Novel Statistical Methods for Missing Data and Multiplicity in Alzheimer's Research

Robert N Montgomery

B.A. Mathematics, University of Kansas, 2014

M.S. Biostatistics, University of Kansas Medical Center 2016

Submitted to the graduate degree program in Biostatistics and the Graduate Faculty of the University of Kansas in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Committee members

Dr. Jonathan Mahnken, Chair

Dr. Byron Gajewski

Dr. Jo Wick

Dr. Devin Koeslter

Dr. Heather Gibbs

Date defended: November 19, 2019

The Dissertation Committee for Robert N Montgomery certifies
that this is the approved version of the following dissertation :

Novel Statistical Methods for Missing Data and Multiplicity in Alzheimer's Research

Dr. Jonathan Mahnken, Chair

Date approved: November 19, 2019

Abstract

The application of statistical procedures to real data sets seldom proceeds as seamlessly as a textbook problem where all assumptions are verified, and sample sizes are adequate. Common issues include lack of adherence to the statistical analysis plan, missing data and in early stage research, small sample sizes and a large number of variables of interest, i.e. multiplicity considerations. We present novel statistical methodologies that have been developed for use in these adverse scenarios with applications to research into Alzheimer's Disease. Specifically, we have developed an approach for the analysis of paired categorical data when the pairing has been lost, in the context of a study examining the effectiveness of a type of therapy on perceptions of Alzheimer's. We used a weighted bootstrap approach to compare the euclidean distance between the pre and post centers of mass the pre and post therapy groups and despite the loss of the pairing, were able to make conclusions about the research hypothesis.

In addition, we developed a new global hypothesis test, the Prediction Test, which is intended for use in early stage research when the sample size is small and the number of endpoints of interest is large. We utilize researcher's predictions about the direction different endpoints will move, e.g. increase/decrease, and weight these predictions based on the sample correlation matrix. Using this test, we are able to come to a go/no-go decision concerning the feasibility of continuing to study the current research hypothesis, a common concern in early and exploratory studies. The prediction test had good power properties even for very small sample sizes and a large number of variables of interest, a situation in which most tests fail, while also controlling the Type I error rate. We demonstrate the methodology with a data set consisting of Arterial Spin Labeling (ASL) measures on older adults before and after a 12-week exercise regimen. The research hypothesis for this study was that the exercise intervention would alter the structural/functional aspects of the brain, specifically that ASL would increase in the different regions of the brain.

We then provide extensions to the predictions that can be made in the Prediction Test and compare the method to a Linear Mixed Model and a set of t-test on a data set consisting of Dif-

fusion Tensor Imaging (DTI) measures on pre and post kidney transplant patients. The research hypothesis of this study is that kidney transplantation will lead to a normalization of DTI measures, which are emerging bio markers for cognition and Alzheimer's Disease. We also discuss power calculations and conduct a simulation comparison between a set of t-tests and the prediction test.

Acknowledgements

As I neared the end of my undergrad with a Bachelors in Math, I realized that after four years in college I had developed almost no marketable skills and was unsure of what I would do after college. Luckily, I found out about biostatistics at a career fair my senior year and quickly became very interested in applied statistics. I feel incredibly fortunate to have stumbled into a career that I really enjoy, and I'd like to thank some of the people who have helped me along the way.

First, I'd like to thank Dr. Mayo for admitting me to the program, for his support throughout my time here and for giving the opportunity to work within the department after finishing my masters, the experience I've been able to gain has been invaluable.

I'd like to thank Dr. Mahnken, my GTA advisor and Chair of my dissertation committee. It has been a great benefit for me to work alongside you for the past five years; I've learned a lot. You have been a great mentor to me and have provided an excellent example of how to be an effective statistician. I'm incredibly thankful for your patience and guidance and for giving me the opportunity to practice working more independently as I gained experience.

I'd like to thank all my committee members for their support over the years. Dr. Wick, thank you for your guidance, you have always been willing to provide advice and encouragement on any question whether statistical or professional. The culture of the Department is overwhelmingly positive and supportive of students and I think you play a very large role in that. Dr. Gajewski, I think after every talk I have given at the department you have come by to discuss it with me and provide feedback and ideas for other directions, that kind of interaction with someone who actually knows what they're doing has been very encouraging and thought provoking. Dr. Koeslter, most of all I need to thank you for your encouragement when discussing my final project for DOE, the "grant proposal", at the time I was still planning to stop with a masters but your comments about what you thought of my statistical abilities made me think for the first time that maybe I should continue on and get PhD, something I'm very grateful for doing. Dr. Gibbs I want to thank you for taking the time to be on my committee, your questions at my proposal about not only the statistical

methods, but essentially asking what the point of the different methods is, have helped to make this dissertation's focus clearer.

I want to thank my family, my parents and sister for their support and the excellent examples they set throughout my life. Most importantly I want to thank my wife Emily for her support over the last five years, including when you were paying all the bills and I was just rolling into class at 10 am. You have always believed in me and encouraged me to finish my degree and have done an incredible job taking care of our kids while I have been working. I'm looking forward to not working on my dissertation at night anymore so we can spend more time together. To our kids, Rosie and Walter, you are both such a joy to come home to and although can't believe I'm still in school with two kids you have given me even more reason to work hard and are such a blessing.

I want to thank my co-authors on these papers as well. My co-authors for the first paper, Evaluatin paired categorical data when the pairing is lost are Amber Watts, Nicole Burns, Eric Vidoni and Jonathan Mahnekn. For the second paper my co-author is my advisor Jonathan Mahnken. For the third paper my co-authors are Aditi Gupta, Rebecca Lepping and Jonathan Mahnken, thank you all for your help and suggestions.

I'd also like to thank the various organizations that funded this research. The work for Chapter 1, Evaluating Paired Categorical Data when the Pairing is Lost, was funded in part by a grant from the U.S. National Institute of Aging (P30 AG035982). The methodological work on developing the Prediction test was supported in part by a National Institutes of Health Clinical and Translational Science award grant (ULI TR002366) awarded to the University of Kansas Medical Center, by the Department of Biostatistics at the University of Kansas Medical Center, and by a National Institutes of Health grant (P30AG035982) to the University of Kansas Alzheimer's Disease Center. I'd also like to thank Dr. Eric Vidoni for allowing use of the ASL data set as an example of the prediction test, the collection of which was supported in part by the Foundation for Physical Therapy through the Magistro Family Foundation Research Grant, a CTSA grant from NCRR and NCATS awarded to the University of Kansas Medical Center for Frontiers: The Heartland Institute for Clinical and Translational Research UL1 TR000001 (formerly UL1 RR033179). The collection of the DTI

measures in chapter 4 was supported in part by grant K23-AG055666 from the National Institutes of Health.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Alzheimer's Disease (AD) is one of the most common diseases in the United States, with estimates of 5.7 million cases, and healthcare costs in the hundreds of millions of dollars with projections showing both the number of cases and the overall costs will increase. It is the fifth leading cause of death for adults over 65 (Alzheimer's Association, 2018) but there is currently no treatment that can either cure or stop the disease progression and little evidence to suggest that the few available treatments are effective at slowing disease progression (Townsend, 2011). In fact, even in treatments that have been deemed effective, the clinical relevance of these treatments has been called into question (Winslow et al., 2011). In short Alzheimer's Disease is a major public health crisis that given the ageing population of the United States and the lack of effective treatment will most likely get worse before it gets better.

With Alzheimer's disease the patient is not the only one who experiences adverse health conditions, the primary care giver of Alzheimer's patients also experience significant health burdens (Dauphinot et al., 2015) and some studies have shown that a large proportion, 49%, of the general population is equally afraid of developing Alzheimer's themselves as they are of becoming a caretaker (Anderson et al., 2009). The statistical methodologies we have developed were used to address these two concerns with AD; the need for methods to cope with the disease and the need to understand the disease better in order to develop more effective treatments.

In Chapter 2 we discuss the methodology developed to deal with problems encountered during a study aimed at evaluating the effectiveness of a type of therapy called Self Revelatory Performance (SRP) (Emunah, 2015). SRP is a type of theater performance in which audience members participate in what is essentially a group therapy session. The impetus for the study came from Arts and AGEing KC, a theater company that encourages healthy ageing through the arts. Arts and Ageing KC has conducted some informal surveys of participants in their programs, which consist

primarily of elderly adults, that led them to believe Alzheimer's disease was the largest source of concern. In an attempt to meet their community's need they reached out to the Alzheimer's Disease Center (ADC) at the University of Kansas Medical Center (KUMC) about support for an experiment they wanted to conduct using SRP (Burns et al., 2018). The result was a simple observational study, subjects who came to the performance would be asked to complete a pre and post survey that measured their attitudes and feelings toward AD. The research hypothesis was that the SRP would lead to an improvement in perception of Alzheimer's disease. In this observational study, over 80% of the participants knew someone with Alzheimer's and over 20% were primary caregivers for someone with AD.

The study design called for pre-post data comparisons on the observational sample. It was also determined that the KU ADC could not help with the collection of data but would provide support for the development of the survey and the analysis of the data. Unfortunately, due to a miscommunication between KU ADC and Arts and AGEing KC, during data collection the pairing was lost, i.e. the pre and post surveys were placed into two piles with no identifier available to connect a single participant's responses. We were presented with collected data that was intended to be paired but was not. The methodology we developed focuses on one specific question from the survey, a question asking participants to use an Affect grid, essentially a two-dimensional Likert item. We observed 93 completed pre Affect grids and 87 completed post Affect grids and based on the study design we expected that there were dependencies between these two groups, i.e. we could not consider them independent. We developed a modified bootstrap approach to determine how extreme the differences were between the centers of mass of the two groups under the null hypothesis that the SRP was ineffective. Despite the lost data, and the need to throw out the original statistical analysis plan we were able to come to a conclusion concerning the research hypothesis.

In Chapter 3 we present a novel hypothesis test called the Prediction Test. This test is intended for use in early stage research, specifically when the number of measures collected relative to the sample size is large. In early stage research the goal is often to determine whether or not a given researcher hypothesis is worth pursuing further, that is, given preliminary findings is the hypothesis

being borne out to such a degree that a larger more targeted study is warranted. We make use of researcher's predictions about the result of each measure as part of our test statistic. For instance, if a researcher collected measures such as BMI, VO2 max, etc. for participants of a weight loss study with an exercise intervention, a natural prediction would be that BMI would decrease from baseline and VO2 max would increase relative to baseline. If the observed difference from baseline was a decrease in BMI and an increase in VO2 max, we would consider these two predictions to be correct. Our test statistic is the sum of the correct predictions, weighted using the sample correlation matrix, this insures that predictions for variables that are highly similar will be down weighted.

We show that the prediction test has good power and type I error control, even for very small sample sizes, and that the power increases as the number of measures increases for a fixed sample size. In addition, we discuss the exact distribution of the test statistic and a normal approximation to it when the computational requirements become too intensive. We also include an example showing how the test works with a real data set consisting of ASL measures. We conclude that the prediction test can be a powerful tool for making a go/no-go decision in early stage research, and that the test performs best in scenarios when many methods fail, i.e. small samples but many variables of interest, a common problem for exploratory and feasibility studies.

In Chapter 4 we provide extensions to the Predictions Test as well as a detailed example of using the test with a different data set concerning the effect of kidney transplantation on functional/structural changes in the brain. We use a real data set to compare our method to two other potential types of analyses, a linear mixed model, which would be the ideal way to analyze the data given a large enough sample, and a set of t-tests on the endpoints with a specified primary endpoint, a common choice of analysis for studies with many measures and small sample sizes. We also discuss empirical power estimates for the test, conduct a simulation study comparing the method to a set of t-tests and discuss best practices for the prediction test.

Finally, in Chapter five we discuss the advantages and disadvantages of the new methodologies, including directions for future work. The Appendices include code for the Prediction Test and an

overview of a Shiny Web application for the Prediction test.

Chapter 2

Evaluating Paired Categorical Data when the Pairing is Lost

This chapter has previously been published in whole without any adaptations since publication and is reprinted here with permission. R. N. Montgomery, A. S. Watts, N. C. Burns, E. D. Vidoni & J. D. Mahnken (2019) Evaluating paired categorical data when the pairing is lost, Journal of Applied Statistics, 46:2, 351-363, DOI: 10.1080/02664763.2018.1485013

Abstract

We encountered a problem in which a study's experimental design called for the use of paired data, but the pairing between subjects had been lost during the data collection procedure. Thus we were presented with a data set consisting of pre and post responses but with no way of determining the dependencies between our observed pre and post values. The aim of the study was to assess whether an intervention called Self-Revelatory Performance had an impact on participant's perceptions of Alzheimer's disease. The participant's responses were measured on an Affect grid before the intervention and on a separate grid after. To address the underlying question in light of the lost pairing we utilized a modified bootstrap approach to create a null hypothesized distribution for our test statistic, which was the distance between the two Affect Grids' Centers of Mass. Using this approach we were able to reject our null hypothesis and conclude that there was evidence the intervention influenced perceptions about the disease.

2.1 Introduction

A joint study between The University of Kansas Alzheimer's Disease Center (ADC) and Arts and AGEing KC investigated whether a particular type of therapy called Self Revelatory Performance (SRP) (Emunah, 2015) had an impact on participant's perceptions of Alzheimer's disease. The

research hypothesis was that the SRP had a positive effect on participant's emotional stance. The study's design called for collecting a survey, which consisted of two 5-point Likert items (Likert, 1932) and an Affect grid (Russell et al., 1989), from each participant before and after the performance. Using this paired data the goal was to analyze whether or not each individual's response was affected by the performance and quantify the average individual's shift in perceptions about the disease. Unfortunately, the surveys were collected in such a way (they were put into two piles) that the pairing between subjects was lost. Due to this issue, the original analysis was no longer viable and we were brought on in an attempt to salvage as much information from the data as possible in order to address the research question. This paper focuses on the methodology used to analyze the responses to the Affect grid. The Likert items were analyzed by permuting 10,000 possible unique pairings and using an Ordinal-Quasi Symmetry model as a sensitivity analysis to the effect of the lost pairing. All statistical analyses and data management procedures were conducted in R (R Core Team, 2018).

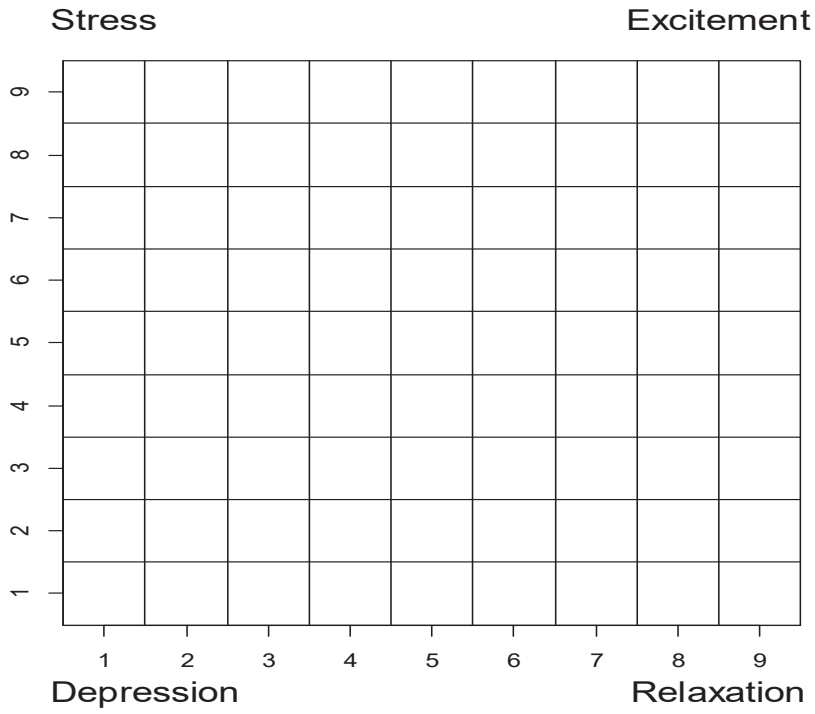


Figure 2.1: The Affect Grid with added row and column numbers

2.1.1 The Affect Grid

Participants were presented with an Affect Grid identical to the one in Figure 2.1 with the exception that we added the row and column labels 1 through 9 after data collection. These labels were added in order to analyze the results and we have no reason to believe the choice of these values affected the analysis. The Affect Grid is an item that addresses two questions at once; the horizontal axis measures valence (on a range from unpleasant feelings to pleasant feelings), while the vertical axis measures arousal (on a range from sleepiness to high arousal) (Kinsinger, 2004). In an Affect Grid participants are prompted to mark a cell that best describes their current combination of valence and arousal. Affect Grids were originally designed to measure a single instance of a participant's emotional state and in these instances has been shown to be a "moderately valid measure of the general dimensions of pleasure and arousal" (Killgore, 1998). However, many authors have used the Affect Grid in paired and longitudinal studies despite the lack of validation in these scenarios. While some work has been done on specific issues related to multiple measurements on the Affect Grid in response to participant's tendency to exaggerate (Russell & Gobet, 2012), we are unaware of any studies that address multiple paired measurements using an Affect Grid to study population wide changes. The original developers of the Affect Grid intended for the item to be scored as two separate measures, which is similar to simply using the Affect Grid to graphically visualize the combination of two Likert items. However, for our analysis we treated the responses to the Affect Grid as a pair, using the coordinates as a combined measure of emotional state. Despite the parsimony that could be achieved by using responses to the Affect Grid as a single response, no validation studies for its use in this scenario have been undertaken. By analyzing the data as a single item we could have unknowingly introduced bias into the responses.

Survey items are often susceptible to extreme responding bias (ER) and central tendency bias (CTB) (Furnham, 1986), which are both concerns with this type of two-dimensional Likert item. The other types of bias often encountered such as acquiescence, or social desirability bias would not be of concern due to the lack of "yes" or "no" questions and the apolitical nature of our topic respectively. If our approach had introduced ER or CTB we would expect to see the responses on

the extremes of the grid or clustered near the middle. The pre-responses In Figure 2.2a show some evidence of ER, with most responses toward the left side of the grid; however the post responses in Figure 2.2b appear to show evidence of CTB with most responses clustered in the middle of the grid. Taken together these provide no evidence of either ER or CTB bias, which cannot occur at the same time. The more likely explanation for the apparent pattern of responses is that the SRP had an effect and shifted the participants more extreme responses to more neutral responses. Overall we cannot rule out the possibility of response bias, however there is no evidence bias was introduced by treating the Affect grid as a single item.

