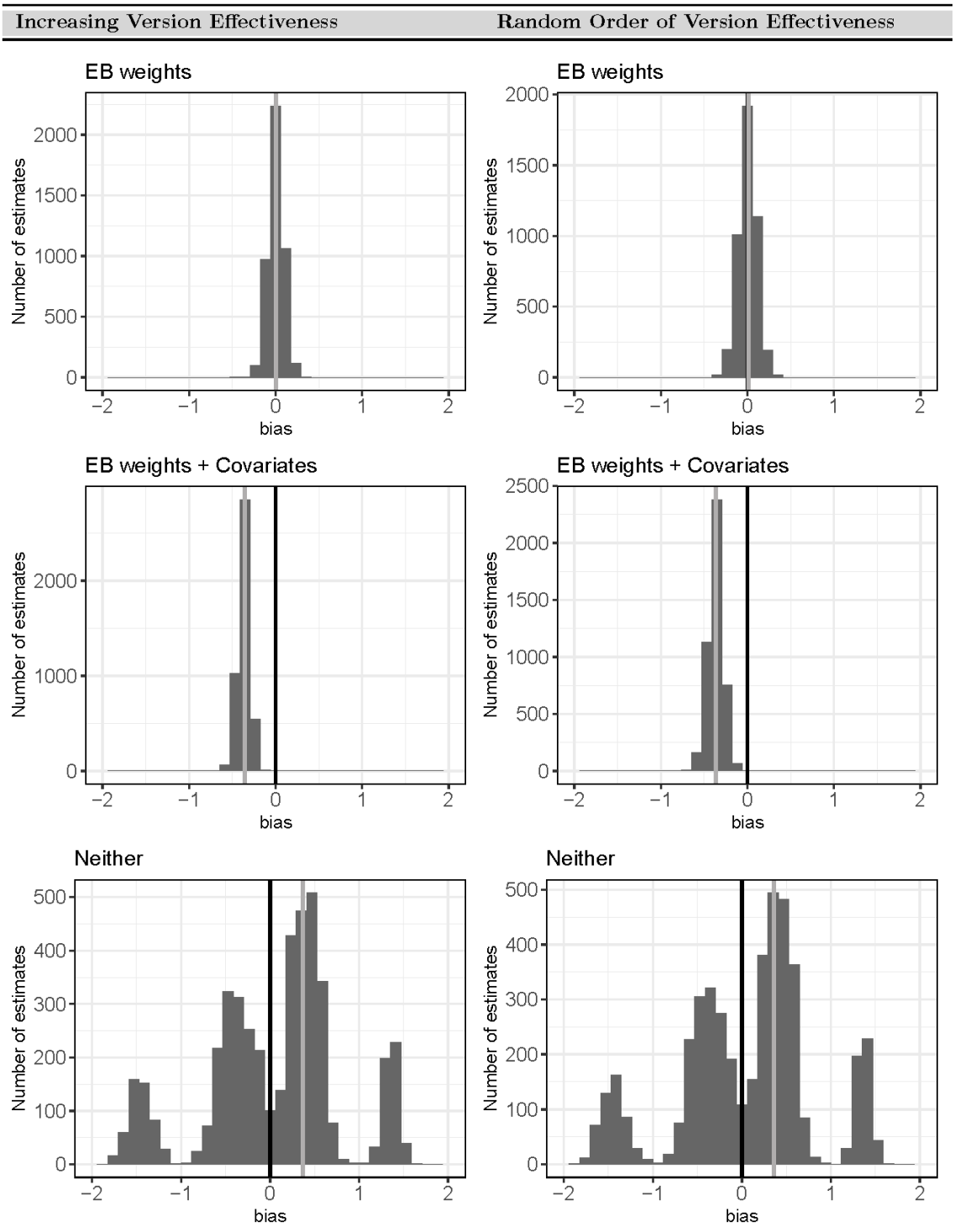


Figure 4.3: Bias in estimates for non-linear outcome design.



CHAPTER 5

Adaptome Framework: Application to School-based Resilience Program

5.1 Description of School-Based Trauma-Informed Preventive Intervention

The preventive intervention called Families OverComing Under Stress (FOCUS) was developed and implemented as a suite of resilience services for military-connected families and youth (Beardslee et al., 2011; Lester et al., 2011; Lester et al., 2013). Subsequently, the FOCUS Resilience Curriculum (FRC) was created using the core trauma-informed, family-centered components of the FOCUS intervention, but adapted of for delivery as a school-based curriculum. It aimed to promote resilience among students facing adversity through modularized skill-building sessions delivered in small groups or classroom settings. Each classroom module taught skills such as goal setting, problem solving, communication, and emotional regulation. Initially, this intervention was delivered by trained school social work interns to military-connected students (Garcia and et al, 2015). In 2015, the intervention was further adapted to meet the needs of minoritized students living in under-resourced communities, who were attending a large urban school district with a high prevalence of trauma exposure. It featured a community-participatory methodology with youth, parent and education stakeholders (Ijadi-Magshoodi et al., 2017).

An early evaluation of the FOCUS Resilience Curriculum (FRC) indicated that the classroom-based intervention was associated with improved student internal resilience, particularly in the areas of problem solving and empathy (Ijadi-Maghsoodi et al., 2017). Fol-