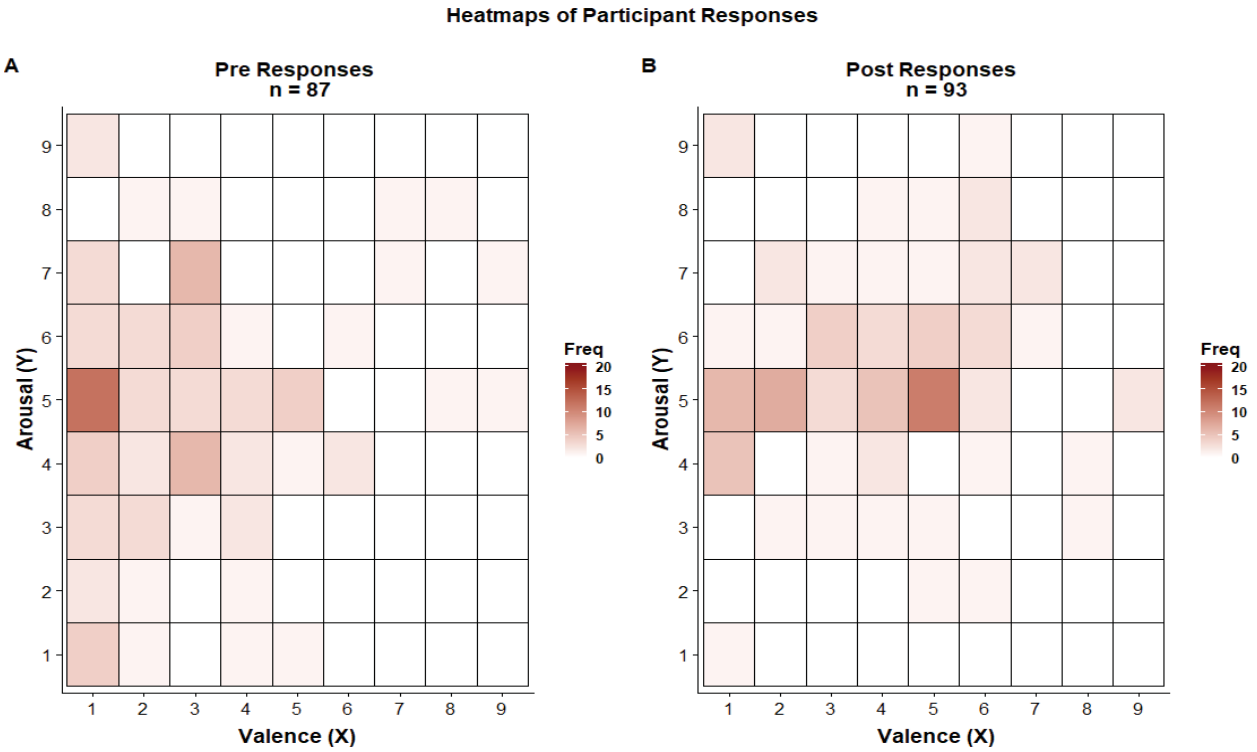


Figure 2.2: Heatmaps of participants responses before and after the SRP

We had a total of 180 observations, 93 completed pre Affect Grids, and 87 completed post Affect Grids. In light of the lost pairing we modified the research hypothesis to the more broad statement that "The SRP had an effect on participant's emotional state concerning Alzheimer's disease." If we concluded there was an effect, then based on the shifts in the Affect Grid we would claim that the shift appeared to be in a certain direction, while noting that we did not test for a directional shift but merely any shift. Thus our null hypothesis (H_0) was that the SRP had no effect

on the participant's emotional state. Under H_0 any differences between the two samples (the pre and post) were assumed to be due to sampling error and therefore the two samples can be viewed as draws from some true distribution of the participant's emotional state, which we refer to as the null distribution. To analyze the data we reformed H_0 to the equivalent statement that the pre and post responses are samples from this null distribution. Under this hypothesis we were able to get an estimate of the null distribution.

2.2 Methods

2.2.1 The Estimated Null Distribution

Without loss of generality, the counts of the cells in the Affect Grid follow a multinomial distribution, where each cell has a certain probability of being chosen such that the total probability equals 1. To form our null distribution these cell probabilities were estimated using the frequency of the total (pre and post) observed cell responses divided by the total number of responses. For example there were 17 total pre and post participants who marked the cell (1,5), thus the relative frequency for this cell was $\frac{17}{170} = 0.10$. We justified combining the pre and post responses to estimate the distribution for two reasons: 1) under our Null distribution these samples are draws from the same distribution, and 2) the sample size was relatively small in relation to the number of cell probabilities we needed to estimate.

There were 93 completed pre Affect Grids and 87 completed post Affect Grids, the combinations of which are graphically displayed by discrete heat maps in Figure 2.2a and 2.2b. Note that both the pre and post samples contained 43 empty cells out of the 81 total cells. Combining these samples into the estimated null distribution resulted in 54 cells with at least one observation; however, even after this combination there were still 27 cells with no responses.

Figure 2.3a displays a heat map of the estimated null distribution with the relative frequencies for each cell. In estimating this distribution we did not want to force the cells with no observations to have an expected value of 0, which would result in a degenerate conditional binomial distribu-

tion for those cells, which we did not believe was an accurate description of the true null distribution. Instead we believe these cells are sampling 0's and following the advice of Agresti (Agresti, 2013) we adjusted the observed null distribution by adding small constants. We added 0.0005 to the estimated probability of the 27 empty cells and reduced the estimated probability of the 54 cells with an observation by 0.00025 so that the total probability would sum to 1. We considered more sophisticated approaches to dealing with these empty cells such as assigning probabilities by weighting the responses near that cell, or by using a function of the row probability multiplied by the column probability. Nevertheless, we felt these approaches would have smoothed the distribution too much, especially given the scarcity of the data. By adding small equal constants to each cell we avoided the issue of having cell probabilities of 0, while staying as close as possible to the observed values.

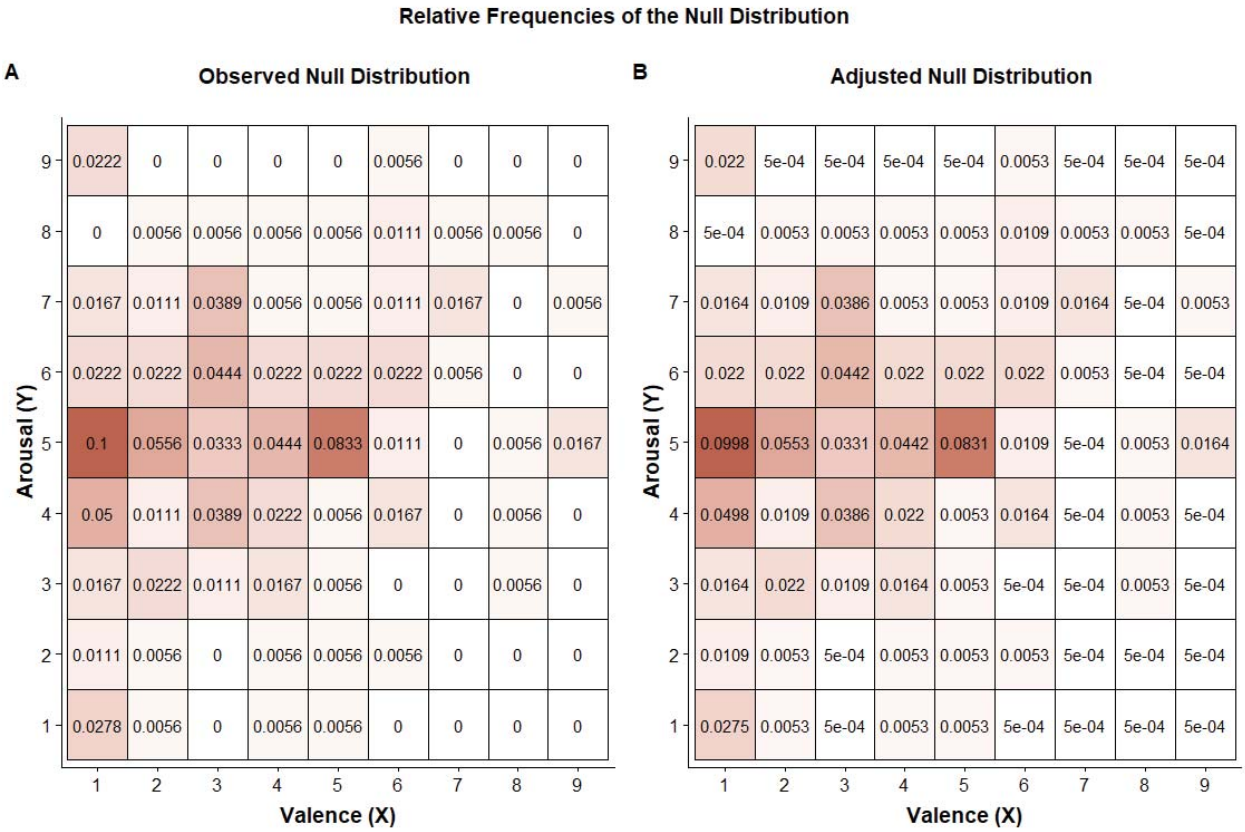


Figure 2.3: Observed and adjusted estimates of the null distriubtion.

The cells with observed responses that would be most influenced by having their relative fre-

quency reduced were the cells with only 1 response. We can see from Figure 2.3a that these cells had a relative frequency of approximately 0.0056. After reducing the observed cell's relative frequency by 0.00025 the relative frequency of the cells with only 1 observation was approximately 0.0053, a reduction of 4.5 percent. While this percent reduction is much greater than the 0.25 percent reduction in relative frequency experienced by the most common cell response (1,5), we felt comfortable that overall it was a fairly modest reduction. Given that we added such small constants to the formerly empty cells, the expected value of a response occurring in any of these previously empty cells is only 1.35 percent. Figure 2.3b shows the adjusted null distribution on the same scale as the observed estimate of null distribution. A sensitivity analysis on the effect of different constants is included in the Results section.

2.2.2 Bootstrap Testing Procedure

In order to analyze our observed data we needed to measure the central tendency of the pre and post samples. To do this we used the center of mass (COM). In a physical system the COM is the point at which the system balances or rests. If we had a physical 9x9 grid with weights corresponding to the participant's responses, the COM would be the point at which we could put a fulcrum and balance the grid. If the responses on the grid shifted so too would the COM. For a two dimensional discrete system of points the center of mass is (x_{cm} , y_{cm}) where

$$x_{cm} = \sum_{i=1}^{n} \frac{m_i x_i}{n} \text{ and } y_{cm} = \sum_{i=1}^{n} \frac{m_i y_i}{n} \tag{2.1}$$

and m_i represents the mass at each point (Protter & Protter, 2009). We summed over the n = 93 participants for the pre center of mass and n = 87 participants for the post center of mass to get the corresponding center of mass (x_{cm} , y_{cm}) for the observed pre and post responses.

To test our null hypothesis that the performance had no effect on participant's responses, we used the Euclidean distance (Deza & Deza, 2009) between the COM of the pre and post samples. We refer to this distance as COMD. Under our null hypothesis these samples come from the same

distribution and therefore we would expect COMD to be small, attributing the difference in the central tendencies of the two samples to random error. To get a measure of the distribution of distances under the null we used a bootstrap approach (Efron, 1979) on our adjusted null distribution. We drew 10,000 samples of size 93 from the modified null distribution and 10,000 samples of size 87 to function as bootstrapped pre and post samples. We then calculated the distance between each of the 10,000 simulated pre and post samples and compared our observed COMD to the bootstrapped distribution of COMD generated under H_0. Our observed pre COM was (2.796, 4.774) and the COM of our post sample was (3.920, 5.333), thus our observed distance was 1.255.

We calculated a bootstrap p-value by summing the total number of bootstrapped distances that were as or more extreme than our observed distance divided by the total number of bootstraps. Due to the inherent sampling variation of a bootstrap p-value we followed the advice of Li, Tai, and Nott (Li et al., 2009) and calculated a confidence interval for our bootstrap p-value. We used the confidence interval originally suggested by Wilson (Wilson, 1927), with a continuity correction (Newcombe, 1998) the confidence interval (L,U) is given as

$$L = max\left\{0, \frac{2n\hat{p} + z^2_{\alpha/2} - [z_{\alpha/2}\sqrt{z^2_{\alpha/2} - \frac{1}{n} + 4n\hat{p}(1-\hat{p}) + (4\hat{p}-2)} + 1]}{2(n + z^2_{\alpha/2})}\right\} \tag{2.2}$$

$$U = min\left\{1, \frac{2n\hat{p} + z^2_{\alpha/2} + [z_{\alpha/2}\sqrt{z^2_{\alpha/2} - \frac{1}{n} + 4n\hat{p}(1-\hat{p}) - (4\hat{p}-2)} + 1]}{2(n + z^2_{\alpha/2})}\right\} \tag{2.3}$$

where $\hat{p}$ represents the estimated p-value, and n is the total number of bootstrap samples. In order to get a good estimate of our p-value we used 10,000 bootstrap samples, which with the assumption that the p-value would be near 0.05 results in an estimated length of the 95% confidence interval being less that 0.01.

2.3 Results

Our bootstrapped distances ranged from .0025 to 1.339; however, only 2 of the 10,000 distances were as or more extreme (i.e. larger) than our observed COMD value of 1.255, resulting in a bootstrap p-value of .0002. Our bootstrap confidence interval for the p-value was (0.00003, 0.0008).

Figure 2.4 displays a histogram of the simulated distances and a dotted line marking our observed distance. While the theoretical distribution of the test statistic COMD is unknown, the observed value of 1.255 clearly lies in the extreme tail of the empirically derived distribution. Under the null distribution we would not expect to draw such different samples from the same distribution.

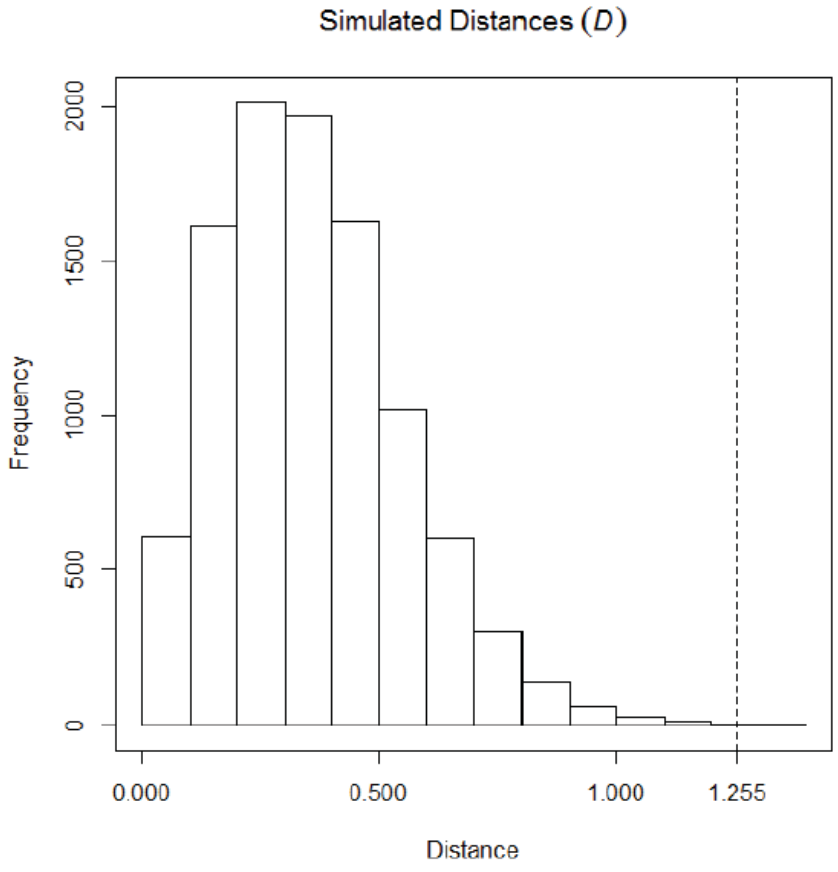


Figure 2.4: Simulated Distances (with observed distance represented by dotted line).

Thus we rejected our null hypothesis that the samples came from the same distribution and therefore rejected the hypothesis that the SRP had no effect. The original research question was whether

the performance had a positive effect. While we did not test for a shift in a certain direction, given that we found evidence a shift had occurred and by examining the observed responses in Figure 2.2, we concluded that there appeared to be a shift in the positive direction (i.e. toward the right side of the grid). We take this as evidence that the SRP had an effect, specifically a positive shift in participant's perceptions about Alzheimer's disease.

2.3.1 Sensitivity Analysis

To evaluate whether either the value we chose to add to the empty cells or possible dependencies between the pre and post responses could have affected our results we conducted a sensitivity analysis for both potential problems.

For the analysis of the effect of added constant we allowed the value chosen to vary among six different levels, the smallest being 0.0001 and the largest being 0.011. As the weight increases so does the amount of smoothing we are applying to the observed data. We would expect that smaller values would have little effect on our results since they would be closest to the actual observed data, which is why only two of the new values are below our original value of 0.0005. The maximum weight we looked at was restricted to 0.011, the reason for this is that we did not want to smooth the data too much and restricted ourselves to removing an equal probability from each of the cells that already contained an observation. The cells that only had one observation had an estimated probability of 0.0056 and thus we could take a maximum of .0055 (allowing for a very small remaining probability) away from the cells with one observation. This corresponds to adding a value of 0.011 to the empty cells. We drew 10,000 samples of size 93 for the pre sample and size 87 for the post sample, calculated their center of mass, and recorded how many of them were as or more extreme than our observed distance.

Table 2.1: Sensitivity Analysis

Weight	Empirical P-value
0.00010	0.0001
0.00025	0.0001
0.00075	0.0005
0.00100	0.0004
0.00250	0.0006
0.01100	0.0024

Our original weight of 0.0005 resulted in an empirical p-value of 0.0002.

Table 1 shows that as the weight increased so did the empirical p-value; nevertheless for all the weights chosen the p-value was still far below the significance level of 0.05. Of course by making the weight arbitrarily large the p-value could potentially be greater than our significance threshold, but the weight would need to be larger than 0.011.Thus we would need to reduce the estimated probability of the cells with observations in an unequal manner (taking more away from cells with more observations, or perhaps based upon regions of observations). There is no evidence to suggest that such an invasive redistribution of the weight of observed values is warranted, thus we concluded that our analysis was not sensitive to the choice of weight.

In addition we also looked at whether different levels of correlation among the responses could have had an effect on the results. We generated 100,000 pre and post draws from the null distribution, recorded the Pearson correlation between the pre and post x, pre and post y and also the distance between the pre center of mass and post center of mass. We then fit several linear models with COMD as the response and main effects and interactions of both the correlation in the x axis and the correlation in the y axis. We fit this model with the full 100,000 draws and also a subset of the draws which resulted in large, i.e. greater than 1, values of COMD. In both of these

models neither the main effects or the interaction term were statistically significant. Thus we don't believe that possible correlation between the pre and post responses would have greatly impacted our results.

2.4 Discussion

2.4.1 Limitations

While we feel confident in the result of this study there are limitations to the method we used to analyze the data. One concern is that the lack of validation of the Affect Grid as a single item introduced response bias that we are unaware of. However, we did not see any evidence of the most common types of bias. We reiterate that we only tested for a shift in perceptions, we found evidence that this shift occurred and based on the apparent positive direction of the shift we feel the SRP had a positive impact. While it is highly unlikely that the Affect Grid would be biased in such a way that the apparent shift toward the positive side of the grid was actually indicative of a different type of response this possibility must be entertained and is an unavoidable symptom of the lost pairing.

Another issue with our approach is that it has poor power to detect specific types of shifts. As an example, Figure 2.5a displays a heat map of possible pre responses that would indicate the population had extreme feelings about Alzheimer's disease. Figure 2.5b displays a heat map of possible post responses that would indicate the population had very neutral feelings. While these two populations clearly differ in their observed emotional state they have the same center of mass, and therefore our test statistic would detect no difference between them. In general, any systematic shift or rotation around the pre center of mass would result in an identical or very similar post center of mass. In theory, there could be many shifts within the population so long as they did not alter the center of mass, none of which would be detected by our test statistic. It is important to note that the shift could merely be a change in the dispersion or variance between our two samples. If the both the dispersion and the COM changed we would detect this difference, but if the dispersion

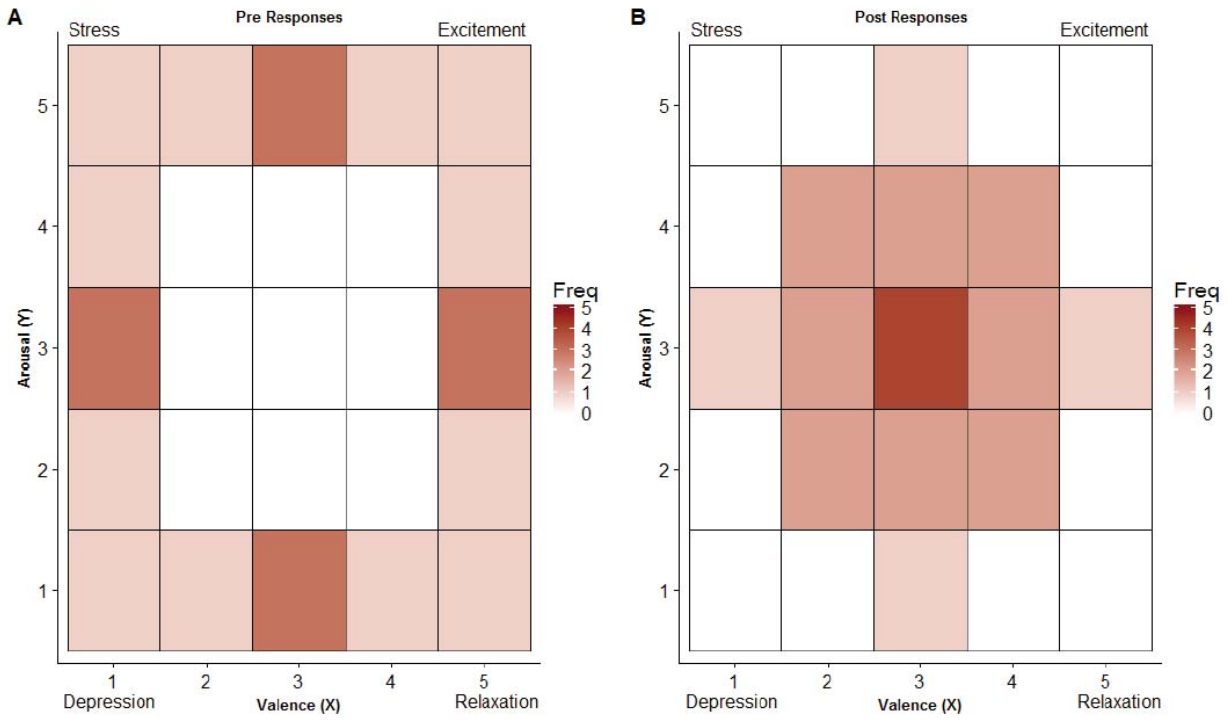


Figure 2.5: Theoretical shift which would result in COMD = 0 despite obvious differences in the samples

changed from pre to post while the COM remained unchanged our approach would not detect this difference because we were concerned with a change in the central tendency of the samples. If this method was used and the null hypothesis was not rejected our method provides no way to definitively know whether there was no evidence of a shift or that there was an undetectable shift. In our study we rejected the null and thus this limitation was not an issue for this particular data set.

2.4.2 Validation and Generalizations

In order to check the validity of our method we did a simulation study to compare its power and type I error rate to that of Hotelling's T^2 (Hotelling, 1931). We chose Hotellings T^2 which is a generalization of the t-statistics for multivariate testing because we were testing the change of central tendency across two dimensions. There is no standard test designed for the situation we found ourselves in. However Hotelling's T^2 is a logical choice in the same way that the on

dimensional t-test is often used to analyze Likert data. While Hotellings T^2 assumes Normality and data from the Affect grid is clearly discrete, it is fairly robust to departures from Normality (Mardia, 1975). In addition we used sample sizes of 100 per sample so that by the multidimensional central limit theorem the samples of coordinate pairs should be fairly Bivariate Normal.

To compare COMD with Hotellings T^2 we created nine distributions that we felt were representative of what might be encountered when using the Affect Grid. Figure 2.6 shows heatmaps of the distributions. These were created by thinking of the data as a combination of two 9-point Likert items which correspond to the x and y axes. Often Likert data exhibit either extreme responses or central tendency bias, which provides three options for each of the axes. Values can be positive skewed, negative skewed or centered in the middle of the nine points. For instance in Figure 2.6 the distribution (Positive, Center) is a combination of right skewed responses on the x-axis and centered responses on the y-axis. The vectors of cell probabilities for Positive, Negative, and Centered are found in Table 2 and thus the cell probabilities for these distributions can be found by multiplying those vectors in the appropriate way. For instance the cell probabilities for the Distribution titled (Right, Centered) are found by calculating $R^T C$. The cell probabilities for the other eight distributions in Figure 2.6 can be calculated similarly.

Table 2.2: Probability Vectors used to Generate Distributions

	1	2	3	4	5	6	7	8	9
R^T	.22	.20	.15	.135	.085	.06	.055	.05	.045
C^T	.055	.075	.1	.16	.22	.16	.1	.075	.055
L^T	0.045	.05	.055	.06	.085	.135	.15	.20	.22

These values represent the probability of a given number 1,...,9 being chosen. Using these vectors we can get the cell probabilities for the distributions.

We wanted to compare how COMD performed relative to Hotelling's T^2 in a scenario where we were drawing equal sample sizes from two distributions. For the distributions shown in Figure

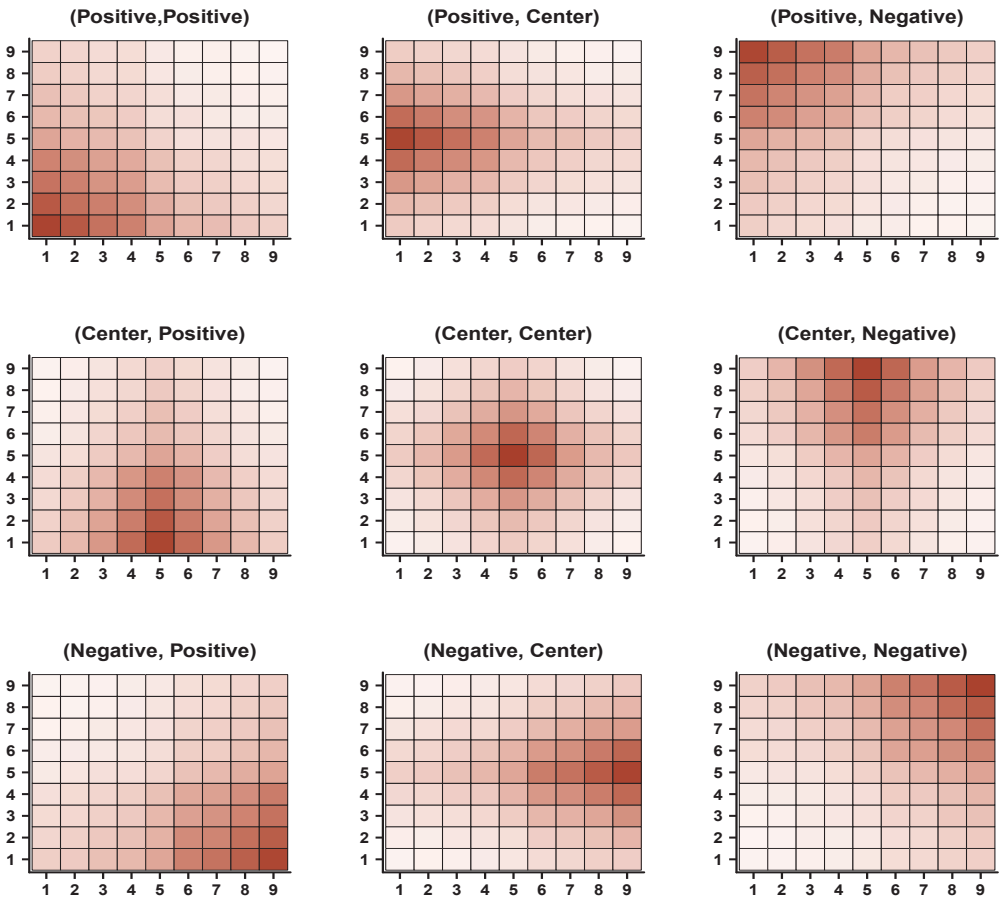


Figure 2.6: Heatmaps showing the differences in cell probabilities for the nine distributions

2.6 there are 36 combinations of the distributions and the 9 scenarios where both samples are actually coming from the same distribution. For the 36 combinations of different distributions we estimated the power for COMD and Hotelling's T^2 by drawing $1,000$ random samples from each distribution and calculating the p-value for Hotelling's T^2 and the bootstrapped p-value for COMD. Our estimate of power was then the percentage of those simulations that had resulted in a p-value ≤ 0.05. Similarly for the estimate of Type I error, we took the percentage of significant results out of the total number of simulations. For the COMD method on each of the $1,000$ samples we used a bootstrap of size $1,000$.

Table 3 provides estimates of the power in detecting differences between the pairs of distributions. The estimated power of COMD is comparable to Hotelling's T^2 in every instance we looked

Table 2.3: Estimated Power for COMD and Hotellings T^2

	PC	PN	CP	CC	CN	NP	NC	NN
PP	.914(.915)	1(1)	.918(.929)	.999(.999)	1(1)	1(1)	1(1)	.909(.910)
PC	—	.908(.913)	1(1)	.943(.924)	.996(.996)	1(1)	1(1)	.917(.918)
PN		—	1(1)	.9(.912)	.907(.911)	1(1)	1(1)	1(1)
CP			—	.927(.915)	1(1)	.920(.926)	.998(.998)	1(1)
CC				—	.934(.921)	.999(.999)	.931(.923)	.999(.999)
CN					—	1(1)	.999(.999)	.907(.911)
NP						—	.913(.921)	1(1)
NC							—	.915(.920)

Table 2.4: Estimated Type I error for COMD and Hotellings T^2

	PP	PC	PN	CP	CC	CN	NP	NC	NN
Hotellings T^2	0.062	0.039	0.044	0.050	0.036	0.045	0.044	0.055	0.052
COMD	0.058	0.045	0.045	0.053	0.039	0.054	0.047	0.060	0.054

at. Both methods achieved greater than 90% power for comparing these specific distributions. A more comprehensive simulation study that drew samples from more similar distributions or looked at different sample sizes would have yielded different and probably lower power estimates for both COMD and T^2. Nevertheless, both methods performed very well under these conditions.

Similarly, Table 4 provides estimates of the Type I error rates, when the two samples came from the same distribution. Again we see that COMD and Hotellings T^2 provided very similar results. The Type I error rate was controlled around the nominal 0.05 level for both COMD and Hotellings T^2. A more extensive simulation study comparing the two methods is needed to determine which method performs better in certain scenarios. In addition more work on the operating characteristic of COMD would need to be done in order to be comfortable using this approach outside of this specific instance. Nevertheless, in this set of simulations it has shown itself to be at least as good as Hotelling's T^2 while not making as many assumptions about the data. We believe it shows promise as a way to analyze discrete grid data, specifically data that can be thought of as combinations of Likert items.

2.5 Conclusions

The loss of pairing and the type of data encountered in this problem presented a unique challenge for analysis. Using our methodology we were able to address the research hypothesis and provide evidence in its favor. The extreme nature of the observed COMD and the apparent direction of the shift toward the right side of the grid lead us to believe that Self Revelatory Performance had an effect on participant's responses, specifically a positive effect. More work would need to be done to validate SRP as an effective tool for educating seniors about Alzheimer's disease, but both the therapy and the methodology developed show promise for application outside of this study.

Chapter 3

The Prediction Test: A go/no-go hypothesis test for early stage research studies

Abstract

This paper introduces a global hypothesis test for studies with many measures intended for use at early stages of research when a go/no go decision is needed on whether to continue on the current research track. We believe this will be especially useful when the number of measures collected is large and the sample size small. Our test makes use of a priori predictions about the direction of the result for each measurement provided by the researcher, we then weight these predictions using the sample correlation matrix. The global alternative hypothesis is that the researcher's ability to predict the results of each measure, ϕ, is greater than a null hypothesized value ϕ_0. If a researcher is able to successfully predict many of the study measures this would provide evidence that the researcher's theory about the underlying phenomenon or process is correct and worthy of more study. We show that this method has adequate power and type I error control even in situations where the number of measurements relative to the sample size is large and the measures are correlated, a situation under which traditional approaches have poor power.

3.1 Introduction

In contemporary biomedical science many studies involve multiple outcomes. While significant advances in the methodologies to deal with multiplicity have been developed, the tendency in biomedical research to measure as many endpoints of interest as possible within each study is not well suited to balancing sufficient multiplicity adjustments along with sample size, financial, and logistic constraints, especially in early stage research. We present a paradigm for experimental testing that controls the type I error rate of the overall experiment (or more specifically of the

global research hypothesis) while maintaining adequate power. We believe this methodology will be especially useful for early stage research.

The scenarios in which multiple endpoints arise are varied, ranging from trials with distinct outcomes of interest, such as endpoints for both efficacy and safety, to trials where the complex nature of or the inability to directly measure the hypothesis of interest necessitates multiple endpoints that can be used as proxies. (FDA, 2017) Having more than one endpoint can have serious implications on the operating characteristics of a study and typically needs to be addressed. This has led to a wealth of literature about the topic (Dmitrienko & D'Agostino, 2017), including guidance from government agencies about best practices. (EMA, 2017) (FDA, 2017) Common approaches to the issue of multiplicity include single step procedures such as the Sidak adjustment (Sidak, 1967), multi-step procedures such as the Holms method (Holms, 1979), global procedures such as Sime's Global test (O'Briend, 1984) as well as many related methods such as False Discovery Rate, gate keeping procedures, etc. (Benjamini & Hochberg, 1984) (Dmitrienko et al., 2009). However, despite these advances, when faced with multiple endpoints a common default continues to be the designation of a primary endpoint, with all others analyzed as secondary, or exploratory, even if there is not a true "primary" endpoint. Our test provides a method for arriving at a go/no-go decision when there is not a true primary endpoint, a common occurrence in early stage research when the validity of the overall theory or research hypothesis is being tested. We incorporate the correlation of the measured endpoints along with the researcher's predictions to form our test statistic.

Section 3.2 provides a motivating example for our methodology. In section 3.3 we introduce the test statistic and associated hypothesis test, section 3.4 presents an empirical assessment of the Normal approximation, Section 3.5 provides simulated empirical power and type I error estimates over various experimental setting, in Section 3.6 we look at the sensitivity of using a small sample correlation matrix, Section 3.7 discusses the choice of the parameter ϕ and maximum hypothesized values for a given number of measures, Section 3.8 applies our methodology to an example data set and in Section 3.9 we discuss the methodology presented including its limitations and areas

for future work. The supplemental material provides R code for implementation of the Prediction Test, a proof showing the Lindberg condition of the central limit theorem holds and more detail on the analysis of the example data set.

3.2 A simple example

Consider an experiment concerning the population mean. Let the null hypothesis be $H_0 : \mu \leq 0$ vs $H_1 : \mu > 0$. At the boundary point, 0, between the null and alternative space, the sampling distribution of the sample mean under the null hypothesis is normal with $\mu = 0$. We consider a simple statistical test with decision rule that would reject the null if we observe $\bar{x} > 0$, and fail to reject the null for $\bar{x} \leq 0$. The rejection region for this test is shown in Figure 3.1 A. Under the null hypothesis the type I error rate for this test is 0.5, which is unacceptably high for most, if not all, conceivable applications. However, the power will also be high since if the population mean is > 0 the sample mean often will be as well. For instance, if the true population were normally distributed with $\mu = 0.5$ and $\sigma = 1$, with $n = 10$, the probability of the sample mean being greater than 0 is almost 95%.

Now consider a similar test of two independent population means defined in the same way. Let the global null hypothesis be that neither mean is greater than 0; $H_0 : \mu_1 \leq 0 \cap \mu_2 \leq 0$. Here we define a statistical test which will reject the null if and only if both observed sample means are greater than zero. This rejection region is shown in Figure 1 B. The type I error of the first test is 0.5, and the Type 1 error of this second test is 0.25. If we added three additional independent, standard normal endpoints and required positive responses from all five measures to reject the null then the type I error would drop below the conventional 0.05 threshold to 0.03125, thus controlling the error rate for this test by simply requiring results in the direction of our alternative hypothesis, no matter how large. Of course, to reject this global null hypothesis we would require all five endpoints to show an increase which is a fairly strict requirement that could lead to a reduction in power.