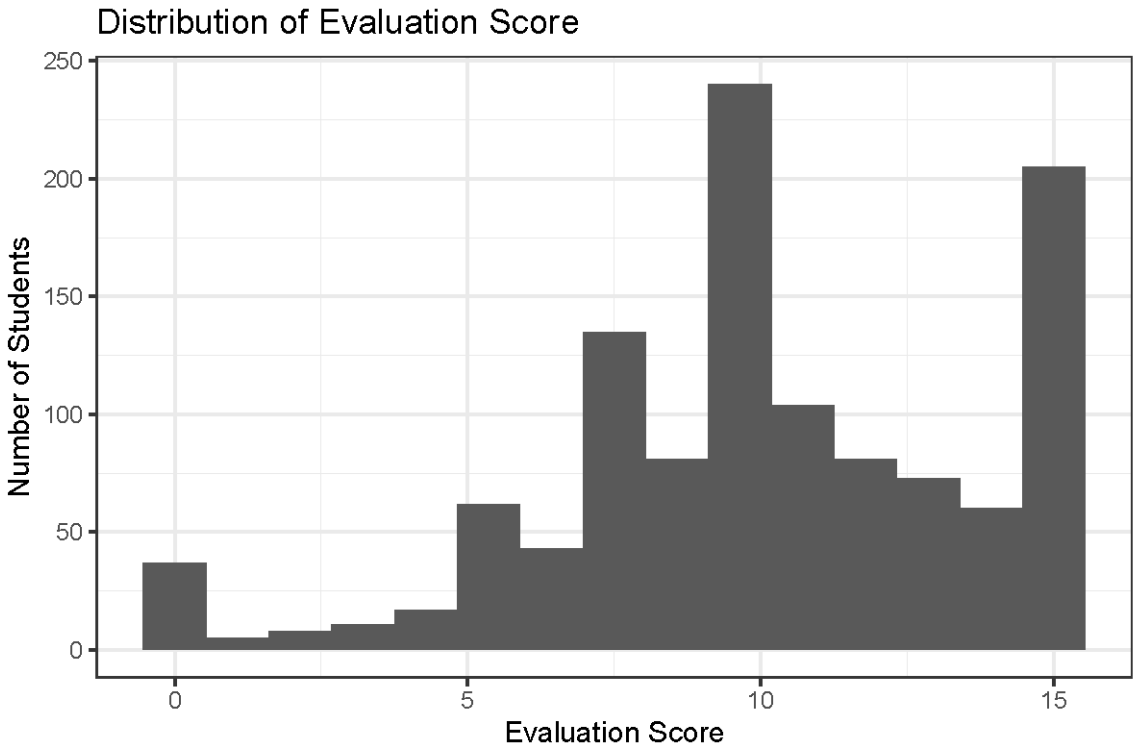
lowing this foundational study, the FRC was implemented on a large scale within the school district. As of February 2022, over 17,000 students in grades 4-12 at over 200 elementary, middle, and high schools had received some form of FRC and implementation is ongoing. Following completion of the curriculum, high school students were electronically administered a short questionnaire which has varied in content over the years but has consistently asked the following set of self-reported learning and satisfaction items:

- I learned ways I can feel less stressed.
- I learned ways to communicate better with others.
- I learned how to set personal goals.
- I learned how to solve problems that come up in my life.
- I would recommend this curriculum to other students.

To each of the above items, students selected a response option ranging from Not At All True (0) to Very Much True (3). Using these responses an Evaluation Score was then calculated by summing the numeric values across all 5 items (range: 0-15). This Evaluation Score, shown in Figure 5.1 demonstrated reasonable variability across students and is the main outcome of interest for the purpose of evaluating intervention effects.

The FRC was delivered to students by facilitator teams, which were comprised of a psychiatric social worker and optional support personnel such as social work interns. Facilitator teams differed greatly in their experience, resources, relationship with students and schools, and comfort administering a classroom-based curriculum. Customization of the intervention occurred dynamically and was encouraged in order to meet the diverse needs of students. Thus, for the currently available data, we consider each group of students who received the FRC from a different facilitator team to have received a different version of the intervention. Current records lack sufficient detail on exact modes of intervention customization that occurred in the classroom to distinguish intervention versions based on the protocol used rather

Figure 5.1: LAUSD Intervention Outcomes.



than the personnel who delivered the intervention. Nevertheless, we expect this information will be recorded and made available to a greater extent going forward.

Broadly, our interest lies in using data that has already been collected to inform the creation of an ongoing analysis plan using the two-step Bayesian analytical framework described herein. This analysis plan could be used to guide future implementation decisions as the FRC program continues to be delivered to students. Furthermore, the District plans to drastically increase the number of students receiving some version of the intervention in the years to come as part of a set of large state- and federally-funded initiatives to support students in recovering from mental health challenges and pandemic-associated learning disruptions.

5.2 Planning Future Analyses Using School-Based Data

5.2.1 Simulation Design

Our sample, used to inform future implementation planning, includes high school students who received the FRC during the 2017-2018 or 2018-2019 school year (prior to the COVID-19 pandemic) and completed the post-intervention questionnaire such that an Evaluation Score could be calculated. The sample was further limited to students who received an intervention from one of 9 different facilitator teams corresponding to the 9 teams that served a minimum of 75 students (Table 5.1). Intervention versions associated with facilitator teams have been labeled 'Version A'- 'Version I.'

Table 5.1: Mean Evaluation Score by intervention version using the retrospective school-based data.

Intervention Version	**Number of Students**	**Mean Evaluation Score**	**Std. Dev.**
Version A	348	10.22	3.53
Version B	129	8.02	4.75
Version C	123	10.07	3.79
Version D	121	11.40	3.29
Version E	95	9.96	3.35
Version F	95	10.99	3.43
Version G	86	10.58	3.13
Version H	90	9.79	3.56
Version I	75	10.93	3.29