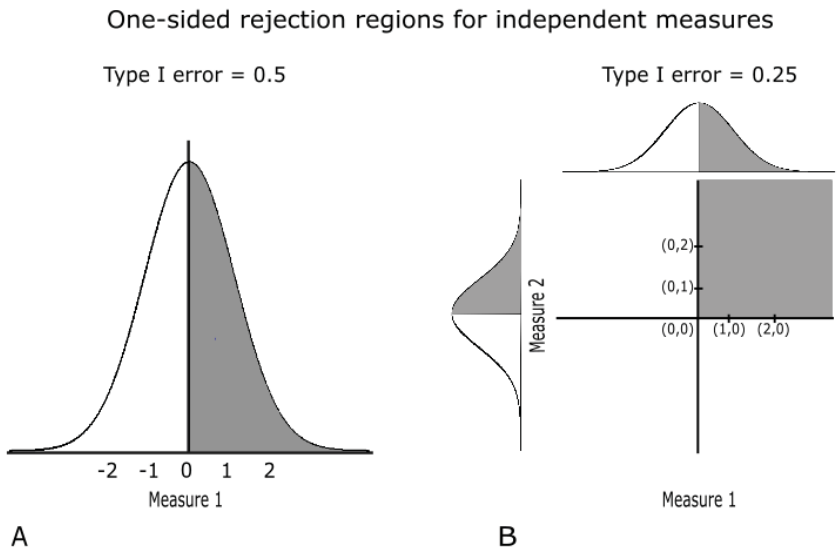


Figure 3.1: For m independent one-sided measures where we would reject the global hypothesis if each measure showed a positive response the Type I error would be 0.5^m. Examples are shown for $m = 1$ and $m = 2$. The density shown over the possible responses is the null distribution, N(0,1).

Finally consider a more realistic scenario of two standard normal endpoints with some level of dependence between them. The type I error rate for the global test would be between 0.25, which would be the case if the two endpoints were independent, and 0.5, if the endpoints were perfectly dependent, i.e. having a pairwise correlation of ± 1. As more endpoints are included and also required to be positive for rejection then the overall type I error rate will continue to decrease, provided that these endpoints are not perfectly dependent. Given a large enough number of endpoints with one sided rejection regions we could control the type I error rate even in situations of extreme multiplicity, however rejection of the null would require that the sample means all be greater than 0. We have developed a statistical test that makes use of the correlations between endpoints to better measure the true underlying distribution for a global test of endpoints under the null hypothesis using one-sided rejection regions. Further we have done so in such a way that does not require the correct directional prediction for all endpoints and thus a rejection criterion can be selected that can balance controlling the type I error rate with maintaining adequate power.

Our methodology asks researchers to make a prediction about the direction of each endpoint, or measure, for example that the sample mean of a measure will be greater than some value, such

as 0. We give the researcher credit for correct predictions and weight these predictions using the sample correlation matrix so that correctly predicting more "independent" endpoints leads to a larger test statistic. Our null hypothesis concerns the researcher's ability to predict the result of the endpoints, which we denote with the parameter ϕ, which can take values in $\{0,1\}$. We hypothesize that $H_0 : \phi \leq \phi_0$. If the researcher can predict the direction of many measures, something unlikely due to chance, we would have evidence that their theory was providing more understanding of the natural phenomenon and conveying an advantage in the predictions, i.e. it was correct, while few correct predictions would cast doubt on the validity of the theory. In addition we weight the predictions depending on how "independent" the underlying measures are, for instance if two measures have a high pairwise correlation, say $r = 0.90$ correctly predicting both of them is only slightly more impressive than correctly predicting one of them, thus we down weight measures that are highly correlated with other measures in the data set. In this way, we require not only a good deal of correct predictions to reject the null hypothesis, but correct predictions of measures that are not simply copies of one another.

3.3 Test statistic and associated test

Let n represent the number of experimental units, and m be the number of measurements taken on each experimental unit. We set our predictions to be one sided, so that without loss of generality we describe these as being predictions of either a positive or negative result for each measure. Let $\mathbf{p}$ be an $m \times 1$ vector of the results of the predictions for the measures, where p_i, the i^{th} value of the vector, is an indicator function that equals 1 if the prediction on measure i is correct, and 0 if the prediction is incorrect. Let $\mathbf{C}$ represent an $m \times m$ correlation matrix between the measures, where ρ_{ij} is the pairwise correlation between measure i and j, which we estimate with the sample correlation r_{ij}. Any type of correlation measure can be used with the choice depending on the underlying data (Maturi & Elsayigh, 2010), for all examples and simulations we use Pearson's correlation coefficient.

For the i^{th} measure we have defined a weight w_i, $i = 1, \ldots, m$, that is the inverse of the sum of the squared pairwise correlations for the i^{th} row of **C**, that is

$$w_i = \left(\sum_{j=1}^{m} r_{i,j}^2 \right)^{-1}$$

Table 3.1: Correlation matrix C between the measures, with weights

C	1	2	.	.	m	Weights
1	r_{11}	r_{12}	.	.	r_{1m}	$w_1 = (r_{11}^2 + \ldots + r_{1m}^2)^{-1}$
2	r_{21}	r_{22}	.	.	.	$w_2 = (r_{21}^2 + \ldots + r_{2m}^2)^{-1}$
.	.	.	.	.	.	.
.	.	.	.	.	.	.
m	r_{m1}	.	.	.	r_{mm}	$w_m = (r_{m1}^2 + \ldots + r_{mm}^2)^{-1}$
						W = $\sum_{i=1}^{m} w_i$

The weight w_i will take a value of $1/m$ for all i when there is a perfect pairwise association (positive or negative) between each measure and will take a value of 1 for all i when the measures are independent. Thus W, the sum of these weights, will equal 1 if the correlation matrix is a matrix of ones and will equal m if the correlation matrix is an identity matrix. Conceptually we view W as an estimate of the number of "unique" or effective endpoints being considered, similar in spirit to the idea of the effective number of variables advanced by several authors in significance threshold correction (Chevrud, 2001) (Li & Ji, 2005). When the measures are perfectly independent $W = m$ implying we have m unique endpoints, while if the measures are perfectly dependent $W = 1$ implying there is effectively only one independent measure. For levels of correlation between these extremes $1 \leq W \leq m$. For example, if data was collected on 10 measures, but W = 4.8, we would take this as an implication that the real number of unique endpoints being examined by the study was between 4 and 5, that is some of the 10 endpoints were measuring the same thing.

Our test statistic increases in value for each correct prediction by the corresponding weight w_i. We define our test statistic as

$$T_m = \sum_{i=1}^{m} p_i w_i$$

where $0 \leq T_m \leq W$. Notably, larger values of T_m indicate experimental results more aligned with the researcher's hypothesized predictions, while those closer to zero imply less concordance between prior predictions and experimental results. It's important to note that for the same number of correct predictions the value of the test statistic will change depending on which measures are correctly predicted, thus predicting measures with higher weights, will result in larger test statistic values. We give greater importance to correctly predicting measures that are more independent of the other variables in the data set, in this way correctly predicting a large amount of highly correlated measures may lead to a relatively small test statistic.

3.3.1 Test Statistic under the Null

Our null hypothesis is that the researcher's predictive ability ϕ is less than or equal to ϕ_0. The parameter ϕ is chosen in regard to the specific experiment of interest. For example, in early discovery experiments if the researcher was able to predict the result of more than $\phi_0 = 0.50$, i.e. 50%, of the measures then perhaps that would be enough to warrant further study because it would indicate that the researcher's theory was able to predict what would happen at a rate better than chance. Under our null hypothesis we assume that the results of each prediction p_i follow a Bernoulli distribution with success parameter ϕ. We also assume that the weights, w_i, are independent of the predictions. If the weights are treated as fixed our test statistic is therefore a weighted sum of Bernoulli random variables, with expected value $E[T_m] = \phi \sum_{i=1}^{m} w_i$ and $Var(T_m) = \phi \cdot (1-\phi) \sum_{i=1}^{m} w_i^2$.

If the correlation matrix $\mathbf{C}$ is such that the off-diagonal values are not all equal then there are a discrete number of unique $w_i's$ and the vector $\mathbf{p}$ can take on 2^m permutations, as each $p_i \in \{0,1\}$. Given a sample correlation matrix we can calculate the exact distribution of the test statistic. In doing this we need to consider that different combinations of correct predictions will result in

different values of the test statistic even if the overall number of correct predictions is the same. Unlike a binomial probability we do care about which specific combination leads to x correct predictions, thus the pmf of our test statistic can be considered as a binomial pmf without the constant $\binom{m}{x}$ where $x = \sum_{i=1}^{m} p_i$, i.e. the number of correct predictions for the observed statistic t_m.

More formally the pmf is

$$f(T_m = t_m) = \phi_0^{\sum p_i}(1-\phi_0)^{m-\sum p_i}$$

When $\phi_0 = 0.50$ the probability of all possible values of T_m will be equal due to the symmetry of the binomial; however as ϕ_0 deviates from 0.50 the probability of observing different values of the test statistic will change.

3.3.2 Special Cases

There are two special cases concerning the distribution of our test statistic, when $\mathbf{C} = \boldsymbol{J}_m$, an $m \times m$ matrix of ones, and when $\mathbf{C} = \mathbf{I}$, the identity matrix.

For $\mathbf{C} = \mathbf{I}$, $w_i = 1$ for all i thus our test statistic can be written as:

$$T_m = 1 \cdot \sum_{i=1}^{m} p_i$$

The sum of independent Bernoulli random variables is a Binomial random variable. Thus, the test statistic would simply follow a Binomial(m, ϕ_0).

If the sample correlation matrix $\mathbf{C} = \boldsymbol{J}_m$ (with the off diagonals being either positive or negative one) then $w_i = 1/m$ for all i. Our test statistic in this scenario can be written as:

$$T_m = \frac{1}{m} \sum_{i=1}^{m} p_i$$

If we let $X = \sum p_i$, and let Y be the transformation $Y = \frac{1}{m}X$ we can show the probability mass function of T_m is:

$$\binom{m}{y \cdot m} p^{y \cdot m} (1-p)^{m - m \cdot y}$$

with support $y = \{0, 1/m, 2/m, \ldots, 1\}$. Thus T_m is simply a linear transformation of a Binomial random variable in the completely dependent case. For sample data the only realistic way either of these scenarios could occur would be through error, or in a contrived way such as measuring the same variable but with different units, for example height in inches, centimeters and meters.

3.3.3 Decision Rule

We are most concerned with whether the researcher's theory provides an advantage in understanding the outcome of different measures, thus we define our null and alternative hypotheses as follows:

$$H_0 : \phi \leq \phi_0$$

$$H_1 : \phi > \phi_0$$

We also require that $T_m > 1$ in order to reject the null, that is we require the sum of correct scores to be greater than 1. This forces the researcher to correctly predict every endpoint when the endpoints are perfectly dependent, thus simply correctly predicting linear, or monotonic combinations of other endpoints provides no advantage, i.e. the null hypothesis cannot be rejected when $\mathbf{C} = \boldsymbol{J}_m$. We consider the sum of T_m to be the number of effective endpoints correctly predicted and thus it is intuitive we would require the researcher to correctly predict endpoints with a value greater than one. In situations where the Type I error rate was of most concern ϕ would be close to 1, and in situations where power was the main concern ϕ would be closer to 0, although we would recommend $\phi_0 \geq 0.5$ without strong reasoning otherwise.

3.4 Normal approximation

The exact distribution for T_m becomes computationally difficult to calculate as the number of endpoints increases, with $m = 30$ the number of possible permutations of $\mathbf{p}$, the prediction vector, is over 1 billion. In the supplementary material we have shown that the sum of the $T_i's$ satisfies the Lindberg CLT, indicating that as the number of endpoints increase the central limit theorem will apply as long as the value of ϕ_0 is bounded away from 0 and 1 (Billingsley, 1995). We show these restrictions on ϕ_0 will always be met as we let $m \to \infty$. We also conducted a simulation study for various values of ϕ and m. We found that if m is large enough and ϕ is not too close to either boundary:

$$T_m \sim Normal(\mu, \sigma)$$

where $\mu = \phi \cdot W$ and $\sigma = \sqrt{\phi(1-\phi) \cdot \sum w_i^2}$, the distribution of our test statistic can be approximated with a Normal distribution.

We looked at values of m ranging from 20 to 70 in increments of 5, with $m > 20$ the calculation of the exact distribution becomes computationally expensive. We generated random correlation matrices using the R (Team, 2017) package clusterGeneration (Qui & Joe, 2015) which uses partial correlations and a recursive method to generate a random m-dimensional covariance matrix which we converted to a correlation matrix (Joe, 2006). The method depends on the dimension of the correlation matrix and on a parameter α_d. We used $\alpha_d = 1$ which is a special case that is uniform over the space of positive definite correlation matrices. For each value of m, we simulated 100 different correlation matrices, for each of these correlation matrices we randomly generated 1,000 sets of predictions from a Bernoulli(ϕ_0) distribution. We used these predictions and the correlation matrix to estimate the exact CDF of T_m which we then compared to the approximate CDF using the normal approximation. We calculated the mean absolute error (MAE), i.e. the sum of absolute differences between the values divided by the number of approximated percentiles, 1,000 in this case. We took the mean of these 100,000 MAE's (1,000 for each of the 100 correlation matrices) to form the grand mean absolute error (GMAE), averaging over the generated correlations matrices

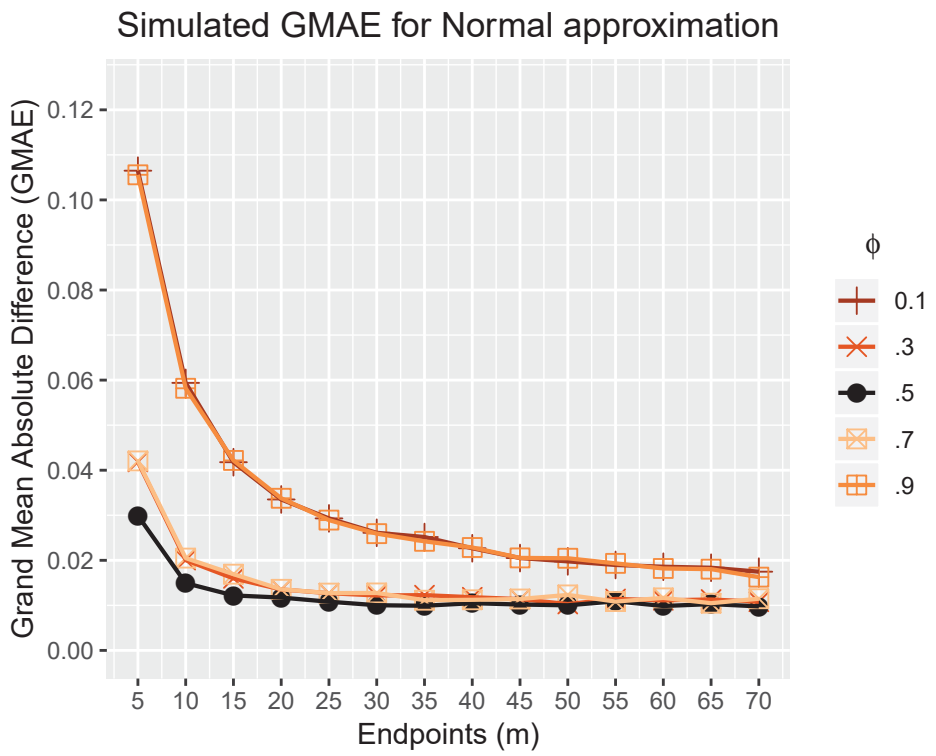


Figure 3.2: For each combination of m and ϕ_0 we calculated the GMAE. We see a clear decrease in GMAE as the number of endpoints increases.

for each combination of m and ϕ. Figure 3.2 shows the results.

When ϕ_0 is near one of the boundaries the approximation is poor especially with a small number of endpoints. However, as the number of endpoints increases the GMAE decreases even when ϕ_0 is near the boundary. With a more moderate value of ϕ_0 the normal approximation is much closer to the value from the exact distribution, with $\phi_0 = 0.50$ the error quickly approaches 0.01 on average.

3.5 Simulation Study of Power and Type I error

Using simulation, we evaluated the empirical type I error and power of our test. We considered situations with between 5 and 40 measures, set our sample size to 20, and considered hypothesized values of $\phi_0 = 0.50$ and $\phi_0 = 0.70$. The simulated sample data were all drawn from a N(1,1), and the predictions were all of an increase so the effect size for each measure is 1. Thus the predictions were all correct but due to the variability of the measures the sample mean will not always be

greater than 0. For each of these hypothesized values we examined the operating characteristic for true values of ϕ in the null and alternative space. For both hypothesized values we examined the operating characteristics of $\phi = \phi_0$ as well as two values such that $\phi < \phi_0$, i.e. the null space and four values such that $\phi > \phi_0$, the alternative space. We plotted the results in Figure 3.3, values in the alternative scale are on a blue scale and values in the null space are on a red scale. For each combination of m, ϕ and ϕ_0 we randomly generated 50 sample data sets and simulated the results 200 times on each sample for a total of 10,000 simulated results for each combination. The plotted values are the number of times out of 10,000 that our test statistic would have led to rejecting the null hypothesis at $\alpha = 0.05$.

With $\phi_0 = 0.50$ we see that empirical power quickly increases as both m or ϕ increase. For example, with $\phi = 0.8$ and $m = 15$ we achieve approximately 80% power to reject the null hypothesis. Conversely the empirical Type I error rate is always at or below the nominal 0.05 level.

For $\phi_0 = 0.70$ the empirical power and type I error estimates when $m = 5$ are both 0, this is due to the discreteness of the test statistic. Under the null of $\phi_0 = 0.70$ about 16%, $P(\sum_{i=1}^{m} p_i = 5 \mid m = 5, \phi_0 = 0.70)$, of the time all 5 predictions would be correct, thus the null hypothesis cannot be rejected at the typical $\alpha = 0.05$ level for this combination of m and ϕ_0. However, as m increases the power increases and the type I error is controlled at 0.05. It is important to note that for certain hypothesized value of ϕ_0 the power will be 0 unless m is large enough, a table of minimum values of m for common values of ϕ_0 is provided in Section 3.7. The problem of discrete test statistics is not new and when appropriate we can use techniques such as the mid p-value. We also note that the sample size, n, only comes into play indirectly though the sample correlation matrix. For a fixed n, power will be non-decreasing as the number of measures increases, so the inclusion of as many endpoints that are relevant to the research hypothesis as possible can be encouraged.