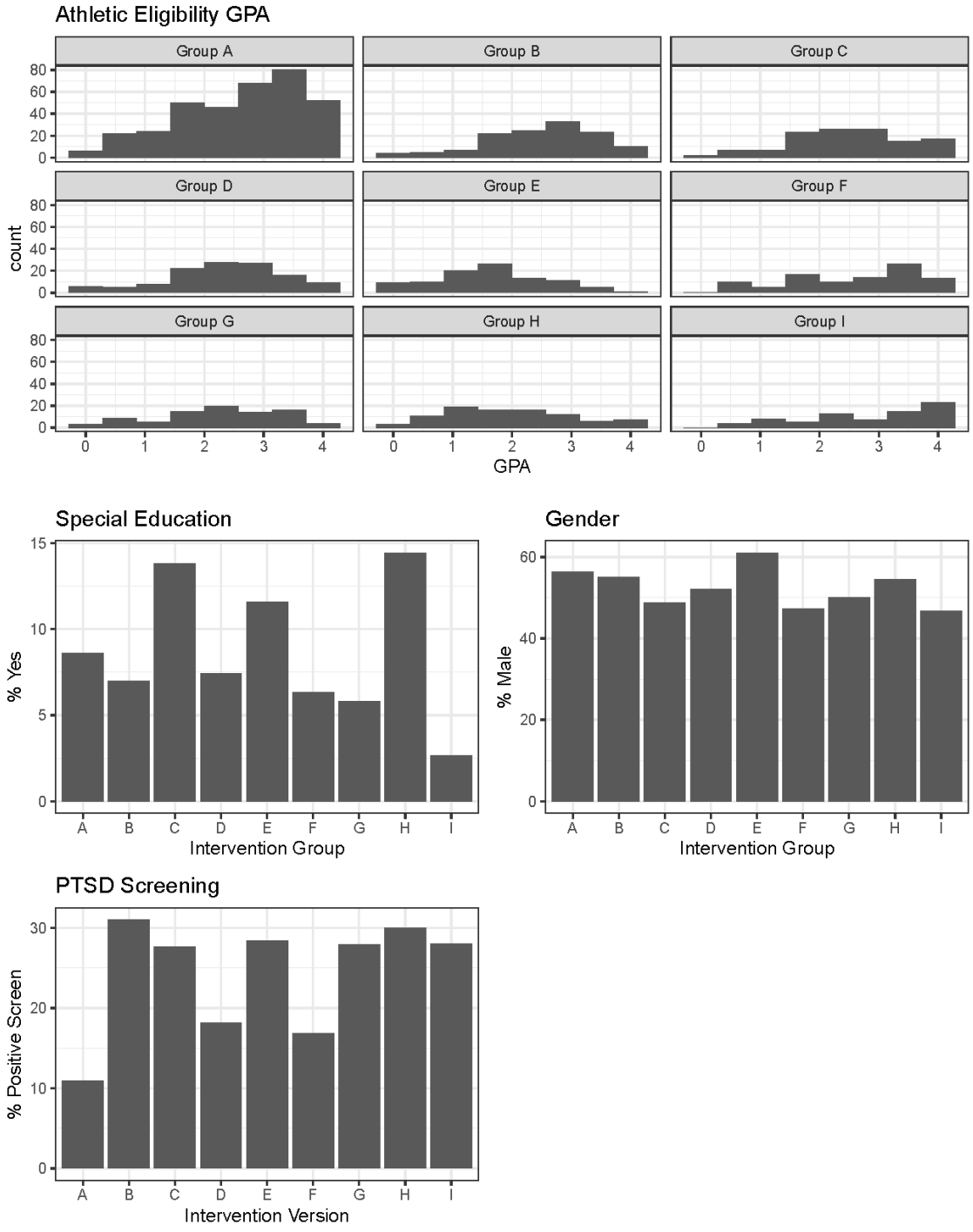
Information on student characteristics was gathered from administrative data provided by the school district as well as routinely administered electronic questionnaires. Student-level characteristics that varied between student groups and which were associated with Evaluation Scores included grade point average (GPA), a Primary Care Post Traumatic Stress Disorder (PC-PTSD) screening result (Cameron and Gusman, 2003), an indicator of whether the student was enrolled in special education, and gender. For the PC-PTSD a verified cut point of ≥ 3 was used to identify a positive screen for PTSD (Vera et al., 2012). In a preliminary analysis, we found several significant associations between student

characteristics and the Evaluation Score. Students with higher GPA tended to report increased Evaluation Scores ($p < 0.001$). Students who screened positive for PTSD tended to report decreased Evaluation Scores ($p < 0.01$) and students in special education programs also tended to report decreased Evaluation Scores ($p = 0.068$). These covariates, shown in Figure 5.2, were chosen as relevant covariates to be balanced in statistical analyses.

An important practical question in putting the Adaptome framework into practice is how often to conduct interim analyses. This parameter can be chosen to optimize statistical performance, and ultimately student outcomes. To answer the question of how frequently to conduct interim analyses we completed a series of simulations and calculated the performance measures described above. We consider frequency to be determined by the number of new participants (students) per intervention group needed before another statistical analysis should take place. Thus, to assess analysis frequency we vary a single simulation input parameter corresponding to the sample size per intervention version per interim analysis with possible options (40, 60, 80, 100, 120, 140, 160). We considered schools' needs to determine other simulation parameters. To allow sufficient speed and flexibility in implementation decisions we use a superiority threshold of 0.8, and an inferiority threshold of 0.2. With school resource constraints in mind, we set a maximum of 4 intervention versions actively enrolling students at any given point in time. For practical purposes, any time an interim analysis results in dropping one or more intervention versions, they are replaced with new intervention versions as long as more adaptations are available. In order to compare the possible values of the Average (intervention) Effect Received for the next 2,000 students receiving the FRC, we set a maximum sample size of 2,000 students for each simulated platform trial.

Simulation inputs corresponding to student characteristics, the distribution of Evaluation Scores, and the relationships between these were derived from the available data. Additionally, relationships between the aforementioned covariates and intervention groups were estimated from the data, as were a set of intervention effects. Table 5.2 gives further detail on how this was achieved. Once these relationships were estimated from the data, the simulation proceeded as follows:

Figure 5.2: Covariate distributions by intervention group in LAUSD.



1. Generate covariates to match the distributions in the available data.
2. Assign intervention versions according to covariate-intervention group relationships.
3. Generate outcomes based on covariates and intervention version received.
4. Perform interim analysis using the data generated for active intervention versions.
5. Drop, add, or switch the preferred intervention version(s) according to the interim analysis results.
6. Repeat until the stopping criteria (maximum total sample size) is reached.

We ran the Bayesian platform trial 100 times for each set of inputs and used the previously described performance measures to compare across different potential sample sizes per intervention version per interim analysis (our metric for interim analysis frequency). We found that for smaller sample sizes, it was sometimes computational infeasible to perform entropy balancing on the generated data. In these instances, the interim analysis was completed without the entropy balancing step, and thus without weights in the Bayesian model. Failing to balance covariate distributions among groups when data is observational is known to produce biased estimates. For this reason, we also consider the proportion of interim analyses that were successfully able to use entropy balancing as a measure for comparison. Although we describe this issue as it arises when implementing entropy balancing, it is an inherent limitation of the data that is likely to cause issues if attempting covariate balancing by any means (matching, propensity scores). In practice, an analyst could address entropy balancing failures by reducing the number of matching constraints, which again poses the risk of increasing bias, or by waiting to perform an interim analysis until more data has been collected.

5.2.2 Simulation Results

Simulated trials encountered no entropy balancing issues when the sample size per intervention version per interim analysis was 100 or greater (Table 5.3). For lower interim