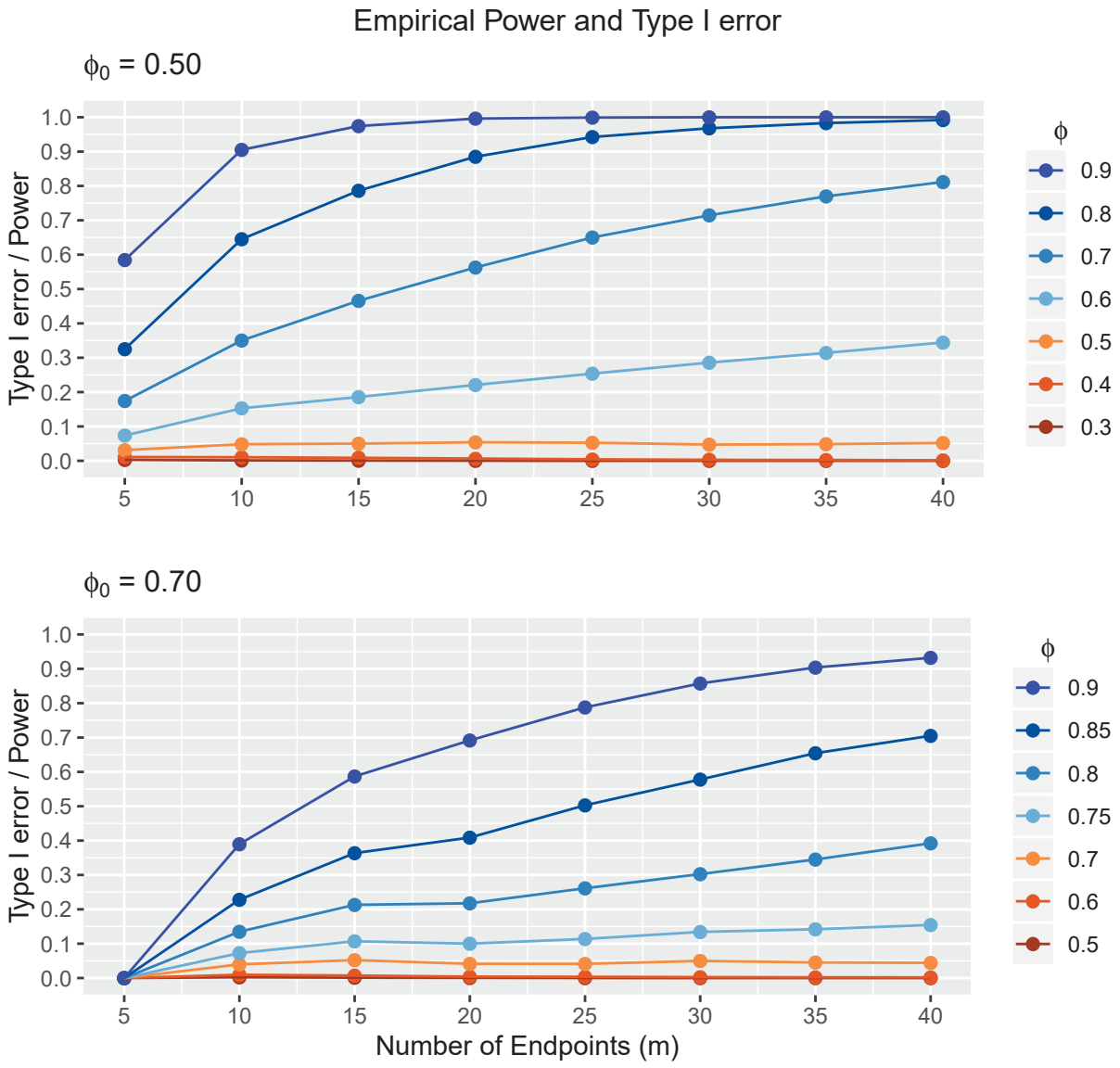


Figure 3.3: Empirical Power and Type I error. These estimates are the percent of times we rejected the null hypothesis of ϕ_0 for combinations of m and ϕ.

3.6 Sensitivity analysis of sample correlation matrix

A concern when applying our method could be the instability of the sample pairwise correlations for a small sample size, n, with some estimates of the required sample size being in the hundreds depending on the application (Bonnett & Wright, 2000). Most of the literature concerning adequate sample sizes for sample correlations has been focused on doing hypothesis testing on a single sample correlation or estimating a confidence interval for a single sample correlation. While we are not directly interested in using the pairwise correlation for testing it is of concern that unstable pairwise correlations could affect our weights and therefore affect the overall test statistic. To examine this situation, we conducted a simulation study. We used the same combinations of ϕ_0, ϕ and m as in the Type I error and Power simulation in Section 5. We generated a "true" correlation matrix, then sampled from a multivariate normal in R using the true correlation matrix to generate a sample, calculated the "true" weights and the sampled weights and simulated our tests for n = 20. We kept track of the proportion of times that our test came to the same conclusion, these values are presented in Table 3.2.

Table 3.2: Proportion of tests coming to the same conclusion for combinations of n, ϕ_0 and m. Average Agreement calculations do not include m = 5 for $\phi_0 = 0.7$ since the null cannot be rejected in this case.

n	ϕ_0	m = 5,	10,	15,	20,	25,	30,	35,	40	Average Agreement
10	0.50	100	97.1	96.6	97.4	97.7	97.9	98.1	98.2	**97.6%**
	0.70	100	97.0	96.4	98.0	97.0	96.3	96.2	96.9	**96.8%**
20	0.50	100	97.9	97.4	97.9	98.1	98.4	98.5	98.4	**98.1%**
	0.70	100	97.8	97.0	98.5	97.8	96.8	96.6	97.2	**97.4%**

The results show that estimating the true correlation matrix with a sample correlation matrix, even for with a relatively small n, leads to the same conclusions as if we had the true correlation matrix. Despite these results if there is concern over the validity of the sample correlation matrix, we would suggest eliciting expert opinions or using historical studies to estimate the matrix, which can then be used for a sensitivity analysis.

3.7 Choice of ϕ

The choice of the hypothesized value ϕ_0 is a critical decision that needs to be made before the data is analyzed. A higher choice of ϕ_0 will lead to a decrease in power. Note that the simulation studies in Section 3.4 all used correlation matrices that were uniform over the set of all positive definite correlation matrices for a given dimension and if this assumption is known to be untrue it could change the power and type I error. In general, we advocate for using a value of ϕ_0 of 0.5. When controlling the Type I error rate at $\alpha = 0.05$ this will still achieve decent power for a small number of measures. With $\phi = 0.8$, our simulations suggest 80% power can be achieved with only 16 measures that have an effect size of 1. This means that regardless of sample size if a researcher measured 16 different things on even a single experimental unit, such as a mouse, and the researcher's ability to predict what will happen is about 80% then they will reject the null hypothesis of $\phi_0 = 0.50$ 80% of the time. This would imply that their understanding of the theory appears to be correct and warrant investigation with much larger sample sizes and more targeted hypotheses. We also note that for $\phi_0 = 0.50$ the accuracy of the Normal approximation is maximized, which will be of interest for studies collecting more than 30 measurements on each experimental unit. For some combinations of m and ϕ_0 it is impossible to reject the null hypothesize due to discreteness, Table 3.3 provides the minimum m that can be chosen for a given ϕ_0 such that the null hypothesis could still be rejected.

Table 3.3: Minimum required m for hypothesized value ϕ_0

ϕ_0	0.50	0.60	0.70	0.80	0.90
Minimum required m	5	6	9	13	29

3.8 Arterial Spin Labeling example

A data set consisting of measures of Arterial Spin Labeling (ASL) in different regions of the brain will be used to demonstrate our hypothesis test. This data comes from Dr. Vidoni at the University

of Kansas Medical Center. The data was collected as part of a larger study examining the relationship between exercise and Alzheimer's disease that measured pre-post changes after a 12-week exercise intervention in 14 older adults. The data we present here consists of ASL measures on six regions of the brain: BA46, Frontal mid, Hippocampus, M1, Superior Parietal and Precuneus, thus $m = 6$. The research hypothesis is that ASL will increase in the six regions of the brain following the intervention. No single region is of greater interest, and we take the observed result across these different regions as a proxy for the more general research hypothesis concerning structural/-functional changes in the brain following the intervention. The researcher provided predictions were that blood flow would increase in every region. For this data we observed the following pre-post sample means [-1.13, 0.77, -1.19, 1.69, 0.97, 0.70] for BA46, Frontal mid, Hippocampus, M1, Superior Parietal and Precuneus respectively. Thus four of the 6 regions did actually increase meaning the predictions were correct on only two thirds of the measures. The results of the predictions are therefore $\mathbf{p} = [0,1,0,1,1,1]^T$. For our data set we calculated the differences in ASL in the six regions before and after the intervention and calculated the following sample correlation matrix of those differences.

We observed the following weights $w_{\text{BA46}} = 0.411$, $w_{\text{FrontalMid}} = 0.408$, $w_{\text{Hippocampus}} = 0.623$, $w_{\text{M1}} = 0.388$, $w_{\text{SuperiorParietal}} = 0.389$ and $w_{\text{Precuneus}} = 0.403$, with $W = \sum_{i=1}^{6} w_i = 2.62$. Our observed test statistic is $t_m = 0 \cdot (0.411) + 1 \cdot (0.408) + 0 \cdot (0.623) + 1 \cdot (0.388) + 1 \cdot (0.389) + 1 \cdot (0.403) = 1.589$.

The weight given to the ASL measure in the Hippocampus is about 150% of all the other measures, indicating that ASL measure in the Hippocampus was more "independent" of the other regions, thus the result of the prediction in that region is given more weight.

There are 2^6 possible combinations of predictions, thus our test statistic can take on 64 different values. The null hypothesis for this study was set at $\phi_0 = 0.50$. The probabilities of correctly predicting {0,1,2,3,4,5,6 } responses are (0.015625, 0.09375, 0.234375, 0.3125, 234375, 0.09375, 0.015625), these are simply binomial probabilities with success parameter 0.50 and $n = 6$. There are (1,6,15,20,15,6,1) combinations respectively of getting the {0,1,2,3,4,5, 6} correct predictions.

Thus, the probability of any single test statistic value depends on the number of correct predictions shown below.

$$\left[\frac{0.015625}{1}, \frac{0.09375}{6}, \frac{0.234375}{15}, \frac{0.3125}{20}, \frac{0.234375}{15}, \frac{0.09375}{6}, \frac{0.015625}{1}\right]$$

$$= \left[0.015625, 0.015625, 0.015625, 0.015625, 0.015625, 0.015625\right]$$

There are 15 ways to have 4 correct predictions out of the 6 total, each one of those sets of four correct predictions will result in a different value of the test statistic, but each one has the same probability of occurring under the null, 0.015625.

In the case where $\phi_0 = 0.5$ the probability of any single test statistic value will be the same as any other value, however this is not the case when $\phi_0 \neq 0.5$. Given the probabilities of all possible test statistic values, we can then calculate the distribution of the test statistic exactly since we can enumerate the probability of every possible value and could calculate every value.

For our observed $t_m = 1.589$ with $\phi_0 = 0.50$ the probability of getting a test statistic as or more extreme under the null, i.e. the p-value, is 0.34, thus we would fail to reject the null hypothesis at the traditional $\alpha = 0.05$ level in favor of the alternative that the research's true predictive ability is ≥ 0.50. Based on this information we would feel less confident about enrolling a larger sample size to look at the interactions between the intervention and structural changes in the brain, i.e. we have little evidence to suggest enrolling a larger sample will show results in the direction the research hypothesis supposes in relation to the ASL data, a "no-go" decision. We note that if t-tests had been carried out on all 6 measures, and one specified as primary, with the others as secondary, it would not matter which one was designated as primary because the p-value for all six tests is > 0.05. Additional analysis of this data set with two hypothetical prediction vectors is presented in the supplemental section.

3.9 Discussion

We have described a new statistical test for a global hypothesis for many measures. Our technique makes use of researcher's predictions and is essentially a test of whether a researcher's understanding of the natural process, i.e. their ability to correctly predict the outcomes, is convincing enough to continue on to larger studies. This test can answer an essential question in early biomedical research, does the researcher understand enough about the natural process to continue supporting that research, or should it be abandoned in favor of a more promising research track. Our test has good power and type I error properties under our simulated scenarios and has a normal approximation for when the exact distribution is difficult to calculate. We believe our testing procedure will be widely useful especially but not limited to exploratory and early stage studies.

3.9.1 Limitations and Future work

For small sample sizes it has been shown that sample correlations can be unstable (Schonbrodt & Perugini, 2013) (Bonnett & Wright, 2000). This is the main concern with treating the weights as fixed since for large sample sizes the sample pairwise correlations and thus the weights will be close to their true values. The literature on the topic has been concerned with the stability of estimates for hypothesis testing while we are interested in the point estimate; nevertheless, for small sample sizes the instability of the pairwise sample correlations could lead to our weights being far from the true value and our test might over or under weight various endpoints. While this is a concern, we note that our test came to the same conclusions over 96% of the time in all scenarios we examined in Section 6. However, for very small sample sizes, and where available we recommend eliciting estimates of the pairwise correlations either from the researcher or from available literature which can then be used as a sensitivity analysis to the sample correlation matrix.

In addition, the current methodology is limited to one-sided predictions, continuous variables and paired data, extensions to these are an avenue for future work.

We believe that not only does our method provide researchers with an extremely useful ap-

proach for early stage research but that it can be greatly extended beyond what we have shown here. Our current work and plans for future work include: extending the methodology beyond one sided predictions to encompass two sided predictions, allowing different prediction probabilities for different endpoints, thereby assigning more importance to different endpoints, extending our method to situations with both discrete and continuous data, and the development of a user friendly tool for researchers to be able to upload their data sets, make predictions and get relevant results and graphs using our methodology.

3.10 Supplemental material for Chapter 2

3.10.1 Asymptotic Normal Approximation

The Lindberg Condition states (Billingsley, 1995) that for a set of independent but not necessarily identically distributed random variables X_i with expected values μ_i and variances σ_i^2 where we let $s_m^2 = \sum_{i=1}^m \sigma_i^2$ that

$$\sum_{i=1}^{m} \frac{X_i - \mu_i}{\sqrt{\sum_{i=1}^m \sigma_i^2}} \xrightarrow{d} N(0,1)$$

as long as the following condition holds for all $\varepsilon > 0$

$$\lim_{m\to\infty} \frac{1}{s_m^2} \sum_{i=1}^{m} E\left[(X_i - \mu_i)^2 \cdot 1_{|X_i-\mu_i|>\varepsilon\cdot s_m}\right] = 0$$

We show that as long as ϕ_0 is bounded away from 0 and 1 this condition will hold. We consider the approximation for $\phi_0 \in (0,1)$ and we do not consider the cases where $C = I$ or $C = J_m$ since we have shown that in both cases the test statistic follows a binomial distribution and thus has a well known normal approximation. We also make the assumption that as $m \to \infty$ the number of unique $w_i's$, the weights, also goes to infinity. By assuming this we restrict this approximation to settings where the measure of interest are not all equal, or for instance where all but 1, or 2 are all equal. We assume here that the number of unique measures grows.

Let $X_i = p_i w_i$ then under the null hypothesis and treating the weights as fixed leads to $E[X_i] = \phi_0 w_i$ and $\sigma^2_{X_i} = \phi_0(1-\phi_0)w_i^2$ and $s_n = \sqrt{\phi_0(1-\phi_0)\sum_{i=1}^m w_i^2}$ gives the following formulation of the Lindberg condition

$$\lim_{m\to\infty} \frac{1}{\phi_0(1-\phi_0)\sum_{i=1}^m w_i^2} \sum_{i=1}^m E\left[(p_i w_i - \phi_0 w_i)^2 \cdot 1_{|p_i w_i - \phi_0 w_i| > \varepsilon \cdot \sqrt{\phi_0(1-\phi_0)\sum_{i=1}^m w_i^2}}\right] = 0$$

The LHS of the limit $\left(\frac{1}{\phi_0(1-\phi_0)\sum_{i=1}^m w_i^2}\right)$ will always go to 0 as $m \to \infty$ because since $w_i > 0$ for all i and we assume the weights are not all identical the sum of the squared weights will go to ∞ as $m \to \infty$.

Thus we focus on the expectation and specifically on the indicator function. We show that for given restriction of ϕ_0 the indicator will always be 0, thus the expectation will always be 0 and thus the limit as $m \to \infty$ equals 0, satisfying the condition and we show that these restrictions on ϕ_0 will always be satisfied as $m \to \infty$.

There are two scenarios to consider the indicator function under, when $p_i = 0$ and when $p_i = 1$.

For all i such that $\mathbf{p_i = 0}$:

Note that w_i and ϕ_0 are always > 0. The indicator function depends on

$$\begin{aligned}
|-\phi_0 w_i| &> \varepsilon\sqrt{\phi_0(1-\phi_0)\sum_{i=1}^m w_i^2} && (3.1)\\
= \phi_0^2 w_i^2 &> \varepsilon^2\phi_0(1-\phi_0)\sum_{i=1}^m w_i^2 && (3.2)\\
= \phi_0 w_i^2 &> \varepsilon^2(1-\phi_0)\sum_{i=1}^m w_i^2 && (3.3)\\
= \phi_0 w_i^2 &> \varepsilon^2\sum_{i=1}^m w_i^2 - \phi_0\varepsilon^2\sum_{i=1}^m w_i^2 && (3.4)\\
-\phi_0(w_i^2 + \varepsilon^2\sum_{i=1}^m w_i^2) &> \varepsilon^2\sum_{i=1}^m w_i^2 && (3.5)\\
& && (3.6)\\
\Rightarrow \phi_0 &> \frac{\varepsilon^2\sum_{i=1}^m w_i^2}{(w_i^2 + \varepsilon^2\sum_{i=1}^m w_i^2)} && (3.7)
\end{aligned}$$

Thus when $p_i = 0$ the indicator function will be 1 for all i if the above inequality holds, thus the indicator function will be 0 for all i such that $p_i = 1$ if the following inequality holds.

$$\phi_0 < \frac{\varepsilon^2 \sum_{i=1}^{m} w_i^2}{(w_i^2 + \varepsilon^2 \sum_{i=1}^{m} w_i^2)}$$

For all i such that $\mathbf{p_i = 1}$:

Note that w_i and ϕ_0 are always > 0. The indicator function depends on

$$| w_i - \phi_0 w_i | > \varepsilon \sqrt{\phi_0(1-\phi_0) \sum_{i=1}^{m} w_i^2} \tag{3.8}$$

$$| w_i(1-\phi_0) | > \varepsilon \sqrt{\phi_0(1-\phi_0) \sum_{i=1}^{m} w_i^2} \tag{3.9}$$

$$w_i^2(1-\phi_0)^2 > \varepsilon^2 \phi_0(1-\phi_0) \sum_{i=1}^{m} w_i^2 \tag{3.10}$$

$$w_i^2(1-\phi_0) > \varepsilon^2 \phi_0 \sum_{i=1}^{m} w_i^2 \tag{3.11}$$

$$w_i^2 - w_i^2 \phi_0 > \varepsilon^2 \phi_0 \sum_{i=1}^{m} w_i^2 \tag{3.12}$$

$$w_i^2 > (\varepsilon^2 \sum_{i=1}^{m} w_i^2 + w_i^2)\phi_0 \tag{3.13}$$

$$\frac{w_i^2}{\varepsilon^2 \sum_{i=1}^{m} w_i^2 + w_i^2} > \phi_0 \tag{3.14}$$

Thus when $p_i = 1$ the indicator function will be 1 for all i if the above inequality holds, therefore the indicator function will equal 0 for all i such that $p_i = 1$ if the following inequality holds.

$$\frac{w_i^2}{\varepsilon^2 \sum_{i=1}^{m} w_i^2 + w_i^2} < \phi_0$$

These two scenarios, that for a given i $p_i = 0$ or $p_i = 1$ cover all possible outcomes. When we combine these two restrictions on ϕ_0 they give an upper and lower bound for ϕ_0 such that the

indicator function will equal 0 if for all i ϕ_0 is within

$$\left(\frac{w_i^2}{\varepsilon^2 \sum_{i=1}^m w_i^2 + w_i^2}, \frac{\varepsilon^2 \sum_{i=1}^m w_i^2}{(w_i^2 + \varepsilon^2 \sum_{i=1}^m w_i^2} \right)$$

If this holds then the indicator function will always be 0 and the limit will equal 0, thus the Lindberg condition will hold. We note that as $m \to \infty$ the summation $\sum_{i=1}^m w_i^2$ will tend to infinity so long that as m increases the number of unique weights also increases (i.e. the increase in m is not solely driven by adding copies of the same measure). Thus as the lower bound will tend to 0, and the upper bound will tend to 1, so the restriction on ϕ_0 tends to (0,1) as $m \to \infty$. Thus asymptotically the normal approximation will hold so long as ϕ_0 is not set equal to 0, or 1.