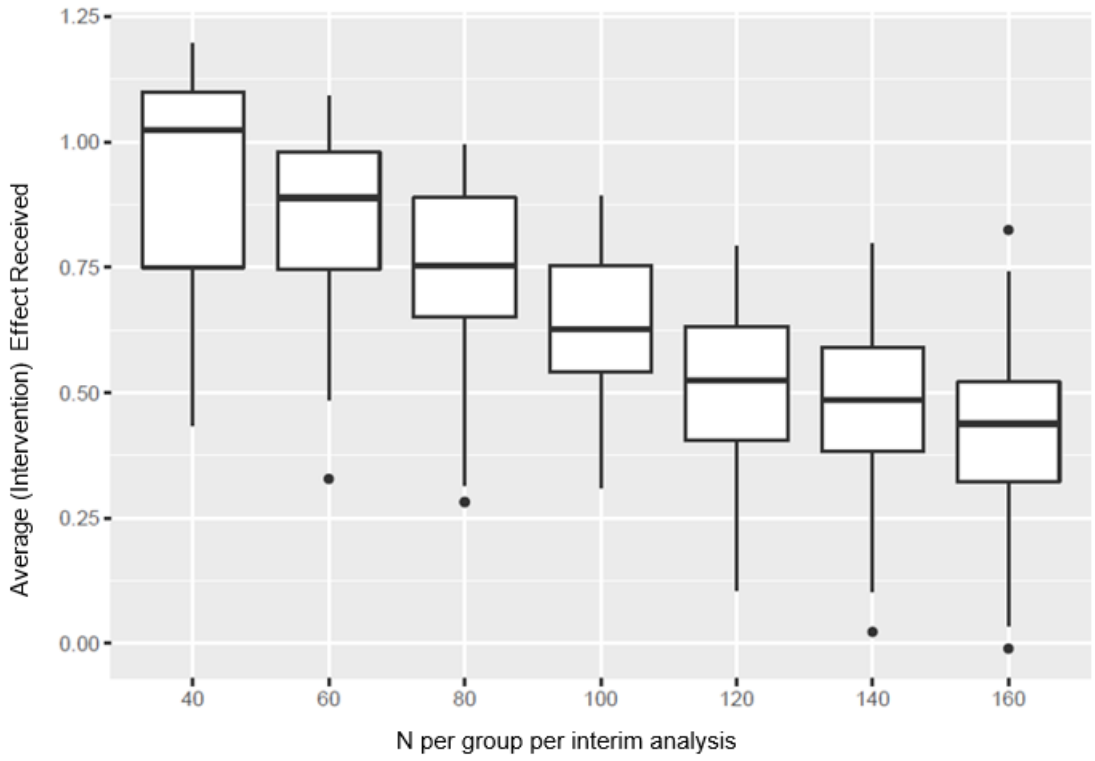
sample sizes, however, as few as 89% of interim analyses successfully used entropy balancing. Positive Action Probabilities increased as a function of interim sample size (range = [0.868, 0.955]). The percent of simulated trials that identify the best intervention version increases with increased interim sample size, reaches a maximum, and then decreases again. The decrease occurs with larger sample sizes because of the set maximum total sample size (2,000 students). With larger interim sample sizes, fewer interim analyses can occur before the maximum is reached and thus there are fewer chances to successfully identify the top intervention version, particularly if the best version is not one of the first versions introduced into the trial.

Estimates of the Average (intervention) Effect Received by students were plotted across simulation scenarios with different sample sizes per intervention version per interim analysis. The Average Effect Received tended to be higher for trials with more frequent interim analyses (Figure 5.3, range = [0.0, 1.2]). This suggests that expected population-level benefit increases when intervention versions are dropped and added more frequently. Based on examination of the three performance measures (Positive Action Probability, Percent Top One, and Average Effect Received), in combination with our understanding of the importance of successful of covariate balancing, we would propose creating an analysis plan that in which interim analyses are conducted after the accrual of 100 students per active intervention version. This analysis plan would require 4-5 interim analyses over the time period during which 2,000 students receive the FRC.

5.3 Practical Steps for Creating an Analysis Plan

To facilitate application of the information presented, we describe a few practical steps investigators and analysts can follow, as well as considerations that should be made throughout the process. Ideally, the steps described below should be taken after an intervention has been successfully developed and proven efficacious through a randomized control trial, as it begins to be translated to real-world settings where effectiveness has yet to be determined. At this pivotal moment, investigators should ensure data collection systems are in

Figure 5.3: Average (intervention) effect received by sample size per intervention version per interim analysis for the application to school-based data.



place such that necessary information will be collected from all intervention participants. Necessary information includes not only intervention outcomes, but also relevant covariates, information about how the intervention is being adapted and which participants receive each adapted version. Collection of information in a timely fashion is also critical in order to conduct interim analyses that can guide implementation of the intervention for the next set of participants. Lastly, investigators should establish a mechanism for routine dissemination of interim results. To meaningfully impact populations targeted by the intervention, the results of interim analyses need to be shared with stakeholders responsible for driving allocation of new potential participants to available intervention versions and, ultimately, the discontinuation of inferior versions.

Once these preliminary considerations have been addressed, an analyst in conjunction with other stakeholders should take the following steps to set up a simulation study. The results of the simulations should lead to creation of a statistical analysis plan with optimal performance.

1. Establish which input parameters will be varied and what values are feasible for each. These parameters may include: inferiority threshold, superiority threshold, or sample size per intervention version per interim analysis.

2. Determine a reasonable plan for the introduction of intervention versions. Will the number of active versions vary randomly according to some distribution? Will new versions be added immediately upon the dropping versions for futility?

3. Determine one or more arbitrary end dates/stopping rules to dictate when to terminate simulated trials and assess performance.

4. Use existing data or expert opinion to determine the following:

   (a) Specific covariates that will need to balanced

   (b) Anticipated covariate distributions

(c) Correlation or another associational measure of the relationship between each covariate and a hypothesized set of intervention versions

(d) Partial correlation or another adjusted associational measure of the relationship between each covariate and the intervention outcome, controlling for other covariates

(e) A series of feasible intervention effects, controlling for relevant covariates

5. Run simulations as described above while varying the input parameters specified in (1). Results will include the set of performance measures described in Table 4.3. (R code available at: https://github.com/tdbufford/Adaptome-Simulations)

6. Make an informed decision that best suits the particular application and relevant priorities of stakeholders.

Conceptually, this process is similar to the way an investigator might vary input parameters used in a traditional power analysis and assess type I error rates in order to plan a clinical trial. We are advocating for the use of similar techniques applied to the ongoing collection of real-world data from intervention implementation. The framework should be in place prior to the initiation of data collection. Accumulation of participant data will trigger the first and all subsequent interim analyses, the results of which will be relayed back to investigators/community partners/stakeholders to enact changes. In theory, this framework can be used in perpetuity, although simulations should be re-run periodically to update performance measure estimates in light of the reality of implementation and make any necessary design changes.

5.4 Limitations

The proposed framework does not take into account differential intervention effects among participants with different characteristics, sometimes referred to as heterogeneity of effects. This would make for a useful extension to further the applicability of the proposed method

in real-world settings. Another limitation of this method is that the performance measures do depend upon the relative effectiveness of the hypothetical new intervention adaptations, the order in which they arise, and the number of different intervention versions that are evaluated at a time. They also depend on the relationships between covariates and number of covariates, which can be difficult to know a priori. Although we cannot predict exactly how this will occur throughout future implementation, historical data can be used to approximate a plausible scenario. If this is not available, we can test multiple scenarios. Regardless, we should still be able to assess the general properties of the proposed analysis plan while keeping potential dependencies in mind.

5.5 Discussion

This paper demonstrates that the proposed two-step Bayesian interim analysis plan can be useful in driving the continuous adaptation process such that beneficial intervention adaptations are seamlessly transitioned into use and subsequent improvements can be introduced, evaluated, and either retained or discarded for the purpose of improving individual outcomes on average over time. This chapter also provides a detailed roadmap for investigators interested in adopting such a framework and understanding its expected performance under conditions relevant to their given setting. It is our hope that these robust methods, which have become increasingly popular in the clinical trials realm, will be implemented to a greater extent among behavioral and other health researchers interested in evaluating dynamically-arising intervention adaptations. This paper serves to make these methods more accessible and demonstrates their immense potential to improve population-level implementation of health interventions without an expensive or logistically infeasible return to the randomized controlled trial setting.

Table 5.2: Additional details for simulations based on school data.

Step 1. Generate Covariates		
Model or procedure for school data	**Estimates from school data**	**Use in simulation**
Mean(Gender)	0.53	Gender $\sim$ Binom($p = 0.53$)
Mean(PTSD)	0.22	PTSD $\sim$ Binom($p = 0.22$)
Mean(SpEdu)	0.09	SpEdu $\sim$ Binom($p = 0.09$)
Mean(GPA)	Mean = 2.5	GPA $\sim$ t($df = 1$) + 2.5
Assess range and shape	Range = [0.0, 4.0]	Truncate to [0.0, 0.4]
Step 2. Assign Intervention Versions		
Model or procedure for school data	**Estimates from school data**	**Use in simulation**
$log\frac{Pr(B)}{Pr(A)} = \beta_{1,1}(Gender)+$ $\beta_{2,1}(PTSD) + \beta_{3,1}(SpEdu)+$ $\beta_{4,1}(GPA)$ $\vdots$ $log\frac{Pr(I)}{Pr(A)} = \beta_{1,8}(Gender)+$ $\beta_{2,8}(PTSD) + \beta_{3,8}(SpEdu)+$ $\beta_{4,8}(GPA)$	$\beta^T =$ $\begin{bmatrix} 0.0 & 1.3 & -0.3 & -0.1 \\ -0.4 & 1.1 & 0.6 & -0.1 \\ -0.3 & 0.5 & -0.3 & -0.3 \\ 0.0 & 0.9 & -0.1 & -0.8 \\ -0.3 & 0.5 & -0.3 & 0.0 \\ -0.3 & 1.1 & -0.8 & -0.4 \\ -0.3 & 1.1 & 0.3 & -0.6 \\ -0.2 & 1.2 & -1.1 & 0.2 \end{bmatrix}$	$Pr(A) = 1/\rho$ $Pr(B) = exp(\beta_{.1}x_{i.} + \epsilon_1)/\rho$ $\vdots$ $Pr(I) = exp(\beta_{.8}x_{i.} + \epsilon_8)/\rho$ $\rho = 1 + exp(\beta_{.1}x_{i.} + \epsilon_1)+$ $... + exp(\beta_{.8}x_{i.} + \epsilon_8)$ $\epsilon_1, ..., \epsilon_8 \sim N(0, 4)$ T $\sim Multinom(Pr(A), ..., Pr(I))$
Step 3. Generate Outcomes		
Model or procedure for school data	**Estimates from school data**	**Use in simulation**
$EvalScore = \alpha_2 t_{i2} + \alpha_3 t_{i3} + ...+$ $\alpha_9 t_{i9} + \psi_1 Gender + \psi_2 PTSD$ $+\psi_3 SpEdu + \psi_4 GPA$	$\boldsymbol{\alpha} = [0, -1.9, 0.07, 1.4, 0.4,$ $0.85, 0.72, 0.11, 0.77]$ $\boldsymbol{\psi} = [0.2, -0.9, 0.1, 0.6]$	$EvalScore = \boldsymbol{X\psi} + \boldsymbol{T\alpha} + \boldsymbol{\eta}$ $\eta_i \sim N(0.3, 3.7)$ Truncate to $[0, 15]$ increase modes at 0, 5, & 10 using additional random variables

Table 5.3: Performance measures by sample size (per intervention version per interim analysis) for the application to school-based data .

N per Intervention Version per Interim Analysis	**Average No. of Analyses**	**Successful Use of EB (%)**	**Percent Top One**	**Positive Action Probability**
40	6.5	89.2	60	0.868
60	6.2	96.9	88	0.913
80	5.2	99.6	88	0.933
100	4.6	100.0	88	0.947
120	3.9	100.0	74	0.956
140	3.1	100.0	71	0.953
160	3.0	100.0	68	0.955

CHAPTER 6

Conclusion

6.1 Correlated Outcomes

In these chapters we have explored novel statistical methods that are well suited for use in analyzing outcomes for Behavioral Health Interventions. The permutation methods described in chapters 2 and 3 deal with the question of determining overall intervention efficacy when analyzing the result of a randomized trial when many primary outcomes have been measured and these outcomes, such as depression, anxiety, PTSD symptoms, and family functioning, are all correlated. The exact correlation structure for the outcome measures is often unknown and estimates may be unreliable. In addition, we require use of complex modeling techniques to assess each individual outcome, and the possibility of different models for different outcomes.

The two versions of permutation tests both aim to control experiment-wise Type I error rate in this situation, while allowing for the needed modeling flexibility and avoiding distributional assumptions. The first permutation test finds a cut point for the number of statistically significant hypothesis tests needed among the M outcomes in order to have enough evidence of an overall intervention effect. The second permutation test takes into account the magnitude of the p-values, rather than simply dichotomizing the p-values into categories of significant or not significant, and calculates a single overall p-value for the intervention. Both tests are successful at controlling Type I error around 0.05, and both are most powerful when the majority of the outcomes have a true underlying difference between intervention and control groups, even if the effect size is relatively small.

The benefits of the proposed permutation methods are that no distributional assumptions

were made for the number of significant findings or the level of correlation between the test statistics. The methods are non-parametric, utilizing an empirical distribution estimated through repeated sampling. We posit that the permutation methods are both valid for data with any underlying correlation structure. In addition, the proposed permutation methods can easily be extended to any type of hypothesis test. They can accommodate not only normally distributed outcomes, but also binary outcomes, count data, survival data, or a mixture of these. This may have use in early stages of a clinical trial when measuring a variety of outcomes related to the new drug/biologic in question, and deciding whether to continue the drug development based on these correlated outcomes. However with any statistical test, we recommend using one-sided p-values whenever possible to ensure the directionality of the estimated outcomes are in line the expected directionality of the intervention effect.