3.10.2 Additional analysis for the example data

The ASL example presented in the paper resulted in a failure to reject the null hypothesis. Here we present an analysis using two hypothesized prediction vectors. Let's compare the results of two different prediction vectors $\mathbf{p_1} = [1,1,0,1,1,1]$ and $\mathbf{p_2} = [1,1,1,0,1,1]$, the only difference being that in $\mathbf{p_1}$ the researcher incorrectly predicts the third measure while in $\mathbf{p_2}$ the researcher incorrectly predicts the fourth measure, otherwise they both consist of 5 correct predictions on the 6 total measures.

3.10.3 Analysis for $\mathbf{p_1}$

Here the observed test staistic would be

$$t_m = 1 \cdot (0.411) + 1 \cdot (0.408) + 0 \cdot (0.623) + 1 \cdot (0.388) + 1 \cdot (0.389) + 1 \cdot (0.403) \approx 2$$

This would results in a p-value of 0.11, in which case we would still fail to reject the null hypothesis.

3.10.4 Analysis for $\mathbf{p_2}$

Consider the result for $\mathbf{p_2}$, here we have a test statistic of

$$t_m = 1 \cdot (0.411) + 1 \cdot (0.408) + 1 \cdot (0.623) + 0 \cdot (0.388) + 1 \cdot (0.389) + 1 \cdot (0.403) = 2.234$$

which results in a p-value of 0.03125, in which case we would reject the null hypothesis of $\phi_0 = 0.05$ in favor of $\phi_0 \geq 0.50$.

3.10.5 Difference between $\mathbf{p_1}$ and $\mathbf{p_2}$

The reason these two prediction vectors come to such different conclusion even with the same number of correct predictions has to do with what was correctly predicted and what was not. In $\mathbf{p_1}$, the one incorrect prediction concerned the ASl region with the highest weight, the Hippocampus which had a weight almost double any of the other measures, similarly the real predictions were also incorrect on this region. This is the region our test considers the most unique and thus correct predictions here are of more weight. Conversely in $\mathbf{p_1}$ the direction in the Hippocampus was correctly predicted while the incorrect prediction was for the M1, the region with the least weight. We can see that depending on which measures are correctly predicted we can either reject or fail to reject the null even with the same number of correct predictions and in this case we can see exactly why. There are $2^6 = 64$ possible outcomes of the test statistic, the top 4 highest possible test statistics would all lead to a p-value > 0.05, there is one way to correctly predict all 6 measures and 6 ways to correctly predict 5 of the 6 measures. Thus, if a researcher correctly predicts all measures or the 3 highest test statistics coming from predicting all but one of the measures we will reject the null hypothesis.

Chapter 4

Comparing the prediction test to other methods via an application to brain imaging data

Abstract

We discuss a new hypothesis test intended for early stage research with small sample sizes and many endpoints called the Prediction Test. The test allows for a go/no-go decision concerning further study of a research hypothesis. The Prediction test is most powerful with a large number of variables of interest. We provide extensions to the types of predictions that can be made and discuss estimating power. In addition, we compare the Prediction test to other typical approaches to paired data with multiple outcome variables, specifically we compare it to a set of t-tests and a linear mixed model with data coming from a recent study of diffusion tensor imaging.

4.1 Introduction

We have developed a hypothesis test intended for use in early stage research. We see our method as being most applicable when sample sizes are small and the number of measures (i.e. variables) observed on each experimental unit is large. Our test allows researchers to make a "go/no-go" decision regarding their research hypothesis. Essentially, we provide a formal way to test whether a hypothesis is being borne out in a small study, if it is then there is evidence to warrant a larger study, if it is not then a reassessment of the research hypothesis is in order. Our test can be applied to any study on which the principal investigator (PI) can make a prediction about the results of the measures based on their hypothesis. These predictions can be one-sided (e.g. predicting that a measure will increase after an intervention) or two-sided (e.g. predicting that a measure will be different, though not necessarily "statistically significantly different" from some value), and

in general the prediction can be of any type so long as the probability of making a successful prediction given that the prediction is actually wrong can be specified. We believe our method has many advantages over other methods of analysis that deal with multiplicity, especially when the sample size is small.

A common approach to dealing with the issue of multiplicity is to designate a primary endpoint and relegate all other endpoints to secondary status, or sometimes co-primary status for certain situations (Dmitrienko et al., 2009). Ideally the primary endpoints are chosen as the endpoints most clinically relevant. Secondary endpoints are then also measured on each experimental unit but are only used in an exploratory sense and even then interpretations of them may not be valid unless the primary endpoint is "statistically significant" (O'Neill, 1997). Secondary endpoints are not typically considered when controlling the Family Wise Error Rate (FWER) or doing power calculations. A study using this approach depends on the outcome of the primary endpoint(s), with a designated "win" criteria, often "statistical significance" of the primary endpoint, while the secondary endpoints are used to help interpret the primary endpoint or as impetus for future work. This technique allows research to be done, even with small sample sizes, on a large number of endpoints but formal inference can only be concluded on one.

Unfortunately, in practice the use of primary and secondary endpoints can be problematic, and this approach is not always appropriate given the underlying data and research hypothesis. It has been well noted in the literature that secondary endpoints are often misinterpreted. That is, given unimpressive results from a primary endpoint, but interesting (i.e. "significant") results in a secondary endpoint the primary endpoint can become "unceremoniously unseated when discovered to be negative at the trial's conclusion ... receiv[ing] little attention in the end" (O'Neill, 1999). There is also cause for concern when there is not a true "primary" endpoint. In studies with multiple endpoints that are of equal clinical relevance, or perhaps are all proxies for some true endpoint of interest that cannot be directly measured then the choice of primary endpoint can be arbitrary and the temptation to over-interpret secondary endpoints, since they are not true secondary endpoints, is even greater. In this scenario if the "win" criteria is that the primary endpoint is "statistically

significant" then an arbitrary choice of that endpoint can easily be the determining factor in the success or failure of a study.

Our hypothesis test is a global test that allows for a single go/no-go decision at the end of the study. Our test is not intended for confirmatory studies or trials, but is focused on early stage research, with small sample sizes, many endpoint and lots of variability in determining the worth of a given research hypothesis. In section 4.2 we present our hypothesis test and discuss its properties including extensions to previous work, section 4.3 provides an overview of properties of our test, section 4.4 illustrates the method with an example on a real data set as well as a comparison to other methods, section 4.5 reports the results of a simulated comparison between our method and a set of t-test, and section 4.6 includes discussion of best practices, conclusions and limitations of our method.

4.2 Hypothesis test

Our null hypothesis concerns the principal investigator's (PI's) predictive ability. The predictions that can be made on the endpoints can be one sided (e.g. predicting the mean response will increase from baseline) or two sided (predicting a change from baseline, or pre-post, or relative to some standard even if that change is not "significant"). In general, the predictions can be of any form such that when we assume the prediction is incorrect, similar to the assumption of a null hypothesis being true, we can specify the probability of getting a correct prediction. The testing procedure is currently limited to paired data due to the need to calculate a sample correlation matrix. Our test makes use of the fact that if the research hypothesis that the study seeks to address is correct then this should provide the PI with a better idea of what will happen on each of the endpoints than random chance alone would. In this case we would have evidence that the research hypothesis is correct if the PI can correctly predict many of the endpoints, while we would have little evidence it was correct if the number of endpoints correctly predicted was relatively small or unimpressive, such as around 50%.

We define $\phi \in [0,1]$ to be a parameter that measures a researcher's ability to predict what will happen on various endpoints of an experiment. The parameter ϕ can be thought of as a binomial success parameter, where a large value of ϕ means we would see more correct predictions on average, while a small ϕ would lead to fewer correct predictions. We let ϕ_0 be the hypothesized value, typically chosen to be $\phi_0 = 0.5$, this would correspond to the researcher having no more understanding about what would happen on the endpoints than would be expected by flipping a fair coin. Our alternative hypotheses is that the researcher's true ability is greater than the hypothesized one, that is $\phi > \phi_0$, more formally the null and alternative hypotheses for our test are:

$$H_0 : \phi \leq \phi_0$$

$$H_1 : \phi > \phi_0$$

If we reject this null hypothesis in favor of the alternative we would say that we have evidence to suggest that $\phi > \phi_0$. For $\phi_0 = 0.5$ this would mean we have evidence that the researcher is able to predict what will happen on different endpoints at a level greater than if they simply guessed on each endpoint, note that for each endpoint there are only two possible outcomes such as an increase vs no increase, or a difference vs no difference. In this case it would seem likely that the research hypothesis, or theory about the underlying natural process would be what was giving the researcher this advantage thus we would consider this hypothesis suitable for further research since the research hypothesis is being borne out, a "go" decision. Other choices of ϕ_0 would lead to similar interpretations, for instance rejecting the null hypothesis of $\phi_0 = 0.70$ would imply that the researcher is able to predict the result of what will happen on more than 70% of the measures. The test statistic is given by

$$T_m = \sum_{i=1}^{m} p_i w_i$$

where m represents the number of endpoints, or variables that are being measures on each experimental unit, p_i is the result of the researcher's prediction on the i^{th} variable (1 for a correct prediction, 0 for an incorrect prediction) and $w_i = (\sum_{j=1}^{m} r_{ij}^2)^{-1}$ is a weight calculated from the

correlation matrix between all endpoints, where r_{ij} is the sample pairwise correlation between endpoint i and j. This weight will give higher values to endpoints that are more "independent", meaning less correlated, from the other endpoints and lesser value to endpoints that are more highly correlated. In this way correctly predicting many endpoints that are highly correlated provides little advantage, for instance if a researcher correctly predicted the results of an exercise intervention on the weight and mass of a set of mice we would down weight these predictions so that the two correct predictions were worth little more than 1 (since mass is a linear function of weight only one unique measure has been correctly predicted) (Montgomery & Mahnken, 2019).

4.2.1 Distribution under the null

Our test statistic has an exact distribution, given the number of measures of interest, m, the number of possible outcomes of the test statistic is simply the possible combinations of predictions. Since every prediction is either correct or incorrect (2 options) there are 2^m total outcomes. Given the observed weights every test statistic value can be enumerated, and exact p-values and other statistics can be calculated. For large m, typically greater than 20 to 30, the exact distribution becomes computationally expensive. In these cases, we can use a Normal approximation. We have shown that

$$T_m \sim N\left(\phi_0 \cdot \sum_{i=1}^{m} w_i, \sqrt{\phi_0(1-\phi_0) \cdot \sum_{i=1}^{m} w_i^2}\right)$$

the average error for this approximation, that is how far the approximated cdf is from the exact cdf has been shown empirically to be small when ϕ_0 is not too close to 0 or 1 (Montgomery & Mahnken, 2019).The test can be conducted using either the exact or approximate distribution, when feasible we recommend the exact distribution.

4.2.2 Types of Predictions

In our previous work we restricted our analyses to one sided predictions, that is we required researchers to predict whether an endpoint would increase or decrease. That is would the value of

an endpoint after some intervention be larger than at baseline, indicating an increase, or smaller than at baseline, indicating a decrease. However, the prediction test is more general and can deal with any type of prediction so long as it can be either categorized as a success or failure at the end of a study and that the probability of success on any single endpoint can be specified. This leads to great flexibility in designing a study since the actual predictions based on the research hypothesis, be it an increase, decrease, difference or something else can be included. Here we discuss directional predictions and how to make a prediction of a difference between endpoints. Based on the need to calculate a sample correlation matrix we assume all the data is paired. That is we have either pre-post measures or approximate pairings between experimental units controlling for different demographics.

4.2.2.1 Directional prediction

A directional prediction is made if the PI believes a measure will increase or decrease, typically this change would be relative to baseline values. As an example we will consider the implications of a positive prediction, a negative prediction would follow directly. If the prediction for an endpoint is positive, that means the PI believes that given some intervention the resulting value will increase. We don't know the actual result, whether or not there is an increase, but by specifying the null hypothesis ϕ_0 we designate how often we expect the researcher to designate the true condition of the measure (an increase). If $\bar{x}_i > 0$ then we would conclude that the prediction of an increase was correct and set $p_i = 1$ for the i^{th} measure. The prediction will be incorrect if the sample mean is on the opposite side of 0 as the prediction, i.e. it's negative for a positive prediction or positive for a negative prediction. An error can occur even if the prediction is correct, that is the true mean μ is on the same side as the prediction but the sample mean $\bar{x}$ is not, which can occur for small samples. When the true value of ϕ is larger than the hypothesized ϕ_0 the researcher will make more correct predictions than would be assumed under the null, and given enough we would reject the null in favor of the alternative.

4.2.2.2 Prediction of a difference

If the prediction on a single measure was that the result of treatment, intervention etc. was non-zero, that is it had some effect on the response then the prediction would be of a difference. Due to random chance the difference before and after some intervention will almost certainly not be exactly zero, so we need to determine how to quantify how far away from 0 the difference needs to be for a prediction of a difference to be deemed successful.

We can take a cue from traditional hypothesis testing where we assume the null hypothesis, the opposite of what we are hoping is true. If our prediction is that there is a difference between two measures one way to define thresholds such that a prediction result can be decided is to assume that the prediction is incorrect, that is there is no difference between the two groups. If the prediction is incorrect we want the researcher to make a "correct' prediction of a difference at less than or equal to $\phi_0 \cdot 100\%$ of the time. If we assume the data is approximately normal, an assumption that may be violated in practice, we can easily define appropriate thresholds. As an example consider the following hypotheses $H_0 : \phi \leq 0.50$ vs $H_1 : \phi > 0.50$ and assume that for a given measure the researcher predicted a difference between two groups. If the values of the difference are approximately normal with mean 0, that is assuming the prediction is incorrect, then there is a 50% chance of observing a value beyond approximately 0.68 standard deviations. For the sample data, a threshold of $0.68 \cdot \frac{s}{\sqrt{n}}$ where s is the sample standard deviation and n is the sample size, will provide thresholds such that 50% of the time the prediction is incorrect the sample mean will be beyond those thresholds. We divide the sample standard deviation by the square root of the sample size because we know that for the sampling distribution of the sample mean $\sigma_{\bar{x}} = \frac{\sigma}{n}$.

If the researcher's true predictive ability is greater than 0.50 then they will correctly predict a difference on greater than 50% of the endpoints that a difference is predicted on. Table 4.1 provides thresholds for various hypothesized values of ϕ_0.

These thresholds will work as expected, that is a correct prediction designation will be made ϕ_0 percent of the time when the prediction is actually false, and a correct prediction designation will be made $> \phi_0$ percent of the time when the prediction is actually true. However, these thresholds

ϕ_0	0.50	0.60	0.70	0.80	0.9
Thresholds	$\pm 0.68 \cdot \frac{s}{n}$	$\pm 0.53 \cdot \frac{s}{n}$	$\pm 0.39 \cdot \frac{s}{n}$	$\pm 0.26 \cdot \frac{s}{n}$	$\pm 0.13 \cdot \frac{s}{n}$

Table 4.1: Recommended threshold for a correct prediction given ϕ_0, s = sample standard deviation, and n = sample size

assume normality of the underlying data which may not be true. If possible, the use of clinically relevant, but not necessarily "statistically significant" thresholds is recommended. In addition it is valid to use a threshold from table 4.1 that is to the left of the recommended threshold for a given ϕ_0, for instance with $\phi_0 = 0.80$ a valid, and conservative, threshold would be $\pm 0.39 \cdot \frac{s}{n}$.

4.3 Properties of our test

In our previous work (Montgomery & Mahnken, 2019) we have shown that the prediction test has good type I error control and competitive power, especially in the adverse situation of small sample sizes and many endpoints. In all of our simulated scenarios the type I error was controlled at the nominal 0.05 level. In addition we have shown that m increase, that is as the number of endpoints being measured increases, the power will increase for a fixed sample size. For example given a sample size of 20, and 10 measures relevant to the research hypothesis, the prediction test will be more powerful if the study had been conducted with additional relevant measures also included. The caveat to this is that the endpoints need to deal the same research hypothesis for the test to provide a valid go/no-go decision about that research hypothesis.

4.3.1 Power calculations

The power of our test depends on the correlation matrix between all endpoints and so cannot be calculated before the data has been collected. However empirical power calculations carried out via simulation can provide good estimates of power. Values need to be chosen for ϕ_0 and ϕ, along with the number of experimental units n and the number of endpoints measured, m. To simulate power estimates three of the four parameters can be fixed, for instance if funding limits $n = 10$,

there are $m = 20$ endpoints of interest and $\phi_0 = 0.5$ then we would be allowing ϕ to vary in order to achieve a given power. It is important to note that unlike traditional testing methods power will not generally increase as the sample size increases unless the effects are on each measure are small; however, power will increase as the number of endpoints increases, we would suggest letting either ϕ_0 or m vary.