The permutation methods also differs from conventional multiple testing methods because they focus on the question of whether there is an underlying intervention effect that causes the observed differences between intervention and control groups for various outcome measures, rather than making inference on individual outcomes. In behavioral health we may have a relatively small sample size, and thus p-values that are not far beyond the traditional 0.05 threshold when an intervention effect is present, but we may have many measured outcomes which can collectively give evidence of an intervention effect. These methods are most useful in detecting an intervention effect when there are many outcomes which display differences between intervention and control groups, as is the case with a well-designed behavioral health intervention.

A potential limitation to both permutation methods are the computational resources required for implementation, which depends on the number of outcomes and the complexity of the outcome models. In cases where we are using a simple statistical test, such as a t-test, the test statistics can be recalculated for 5,000 permutations in under 30 seconds on most machines. However when using more complex models, such as in the FOCUS-EC example, the computation time is much longer. When performing the outcome analysis a single time, or for tweaking and re-fitting the outcome models, the computation time is

usually not an issue, even for relatively complex models, unless the data size is very large. However when multiplying that by the number of permutations, which needs to be at least 1,000 in order to calculate a p-value with 3 decimal points, then the computation time can easily grow to a considerable length. For example, in the FOCUS-EC application, each individual linear mixed effects model took about 30 milliseconds to fit in R. Multiplied by 24 outcomes, the FOCUS EC data models took about 720 milliseconds to produce all estimates and associated p-values. Multiplied by 5,000 permutations, however, and the computation time is approximately 1 hour.

Parallelization of the computation, using the R package "doParallel" (cite), is a relatively simple way to minimize the computation time. Further effort to increase the computational efficiency may not be worthwhile since the permutation test only needs to be run once for a given analysis. The fitting, revising, and refitting of models is done before implementing the permutation test, and the permutations only need to be run using the final statistical models used in the analysis of the non-permuted data. This also helps mitigate the computation time burden. Implementation in this manner does require some coding know-how, but should be within the capability of a typical statistical programmer.

Future work may include streamlining use of the permutation methods through development of R packages that can be easily installed and used by statistical programmers. Development of such code will present a challenge in incorporating enough flexibility in choice of statistical model for each outcome, including regression models, logistic models, longitudinal models, and various others.

An underlying assumption that we have made in the course of these simulation studies is that in the presence of a treatment effect, the effect size would be constant across measures. This may be an acceptable assumption if we believe that the various measures are all attempting to quantify the same underlying quality of family functioning or emotional wellbeing. We can imagine this as a latent variable. This assumption is important because differing effect sizes for different measures, or different groups of measures, will affect our power estimation.

Future work may include further simulations that vary the effect sizes of each outcome for a single alternative hypothesis. For example, we may want to consider that one cluster of outcomes has effect size 0.3 while another cluster of outcomes has effect size 0.2. Simulations can be used to estimate the power of a given study if we can estimate a reasonable effect size for each outcome and have an idea of which outcomes may be clustered together, ie. strongly correlated. We may also want to explore further the performance of these permutation methods when the data have non-normal distributions. Simulating this type of data is more challenging than simulating normal data, however it may be useful to assess estimated power and Type I error rates when the data are, for example, bimodal. We may also wish to compare these power estimates for the permutation test to power estimates for Hotelling's T^2 test when the distributional assumption for the latter is violated. We may find additional use cases for the permutation tests. Finally we could compare the permutation test in chapter 1 with the permutation test in chapter 2 to see if conclusions of overall intervention efficacy are consistent and again compare the relative power of each method.

6.2 Adaptome

In addition to analyzing Behavioral Health Interventions in the controlled trial stage, we also have developed novel methods to assess comparative effectiveness of intervention adaptations that occur afterwards during widespread intervention implementation. Adaptations to intervention protocol naturally arise to meet the needs of various populations and to overcome practical obstacles. In chapters 4 and 5 we have laid a path forward for strategically evaluating interventions being implemented at the population level in a way that maximizes benefit though identification of beneficial adaptations. This new statistical framework involves regular analysis of intervention adaptations over time as new data is collected. These are called interim analyses. We combine methods from platform trials with methods for covariate balancing to make pairwise comparisons between intervention adaptations and estimate their relative effectiveness. We then make decisions to drop an adaptation for futility or designate an adaptation as the current best version. In doing this, more members of the

population receive intervention versions with greater effectiveness in the long run.

This is particularly useful in today's world where large quantities of real-world data are continuously collected over time. We find that by comparing intervention outcomes while balancing covariates among intervention groups and dropping inferior intervention adaptations we can improve overall outcomes. We also find that in implementing the adaptome framework, when an action is taken such as switching which adaptation is the standard for comparison or dropping an adaptation, the probability of making a good decision is very high. Errors typically occur less than 5% of the time using simulated data where all possible confounding factors are measured and balanced using entropy balancing. Stricter superiority and futility thresholds can further reduce error rates, while looser thresholds will allow for swifter action and more flexibility in implementation.

While this method may not account for every nuance in the data, we have shown that implementing this framework for statistical analysis does improve upon population average outcomes that are observed when there is not periodic statistical analysis or strategic decisions made during intervention implementation. It is, unfortunately, too often the case that no form of data analysis is used to guide widespread implementation following a successful randomized controlled trial. In contrast to the randomized controlled trial, we have shown that measurement and balancing of covariates that are related to receipt of the intervention or intervention outcome is critical in being able conduct meaningful data analysis using real-world data. Without accounting for these covariates, our implementation decisions are often misguided. Another important aspect of the adaptome framework is the use of Bayesian statistical techniques. Bayesian analysis provides us a natural way to use knowledge gained from previously collected data at each interim analysis without explicitly re-using data.

A way to further enhance population level benefit may be to consider outcome-adaptive intervention allocation, which increases the ratio of participants who are allocated to intervention versions that produced better outcomes according to the latest interim analysis (Sabo et al., 2013; Sim, 2019). This way more people would receive the best intervention adaptation of those currently available. In doing so, however, we would not want to be

completely deterministic in our approach to intervention group assignment because we still want to allow for further data to be collected for other adaptations and for new adaptations to arise. Future work can include a simulation study that uses a probability for adherence to determine the level of outcome-adaptive intervention allocation. We could compare the Average Effect Received by participants under different adherence probabilities.

Further exploration of the adaptome framework may include simulations that vary hyper-parameters that we held fixed. This could be varying numbers of active intervention versions at a given interim analysis, or using an indefinite number of potential intervention versions where intervention effect sizes as well as outcomes and relationships between these and the covariates all arise from probability distributions.

The framework can also be extended to account for additional structural elements in the data, such as clustering, through typical statistical methods used in Bayesian modeling. In our example, the FRC was delivered at the classroom level, so we may choose to add a random-effect estimate for modeling variability across classrooms if the classroom information is available.

The adaptome framework can be applied not only to behavioral health interventions but also more broadly to health services, in particular those in which extensive electronic health record (EHR) data are available. This may include cancer treatments, diabetes treatments, or other settings in which observational data are readily available and multiple treatment options are currently utilized in clinical practice. If we applied the adaptome framework to EHR data, we may want to similarly account for clustering by clinic or hospital. Future work may also include further investigation into the feasibility of applying the adaptome framework to EHR data.

Additionally, the proposed framework does not take into account differential intervention effects among participants with different characteristics, sometimes referred to as heterogeneity of effects. Further extension of the adaptome framework to include intervention effect heterogeneity could also be useful for the application of this method in real-world settings.

In the case of behavioral health interventions, a crucial aspect of successful implemen-

tation of the adaptome framework is an working flow of communication between those who deliver the intervention and those who assess the outcomes. This can take the form of community-academic partnerships. These partnerships can support population- based data collection, assessment, and quick reporting, including providing online data platforms to community partners. Establishing these partnerships can be a first step in addressing the issue of data analysis during widespread implementation in communities.

An important next step in the adaptome project is implementing the new statistical framework within LAUSD. This will be done as data from the FRC is analyzed at interim analyses over the coming years. We may also want to assess whether there are additional covariates that can be measured and balanced that may affect intervention outcomes.

Bibliography

Aickin, M. and Gensler, H. (1996). Adjusting for multiple testing when reporting research results: the bonferroni vs holm methods. *American Journal of Public Health*, 86(5):726–728.

Anatchkova, M., Donelson, S., Skalicky, A., and et al (2018). Exploring the implementation of patient-reported outcome measures in cancer care: need for more real-world evidence results in the peer reviewed literature. *J Patient Rep Outcomes*, 2(64).

Anderson, M. J. (2001). Permutation tests for univariate or multivariate analysis of variance and regression. *Canadian Journal of Fisheries and Aquatic Sciences*, 58(3):626–639.

Angus, D., Alexander, B., Berry, S., and et al (2019). Adaptive platform trials: definition, design, conduct and reporting considerations. *Nature Reviews Drug Discovery*.

Austin, P. C. (2011). "an introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate behavioral research*, 46.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B (Methodological)*, 57(1):289–300.

Berry, K. J. and Mielke, P. W. (1985). Computation of exact and approximate probability values for a matched pairs permutation test. *Communications in Statistics - Simulation and Computation*, 14(1):229–248.

Bufford, T., Aralis, H., Crespi, C. M., Ijadi-Maghsoodi, R., Kataoka, S., and Lavelle, C. (2022a). Assessing intervention adaptations using real-world evidence and ongoing analysis. *Submitted to the Journal of Educational and Behavioral Statistics*.

Bufford, T., Aralis, H., Kataoka, S., Lee, S.-J., Lavelle, C., and Lester, P. (2022b). Creating a statistical analysis plan to continually evaluate intervention adaptations that arise in real-world implementation. *Submitted to Prevention Science*.

Cameron, R. P. and Gusman, D. (2003). The primare care ptsd screen (pc-ptsd): Development and operating characteristics. *Primary Care Psychiatry*, 9(1):9–14.

Carroll, C., Patterson, M., Wood, S., and et al (2007). A conceptual framework for implementation fidelity. *Implementation Sci*, 2(40).

Chambers, D. A. and Norton, W. E. (2016). The adaptome: Advancing the science of intervention adaptation. *American Journal of Preventive Medicine*, 51(4, Supplement 2):S124–S131. Realizing Population-Level Improvement for All Children's Cognitive, Affective, and Behavioral Health.

Cheverud, J. M. (2001). A simple correction for multiple comparisons in interval mapping genome scans. *Heredity*, 87:52–58.

Corrigan-Curay, J., Sacks, L., and Woodcock, J. (2018). Real-World Evidence and Real-World Data for Evaluating Drug Safety and Effectiveness. *JAMA*, 320(9):867–868.

Davidson, B. A. and et al (2011). Permutation criteria to evaluate multiple clinical endpoints in a proof-of-concept study: lessons from pre-relax-ahf. *Clin Res Cardiol*, 100:745–753.

Drake, C. (1993). Effects of misspecification of the propensity score on estimators of treatment effects. *Biometrics*, 49:1231–1236.

Feng, P., Zhou, X.-H., Zou, Q.-M., Fan, M.-Y., and Li, X.-S. (2012). Generalized propensity score for estimating the average treatment effect of multiple treatments. *Statistics in Medicine*, 31(7):681–697.

Folich, M. (2007). Propensity score matching without conditional independence assumption with an application to the gender wage gap in the united kingdom. *The Econometrics Journal*, 10:359–407.

Franklin, J. M., Pawar, A., Martin, D., Glynn, R. J., Levenson, M., Temple, R., and Schneeweiss, S. (2020). Nonrandomized real-world evidence to support regulatory de-

cision making: Process for a randomized trial replication project. *Clinical Pharmacology & Therapeutics*, 107(4):817–826.

Gao, X., Starmer, J., and Martin, E. R. (2008). A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms. *Genetic Epidemiology*, 32:361–369.

Hainmueller, J. (2012). Entropy balacing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis*, 20:25–46.

Hajal, N., Aralis, H., Kiff, C., Wasserman, M., Paley, B., Milburn, N., Mogil, C., and Lester, P. (2020). Parental wartime deployment and socioemotional adjustment in early childhood: The critical role of military parents' perceived threat during deployment: Parental deployment and early childhood adjustment. *Journal of Traumatic Stress*, 33.

Harwood, J. M., Weiss, R. E., and Comulada, W. S. (2017). Beyond the primary endpoint paradigm: A test of intervention effect in hiv behavioral intervention trials with numerous correlated outcomes. *Prev Sci*, 18:526–533.

Hothorn, T., Hornik, K., van de Wie, M. A., and Zeileis, A. (2008). Implementing a class of permutation tests: The coin package. *Journal of Statistical Software*, 28(8):1–23.