The choice of ϕ_0 will depend on the goal of the study. The Prediction test is intended to provide a go/no-go decision on the research hypothesis, rejecting the null of $\phi_0 = 0.50$ would imply that the researcher's theory is providing an advantage in predicting the results over what chance alone would provide and is an appropriate choice for many situations. If ϕ is also fixed and m allowed to vary then an effect size can determined. For ϕ and ϕ_0 this effect size can be calculated using Cohen's h (Cohen, 1988) which for the Prediction test is defined as

$$h = 2 \cdot \arcsin(\sqrt{\phi}) - 2 \cdot \arcsin(\sqrt{\phi_0})$$

rules of thumb exist such that h = 0.2 is a "small effect", h= 0.50 is a "medium effect", and h= 0.80 is "large effect"; however as with all rules of thumb an effect size made specific to the study at hand is better than an arbitrary rule of thumb. For ϕ_0 = 0.50 and $\phi = 0.80$ the Cohen's h = 0.64 which is classified as between a "medium" and "large" effect.

Once the parameters are fixed then a set of simulated correlation matrices can be generated. For instance the R package clusterGeneration (Qui & Joe, 2015) allows for easy generation of valid, that is positive semi-definite, covariance matrices which can be converted to correlation matrices in R or other software.

The following is an example of how to estimate power. Depending on certain study values, goals and expectations about parameters the number of simulated correlation matrices (Step 1) and number of estimates on each correlation matrix (Step 3) may need to be changed. We present an example where ϕ is free to vary.

1. Choose a set of potential ϕ values, e.g. ($\boldsymbol{\phi}$ = [0.6, 0.7, 0.8, 0.9])

2. For i =1: $\boldsymbol{\phi}_i = 0.6$:

   - Generate 100 correlation matrices, C_j; j = 1, ..., 100.
   - For each of the 100 correlation matrices (i.e. for (j in 1:100)):
     - Calculate the the weights for the correlation matrix C_j by squaring the entire matrix, summing every row and taking the inverse of the row sums. This will result in a vector $\mathbf{w_j}$ of weights.
     - Simulate 100 prediction results vectors, $r_{j,k}$, k = $1, \ldots, 100$. (Simulate draws from Binomial(m, $\phi = \boldsymbol{\phi}_1 = 0.6$) for the j^{th} correlation matrix)
     - Calculate the observed test statistic for the j^{th} correlation matrix and the k^{th} prediction vector for that correlation matrix as $r_{j,k} \cdot w_j$.
     - Calculate the prediction test given C_j, w_j and $r_{j,k}$, and the decision (p-value less than pre-specified α: d = 1, or p-value $> \alpha$: d = 0). Record the value of d.
   - Repeat the above steps for all values of i
   - For each value of $\boldsymbol{\phi}_i$ calculate the sum of all the values of d divided by 10,000 (100 correlation matrices multiplied by 100 prediction vectors for each matrix). This is the empirical power estimate.

A post-hoc power analysis is shown in the following section for the example data set.

4.4 DTI Analysis

We have a data set consisting of diffusion tensor imaging (DTI) measures across 12 regions of the brain. For each region we collected both Mean Diffusivity (MD) and Fractional Anisotropy (FA), resulting in 24 different measurements of interest. The 12 tracts that MD and FA were measured

on are: ATR, CG, CH, CST, FMAJ, FMIN, IFOF, ILF, SLF, SLFT, UF and All tracts which is a combination of the other 11 tracts.

DTI measures the diffusion of water molecules along different tracts (O'Donnell & Westin, 2011). It can be used to measure structural changes in the brain due to the fact that pathologic processes can change the way that water is diffused in tissue (Alexander et al., 2007). MD measures the mean diffusivity while FA is very sensitive to any changes in diffusivity, the two measures are often used together in studies. It is sometimes assumed that MD will be higher in the presence of damaged tissue, while FA will be lower (Soares et al., 2013), thus if the disease of interest was mitigated in some way we would expect MD to decrease and FA to increase.

The intervention in this study was Kidney transplantation; we had paired pre and post data on n = 22 subjects with End Stage Renal Disease (ESRD). The research hypothesis of interest is that kidney function and thus kidney transplantation causes structural changes in the brain that may be related to Alzheimer's disease and/or cognition. As an early step toward understanding the relationship between healthy kidney function and Alzheimer's disease we first seek to understand the structural changes in the brain that are caused by kidney transplantation. By using paired pre post data we can compare the brain structure before transplantation, when kidney function was poor, and after, when kidney function was improved.

Prior to receiving the data set we discussed the directional change the principal investigator expected for the pre-post DTI measures, the predictions were that FA values would increase across all tracts and MD values would decrease, that is they would normalize. To analyze this data set using the Prediction test we set $\alpha = 0.05$ and $\phi_0 = 0.50$ a rejection of this null hypothesis in favor of the alternative that $\phi > 0.50$ would imply the PI's understanding of the structural changes in the brain as measured by all 24 tracts was greater than we would expect due to chance. All analyses were conducted in R (R Core Team, 2018).

Figure 4.1 shows the observed pre-post changes in DTI for all tracts categorized by DTI measure (FA or MD) where the band inside the box is the sample mean. All FA measures except for SLFT had sample means above 0, and all MD measures had sample means below 0, the exact

sample means are displayed in Table 4.2.

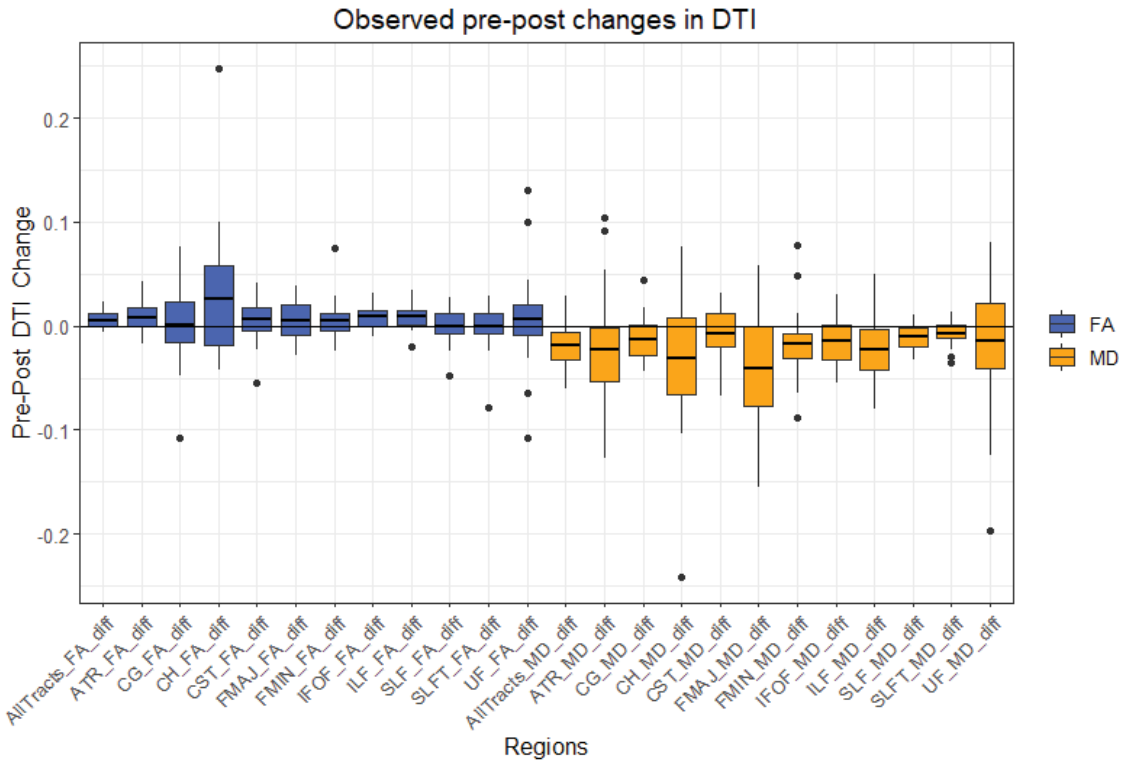


Figure 4.1: Boxplots of observed DTI measures for FA and MD

The calculated weights for each measure are given in Figure 4.2, the tracts that are combinations of the other tracts, All tracts FA and All tracts MD, have the lowest weights as we would expect since they are simply combinations of the other 11 measures for FA and MD respectively. The tract with the highest weight is CG MD, with a weight of 0.59. The weight given to CG MD is more than the weight given to the lowest three measures combined. The reason CG MD is given so much weight is that it is the most "independent" or unique of the measures. For instance, the highest pairwise correlation between CG MD and another tract is 0.28, while All Tracts FA has 14, of 23 pairwise correlations larger than 0.28 and All tracts MD has 13 pairwise combinations greater than 0.28. These regions are given little weight in the test statistic because a correct prediction on them considering correct predictions on other measures which they are highly correlated with is less impressive than a correct prediction on CG MD which is not highly correlated with any other measure. By weighting the endpoints, we insure that correct predictions of many sim-

ilar endpoints do not contribute too greatly to the test statistic since predicting 2 or more highly correlated endpoints is not as impressive as correctly predicting multiple independent endpoints.

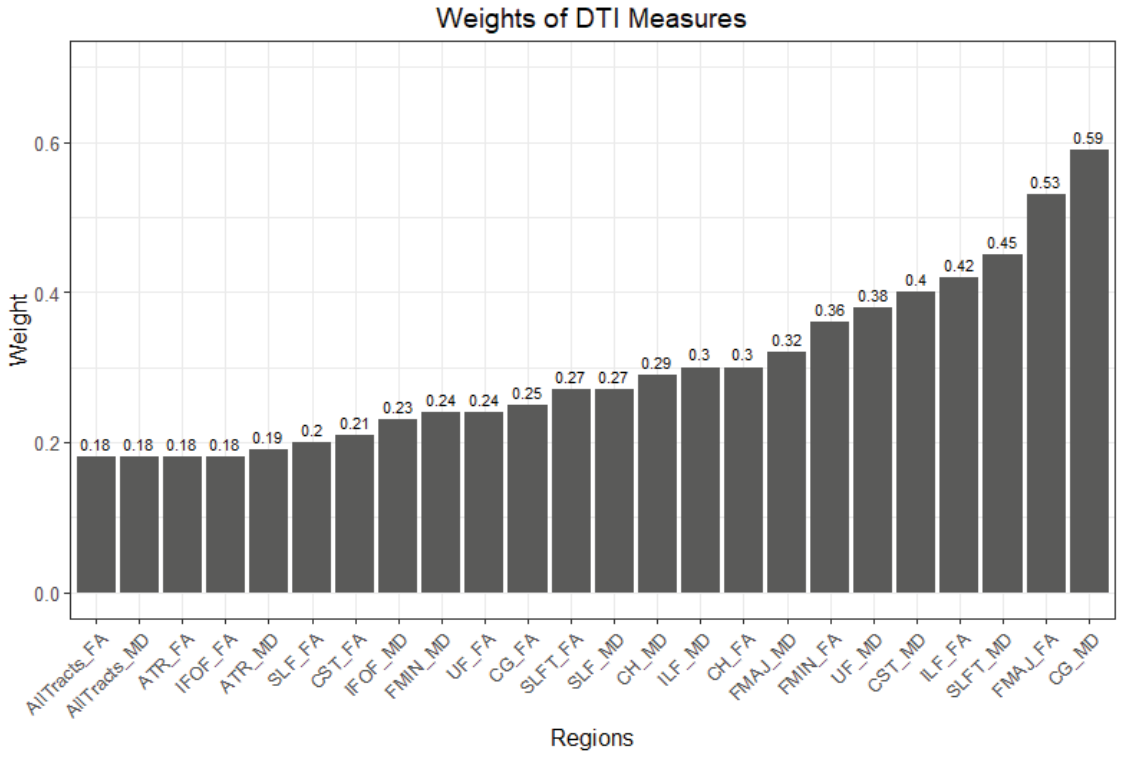


Figure 4.2: Weights for the DTI tracts.

All but one (SLFT FA) of the PI's predictions are correct. The observed test statistic is simply the sum of the weights multiplied by the result of the prediction (1 for correct, 0 for incorrect), $t_m = 6.88$, this is the sum of all the weights in Figure 4.2 except for SLFT FA. For $m = 24$ the exact p-value is difficult to calculate, there are over 16 million possible combinations that need to be computed and ordered, thus we used the Normal approximation, this resulted in a p-value for our test statistic of 0.000011, thus we would reject $H_0 : \phi \leq 0.50$ in favor of $H_1 : \phi > 0.50$. We have evidence that the PI's ability to correctly predict the outcome of different measures informed by the research hypothesis of structural changes in the brain due to kidney transplantation is greater than we would expect if the research hypothesis was false. There is evidence that going from a poorly functioning kidney to a healthy kidney causes structural changes in the brain, thus our resulting decision is a "go". That is the research hypothesis seems promising even with a small sample size

and should be studied further.

4.4.1 Post-hoc power calculation

To illustrate the method of a power analysis for the prediction test we conduct a post-hoc power analysis. We make no interpretations of the results based on this analysis and reiterate that it is simply an example of the process. For the DTI data set, the sample size was n = 22, with m = 24 tracts that would be measured. Before the data was collected we pre-specified $\alpha = 0.05$ and $\phi_0 = 0.50$. Suppose the PI was only interested in rejecting the null hypothesis of $\phi \leq 0.50$ if the effect size was 0.50. Using Cohen's h we can solve and show that for $\phi_0 = 0.50$ we would require $\phi \approx 0.70$ to get h = 0.50. Given $\phi = 0.70$ and the other specified parameters we can use R, or other software, to get an empirical power estimate. For the simulated data we simulated 12 FA measures with means of 0.5, and MD measures with means of -0.5, and standard deviations of 1, the predictions for these simulated data points are all true. This allows us to see how powerful the prediction test would be to detect effects that are one half of the standard deviation. Following the template in Section 4.3.1 we get an empirical power estimate of 0.6318, or approximately 63% power.

Under this simulation all the predictions were assumed to be true, but due to the small simulated means (0.5 and -0.5) the predictions were not correct all the time. Depending on the situation at hand the parameters can be tweaked to better fit the research hypothesis, for instance a certain proportion of the predictions could be simulated as "true", various effect sizes for the measures themselves could be examined, etc. The flexibility of empirical power calculations is one of their main advantages.

4.4.2 Comparison to other methods

The analysis of the DTI data could have been conducted using other methods, we compare our test, and discuss advantages and disadvantages between our test, a set of t-tests and a linear mixed

model.

4.4.2.1 T-tests

T-tests are an obvious candidate for testing hypotheses concerning these endpoints since the endpoints are all continuous and the parameter of interest is the population mean of the pre-post difference. To conduct t-tests on on the 24 different regions we would set the null and alternative hypothesis for each region to $H_0 : \mu = 0$ vs $H_1 : \mu \neq 0$. Due to the large number of tests we would need to do a multiplicity correction; however in situations like this a common approach to dealing with multiplicity is to choose a primary endpoint that will be used for inference and treat all other endpoints as exploratory. Unfortunately, for this particular data set since there is not a true "primary" endpoint since the research hypothesis is about structural changes throughout the brain. Nevertheless, we could choose one of the All tracts measures as the primary endpoint since they are combinations of the other tracts, with $\alpha = 0.05$, and treat all others as secondary, after discussion with the PI All tracts FA was chosen as the primary endpoint. If we conduct a t-test on All tracts FA, we get an observed test statistic value of t = 3.003 and a p-value of 0.006, indicating that we would reject the null hypothesis in favor of the alternative that the mean of the pre-post difference is greater than 0. We treat the response in All tracts FA as a proxy for structural changes in the brain, and thus we have some evidence that structural changes do occur in the brain, and the direction of the change, an increase, implies that normalization is occurring post transplant.

4.4.2.2 Linear Mixed Model

An alternative approach to both the prediction test and a set of t-tests is to use a linear mixed model since patients with ESRD all had pre and post DTI measures on the 24 tracts. Due to the large number of regions and the relatively small sample size we would again pick one model, with All tracts FA as the response, as our primary outcome, and then compute models for the others regions to explore what if anything is driving a pre-post change. The patients all have a pre and post transplant observation; however, the times before and after the transplant are not equal. To

Region	Sample Mean	Weight	Prediction	Result	T-test p-value	Coefficient p-value
All Tracts FA	0.006	0.175	Increase	Correct	0.006*	0.003*
All Tracts MD	-0.02	0.178	Decrease	Correct	< 0.001*	< 0.001*
ATR FA	0.008	0.180	Increase	Correct	0.01*	0.008*
IFOF FA	0.009	0.180	Increase	Correct	0.004*	0.003*
ATR MD	-0.02	0.190	Decrease	Correct	0.07	0.04*
SLF FA	0.00002	0.195	Increase	Correct	0.99	0.67
CST FA	0.006	0.210	Increase	Correct	0.19	0.15
IFOF MD	-0.01	0.233	Decrease	Correct	0.005*	0.003*
FMIN MD	-0.02	0.237	Decrease	Correct	0.03*	0.01*
UF FA	0.006	0.238	Increase	Correct	0.56	0.44
CG FA	0.001	0.249	Increase	Correct	0.87	0.73
SLFT FA	-0.0005	0.268	Increase	**Incorrect**	0.92	0.87
SLF MD	-0.01	0.269	Decrease	Correct	0.002*	< 0.001*
CH MD	-0.03	0.294	Decrease	Correct	0.03*	0.04*
ILF MD	-0.02	0.300	Decrease	Correct	0.002*	0.001*
CH FA	0.026	0.302	Increase	Correct	0.09	0.10
FMAJ MD	-0.04	0.323	Decrease	Correct	0.002*	0.002*
FMIN FA	0.005	0.359	Increase	Correct	0.33	0.17
UF MD	-0.01	0.375	Decrease	Correct	0.30	0.80
CST MD	-0.008	0.401	Decrease	Correct	0.18	0.13
ILF FA	0.009	0.424	Increase	Correct	0.003*	0.002*
SLFT MD	-0.008	0.447	Decrease	Correct	0.02*	0.008*
FMAJ FA	0.006	0.526	Increase	Correct	0.17	0.13
CG MD	-0.01	0.592	Decrease	Correct	0.01*	0.006*

Table 4.2: Results for the predictions, the set of t-tests and the linear mixed models for DTI tracts.

account for the difference in a post transplant observation 3 weeks after transplant compared to one two months after transplant, we let age be the temporal variable instead of a possible arbitrary value such as days before/after transplant. In this way, we also control for the covariate of age. Our covariate of interest was a group variable to distinguish pre vs post. The research hypothesis is that β_{Group} would be positive, that is post transplant patients would have higher All tracts FA scores when accounting for other covariates. By using a linear mixed model we were also able to include clinically relevant covariates: age, gender, race, and education. Fitting the LMM in SAS resulted in $\hat{\beta}_{Group} = 0.006$, with p-value of 0.003, indicating that we should reject the null hypothesis of $\beta_{Group} = 0$ in favor of the alternative that $\beta_{Group} \neq 0$ and given the direction conclude that there is a positive effect of kidney transplant on All tracts FA. The interpretation of this effect is that for all other covariates held constant the post-transplant patient has a mean FA, as measured by All tracts combined, that is 0.006 units higher than a pre-transplant patient.