Hummel, J., Wang, S., and Kirkpatrick, J. (2015). Using simulation to optimize adaptive trial designs: applications in learning and confirmatory phase trials. *Clinical Investigation*, 5:401–413.

Hyndman, R. J. and Fan, Y. (1996). Sample quantiles in statistical packages. *American Statistician*, 50.

Ibrahim, S. and Sidani, S. (2015). Fidelity of intervention implementation: A review of instruments. *Health*, 7:1687–1695.

Jennings, P. A. and et al (2013). Improving classroom learning environments by cultivating awareness and resilience in education (care): Results of a randomized controlled trial. *School Psychology Quarterly*, 28(4):374.

Kent, M. and et al (2011). A resilience-oriented treatment for posttraumatic stress disorder: Results of a preliminary randomized clinical trial. *Journal of Traumatic Stress*, 24(5):591–595.

Kiser, L. J. and et al (2015). Strengthening family coping resources (sfcr): Practice-based evidence for a promising trauma intervention. *Couple and Family Psychology: Research and Practice*, 4(1):49.

Lee, B. K., Lessler, J., and Stuart, E. A. (2010). Improving propensity score weighting using machine learning. *Stat Med*, 29, 3:337–346.

Li, J. and Ji, L. (2005). Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity*, 95:221–227.

Lin, J. and Bunn, V. (2017). Comparison of multi-arm multi-stage design and adaptive randomization in platform clinical trials. *Contemporary Clinical Trials*, 54:48–59.

Luby, J., Lenze, S., and Tillman, R. (2012). A novel early intervention for preschool depression: Findings from a pilot randomized controlled trial. *Journal of Child Psychology and Psychiatry*, 53(3):313–322.

Ludbrook, J. and Dudley, H. (1998). Why permutation tests are superior to t and f tests in biomedical research. *The American Statistician*, 52(2):127–132.

Madariaga, A., Kasherman, L., Karakasis, K., Degendorfer, P., Heesters, A. M., Xu, W., Husain, S., and Oza, A. M. (2021). Optimizing clinical research procedures in public health emergencies. *Medicinal Research Reviews*, 41(2):725–738.

McCaffrey, D. F., Griffin, B. A., Almirall, D., Slaughter, M. E., Ramchand, R., and Burgette,

L. F. (2013). A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Statistics in Medicine*, 32(19):3388–3414.

Mogil, C., Hajal, N., Garcia, E., Kiff, C., Paley, B., Milburn, N., and Lester, P. (2015). Focus for early childhood: A virtual home visiting program for military families with young children. *Contemporary Family Therapy*, 37(3):199–208.

Mogil, C., Paley, B., Doud, T. D., Havens, L., Moore-Tyson, J., Beardslee, W. R., and Lester, P. (2010). Families overcoming under stress (focus) for early childhood: Building resilience for young children in high stress families. *Zero to Three*, pages 10–16.

Müller, H.-H. and Schäfer, H. (2001). Adaptive group sequential designs for clinical trials: Combining the advantages of adaptive and of classical group sequential approaches. *Biometrics*, 57(3):886–891.

Nyholt, D. R. (2004). A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. *American Journal of Human Genetics*, 74:765–769.

Oden, A. and Wedel, H. (1975). Arguments for fisher's permutation test. *Ann. Statist.*, 3(2):518–520.

O'Brien, P. C. and Fleming, T. R. (1979). A multiple testing procedure for clinical trials. *Biometrics*, 35:549–556.

Pesarin, F. and Salmaso, L. (2012). A review and some new results on permutation testing for multivariate problems. *Stat Comp*, 22.

Plummer, M. (2003). Jags: A program for analysis of bayesian graphical models using gibbs sampling.

Pocock, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika*, 64:191–199.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Sabo, R. T., Roberts, C., Toor, A. A., and McCarty, J. M. (2013). An outcome-adaptive allocation method for clinical trials with dual binary objectives. *Statistics in Biopharmaceutical Research*, 5(1):67–78.

Saville, B. and Berry, S. (2016). Efficiencies of platform clinical trials: A vision of the future. *Clin Trials*.

Saville, B. R., Connor, J. T., Ayers, G. D., and Alvarez, J. (2014). The utility of bayesian predictive probabilities for interim monitoring of clinical trials. *Clinical Trials*, 11(4):485–493. PMID: 24872363.

Schneeweiss, S. and Patorno, E. (2021). Conducting Real-world Evidence Studies on the Clinical Outcomes of Diabetes Treatments. *Endocrine Reviews*, 42(5):658–690.

Sim, J. (2019). Outcome-adaptive randomization in clinical trials: issues of participant welfare and autonomy. *Theor Med Bioeth*, 40:83–101.

Solomon, R. and et al (2014). Play project home consultation intervention program for young children with autism spectrum disorders: a randomized controlled trial. *Journal of Developmental and Behavioral Pediatrics*, 35(8):475.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical science : a review journal of the Institute of Mathematical Statistics*, 25, 1:1–21.

Thorlund, K., Golchi, S., Haggstrom, J., and Mills, E. (2019). Highly efficient clinical trials simulator (hect): Software application for planning and simulating platform adaptive trials. *Gates Open Research*, 3.

Vera, M., Juarbe, D., Hernandez, N., Oben, A., Perez-Pedrogo, C., and Chaplin, W. (2012). Probable posttraumatic stress disorder and psychiatric co-morbidity among latino primary care patients in puerto rico. *Journal of Depression & Anxiety*, 1(5):124.

Wang, S. V., Pinheiro, S., Hua, W., Arlett, P., Uyama, Y., Berlin, J. A., Bartels, D. B., Kahler, K. H., Bessette, L. G., and Schneeweiss, S. (2021). Start-rwe: structured template for planning and reporting on the implementation of real world evidence studies. *BMJ*, 372.

Wasserstein, R. L. and Lazar, N. A. (2020). *ASA Statement on Statistical Significance and p-Values*, pages 1–10. Springer International Publishing, Cham.

Whittingham, K. and et al (2014). Interventions to reduce behavioral problems in children with cerebral palsy: an rct. *Pediatrics*, 133(5):e1249–e1257.

Wise, J., Möller, A., Christie, D., Kalra, D., Brodsky, E., Georgieva, E., Jones, G., Smith, I., Greiffenberg, L., McCarthy, M., Arend, M., Luttringer, O., Kloss, S., and Arlington, S. (2018). The positive impacts of real-world data on the challenges facing the evolution of biopharma. *Drug Discovery Today*, 23(4):788–801.