The prediction test, t-tests and linear mixed model all came to the same conclusion, they rejected their respective null hypotheses and all provide evidence to support the research hypothesis of normalization of DTI measures post kidney transplant. However, the conclusions that can be drawn from rejecting these different null hypotheses are different. For the prediction test, we would reject the hypothesis that the researcher's ability to predict the endpoints was equivalent to a guess, and take this as evidence that their research hypothesis was providing an advantage in predictions and is therefore plausibly correct. This would be encouraging for future research concerning the same research hypothesis, with larger sample sizes and statistical methods that address more targeted hypotheses. With the t-test we were able to reject the null hypothesis that the mean pre-post change for All tracts FA was less than or equal to 0. In this case we treat All tracts FA as the primary measure and the other tracts as exploratory. These tracts, with the exception of SLFT FA, all showed a result in the hypothesized direction and 13 of the 24 were "statistically significant" as shown in Table 4.2, however this is with no multiplicity adjustment. For the set of t-tests if we used another common way to address multiplicity, such as Sidak's adjustment where we could compare each p-value to 0.0021 (Sidak, 1967). The only "significant" results were on All Tracts MD and

SLF MD, indicating that the choice of primary endpoint is extremely important.

The linear mixed allows us to control for other covariates, an advantage when possible, and led to the rejection of the null hypothesis of $\beta_{Group} = 0$, thus we conclude that the effect of transplant, when controlling for other covariates had a positive effect on All tracts FA. Similarly to the t-test the other models with the 23 other tracts as the response indicate showed results in the expected directions, higher for FA and lower for MD, and the coefficients for group were "statistically significant" in 14 of the 24 regions.

4.4.3 Advantages and Disadvantages

We were surprised to see the different methods all perform so well given a relatively small sample size (n=22) and a large number of endpoints (m=24) and note that in general it is unlikely these methods will come to the same conclusions, that is each will be more powerful in certain scenarios. We discuss some of the advantages, disadvantages and assumptions of the respective methods.

4.4.4 Advantages and Disadvantages of the Prediction Test

An advantage of the prediction test is that as the number of endpoint under analysis grows and the sample size remains fixed the power increases. The test can be performed with a sample size as small as n=2, the only concerns for extremely small sample sizes is that a prediction of a measure could be correct, but the sample mean could be on the wrong side of 0, or outside the specified thresholds due to variability in extremely small sample sizes or that the estimated weights could be unstable. The prediction test makes few assumptions about the underlying data, namely that a correlation matrix can be constructed, typically through a paired design, and that the measures themselves are continuous. As a global test, it is ideally suited for addressing research hypotheses that rely upon multiple endpoints and actually uses all of the endpoints in coming to a decision. Essentially the test provides a rigorous way to define a go/no-go decision.

The main disadvantages of the prediction test is that is does not provide an inference about the

research hypothesis of interest, such as structural changes in the brain, instead it makes a claim about whether the hypothesis is worthy of more study instead. The test should not be used in a confirmatory setting, and outside of early research where a proof of concept is sought, it's utility is diminished. For certain studies getting predictions on all the different endpoints, or comparisons between them could be time consuming or infeasible. In addition, the test is currently restricted to paired data due to the need to calculate the sample correlation matrix, although we are exploring ways to use the test for unpaired data.

4.4.5 Advantages and Disadvantages of the T-test

The t-test is fairly robust to departures from it's assumptions, which is especially useful in the settings of small sample sizes where the assumptions can be questionable. The t-test is also easy to implement, easily interpretable, and the interpretation concerns the population mean, whereas the prediction tests interpretation concerns the predictive ability of the researcher.

Some disadvantages of the t-test, and specifically we are considering the t-tests conducted on a set of different endpoints not a single endpoint, include the need to choose a primary endpoint or do some sort of multiplicity adjustment. This can be difficult to do if there is no true "primary" and failure to do so will greatly inflate the family wise error rate. In addition, when doing a set of t-tests with a primary endpoint this allows for an inference to be drawn on only a single endpoint. Thus any decision about the research hypothesis, or continuing to study the research hypothesis, relies on a single endpoint, with all other collected variables being used in an exploratory setting. If a multiplicity adjustment is used instead and the endpoints provide mixed signals this can be difficult to interpret.

4.4.6 Advantages and Disadvantages of the Linear Mixed Model

The main advantage of a linear mixed model is the ability to account for the effect of covariates. In addition, the fitted model can be used to predict the effect of an intervention on future experimental

units.

The disadvantage of a Linear Mixed model, especially in early stage research, is that it simply may not be feasible due to small sample sizes, large number of endpoints and potential covariates of interest. Similarly to the t-test, when there are many endpoints of interest a primary one can be chosen and inferences drawn from the primary model with other endpoints being exploratory or a multiplicity correction for the number of endpoints would need to be undertaken.

4.5 Empirical Comparison to the t-test

We also conducted an empirical comparison between the prediction test and a set of t-tests. We excluded the linear mixed model due to the arbitrary nature of simulating covariate values such as race and age which could have a large impact on the results. If more was known about a specific target population for some intervention then an empirical comparison including the linear mixed model could be done. For the comparison we simulated data with sample sizes of n = 5, 10, 15, and 20, and for each sample size simulated m = 5, 10, 15, 20, and 25 measures. For simplicity of the simulation we used directional increases of a prediction on each measure and simulated the pre-post difference as being positive. Specifically, we simulated post data from Normal distributions with means = 0.5 and standard deviations of 1 and pre data from N(0,1) distributions. The effect size for each of the endpoints is thus $\frac{0.5-0}{1} = 0.5$ which commonly used rules of thumb of Cohen's d would classify as "medium effects". The only endpoint that was simulated differently was the first one in the data set, for this endpoint we let the mean be 0.64 and the standard deviation be 1.

For the t-test we picked a primary endpoint, the first simulated endpoint, and also calculated the results using Sidak's adjustment on the α threshold. For the t-test with a primary endpoint, we set the primary endpoint to be the first endpoint in the data set. This was simulated to have the largest mean, i.e. the chosen primary endpoint has the largest effect size $d = 0.64$ of every endpoint in the data set. This effect is between "medium" and "large", and we know that it is in the correct direction. In addition, for the t-test with a primary endpoint we conducted a one-sided

hypothesis test, in essence we halved the threshold for "significance", if the p-value is below this threshold we consider this a finding of a significant result, that is, a rejection of the null hypothesis. This is intended to be somewhat of an ideal scenario for the primary endpoint method. For the set of t-tests using Sidak's adjustment we compare each of the p-values to $\alpha_S = 1 - (1 - 0.05)^{1/m}$. In an attempt to also make these comparisons advantageous for the t-tests we will reject the null hypothesis if any of the m measures are below α_S, for example if only 1 out of 20 measures for a given simulation is below this threshold we will still reject the null hypothesis.

We compare the t-tests to the prediction test when the true predictive ability is $\phi = 0.80$, which gives Cohen's h = 0.64, between a moderate (0.50) and large (.80) effect, the same effect size as for the t-test with primary endpoint. For these comparisons we simulated 1,000 data sets with the specified means and standard deviation. Figure 4.3 displays the empirical power estimates for the three approaches at combinations of n and m.

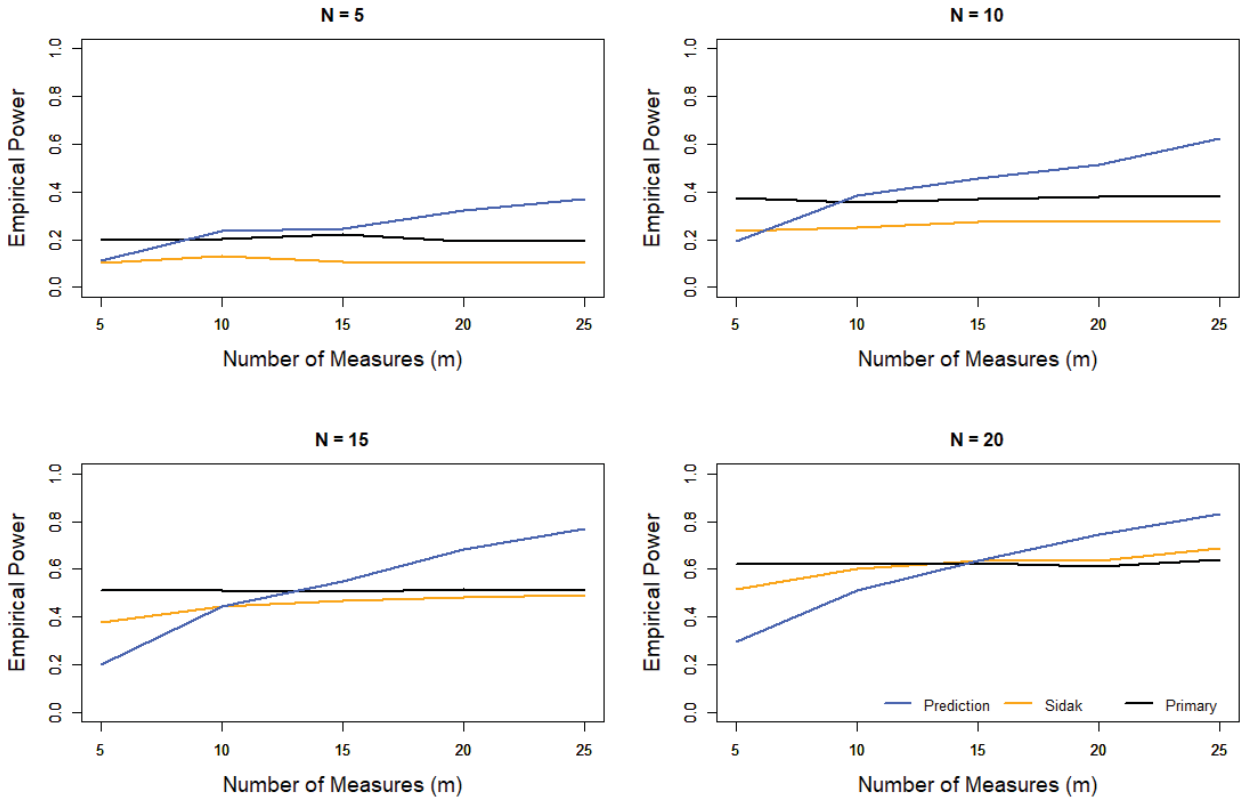


Figure 4.3: Power comparison between the Prediction Test and T-tests

Both of the approaches utilizing the t-test will increase in power as N increases and while

there may be some increase in power for the prediction test as the sample size increases the main increase is due to an increase in the number of measures of interest. The t-test with a primary endpoint seems to be more powerful than the set of t-tests using the Sidak adjustment, which is not surprising since the primary endpoint data was drawn from the most extreme endpoint given the null hypothesis was true. Only when the number of measures of interest, m, are small do the other techniques outperform the Prediction test, which is to be expected since it is intended for use with many endpoints. As m increases so does the power of the prediction test for a fixed sample size. For smaller sample sizes, especially n = 5 and n = 10 the prediction test outperforms the other methods except when m = 5. For n = 15 and n = 20 the prediction test out performs both competitors for $m \geq 15$.

Overall when comparing the methods there were 20 combinations of n and m, the Prediction test achieved the highest empirical power in 14 of the 20 even with the attempts to provide advantages to the t-tests. The prediction test performed well; this provides more evidence that the prediction test should be used for go/no-go decisions in early stage research.

4.6 Discussion

We have discussed a global hypothesis test that can provide a go/no-go decision in early stage research with an application to a real data set and comparisons to other common methods. The prediction test came to the same conclusion as both the set of t-tests and the linear mixed model for the DTI analysis. All analyses showed that the research hypothesis of structural changes in the brain post transplant is promising, there is evidence that this is what is occurring and this should be researchered more fully to understand the interaction between kidney function and Alzheimer's disease.

The prediction test is unique in that it addresses a real issue in early stage research, that of whether to continue following an idea. The point of pilot studies, and other studies with small sample sizes is typically not confirmatory in nature; however the methods that are often used in

these studies seek to draw a conclusive inference despite the fact that even if the null hypothesis is rejected a larger study will still be needed to verify the result. Our test acknowledges that early stage research is primarily seeking to identify promising research hypotheses and weed out ones that are unlikely to be true. By doing this we are able to gain an advantage in power. The prediction test provides a formal mechanism that can be used for very small sample sizes and many measures, if the research hypothesis as predicted by the PI is holding true, regardless of whether the individual endpoints are "statistically significant" then the research hypothesis should be studied in more depth.

4.6.1 Best practices

The applications of the prediction test can be quite broad, we provide a couple best practices in order to insure that the testing procedures performs well. The predictions on the endpoints must be made before the data is collected and cannot be changed post analysis. We believe that our test is especially susceptible to misuse due to the fact that a single change of prediction could result a different conclusion and after the data has been collected and analyzed it can seem easy to understand how predictions went wrong in ways that could still support the research hypothesis. The test is only valid if the predictions are made a-priori. The endpoints that are included in the study need to all deal with the same research hypothesis, if the endpoints are not all somehow related then they should not be analyzed using the prediction test.

4.6.2 Limitations

Like any method there are limitations to our approach. One of the simplest is making predictions, in some settings such as truly exploratory research there may not be a prediction about a given endpoint. All endpoints on which predictions are made need to be influenced by the research hypothesis, that is if the research hypothesis concerns cerebral blood flow as in our example, correctly predicting the results of pre-post differences in views about a political issue would be of little in-

terest to the research hypothesis and thus should not be included in the set of predictions. The test also currently requires paired data. The biggest limitation is that the test cannot make an inference about the research hypothesis, rather it deals with whether the researcher's hypothesis should be studied in more depth.

Chapter 5

Summary and Future Work

We have developed multiple statistical techniques to deal with issues common in biomedical research, and have applied them specifically to Alzheimer's disease studies. Our methods deal with adverse situations when standard approaches may fail. The study of the impact on Self Revelatory Performance could provide a method for mediating negative perceptions of Alzheimer's disease at a community level. More generally, our methodology could be applied to Likert Scale data, since the Affect Grid is essentially a two dimensional Likert Scale. Likert Scale data is often analyzed using a combined summary score and a t-test, either of a difference between groups or against a null hypothesis for a single group. For an n dimensional Likert scale we could compute the distance between the n dimensional center of mass of different groups, or of the distance relative to some hypothesized center of mass. This would take into account all the data instead of simply using a summary scale and would not make assumptions about the data that are not true, such as the data being continuous. As n increases we would expect scarcity to be a problem, thus different approaches to weighting would need to be used, nevertheless this could provide a more sensitive test than using a t-test.

The results of the Prediction Test for the ASL data suggest that the exercise intervention may not have a strong impact on structural changes in the brain in older adults; nevertheless the test we developed provides researchers with a formal decision criteria for whether or not continue studying a given hypothesis. We've shown that the test has good power and type I error control, and specifically that the power increases for a fixed sample size when more endpoints are predicted, an excellent property for early stage research. We believe it has the potential to be a useful tool for researchers.

The analysis of the DTI data was more promising and showed that post Kidney transplantation both FA and MD values appeared to normalize, lending credence to the research hypothesis that

kidney function impacts structural changes in the brain. The next steps on this research track will be to verify those results and determine specific structural changes and what is driving them. In addition, we extended our original predictions to include two sided predictions, and showed via the DTI data and a simulation that the prediction test seems to have advantages for studies with many endpoints and small sample sizes. More work needs to be done to determine specific scenarios when the prediction test should be used.

Future research could include extending the set of possible predictions, for instance including equivalence prediction. We are also investigating ways to use the Prediction Test with unpaired data, with unpaired data we can easily get the result of a prediction, e.g. this group has a larger sample mean than another group; however we cannot calculate a pairwise correlation between unpaired data thus another method of estimating the "correlation" will need to be used. Tt would be interesting to apply the prediction test to a set of early research studies as a supplementary analysis and then, given enough funding, do larger studies of the same hypothesis and see how the prediction test compares to other methods in identifying promising hypotheses. The Shiny application, with early work presented in Appendix A, is another avenue of research. The app currently can calculate the prediction test for several different types of data and predictions, we would like to extend this to more scenarios and make it more user friendly as well as adding the ability to do a power calculation so that researchers would not have to code their own functions or know a specific programming language to use the test.

References

Agresti, A. (2013). *Categorical Data Analysis, Third Edition*. Wiley.

Alexander, A., JE, L., Lazar, M., & Field, A. (2007). Diffusion tensor imaging of the brain. *Neurotherapeutics*, 4(3), 316–329.

Alzheimer's Association (2018). 2018 alzheimer's disease facts and figures. *Alzheimer's & Dementia*, 14(3), 367 – 429.

Anderson, L. A., Day, K. L., Beard, R. L., Reed, P. S., & Wu, B. (2009). The public's perceptions about cognitive health and alzheimer's disease among the u.s. population: A national review. *The Gerontologist*, 49(S1), S3–S11.

Benjamini, Y. & Hochberg, Y. (1984). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *JRSS Series B*, 57(1), 289–300.

Billingsley, P. (1995). *Probability & Measure*. New York, New York: Wiley. ISBN 0471007102'.

Bonnett, D. & Wright, T. (2000). Sample size requirements for estimating pearson, kendall and spearman correlations. *Psychometrika*, 65(1), 23–28.

Burns, N. C., Watts, A., Perales, J., Montgomery, R. N., Morris, J. K., Mahnken, J. D., Lowther, J., & Vidoni, E. D. (2018). The impact of creative arts in alzheimer's disease and dementia public health education. *Journal of Alzheimer's disease : JAD*, 63(2), 457—463.

Chang, W., Cheng, J., Allaire, J., Xie, Y., & McPherson, J. (2018). *shiny: Web Application Framework for R*. R package version 1.1.0.

Chevrud, J. (2001). A simple correction for multiple comparisons in interval mapping genome scans. *Heredity*, 87(1), 52–58.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. New York, New York: Lawrence Erlbaum Associates. ISBN 9780805802832.

Dauphinot, V., Delphin-Combe, F., Mouchoux, C., Dorey, A., Bathsavanis, A., Makaroff, Z., Rouch, I., & Krolak-Salmon, P. (2015). Risk factors of caregiver burden among patients with alzheimer's disease or related disorders: a cross-sectional study. *Journal of Alzheimer's disease : JAD*, 44(3), 907—916.

Deza, M. & Deza, E. (2009). *Encyclopedia of Distances*. Springer.

Dmitrienko, A. & D'Agostino, R. (2017). Editorial: Multiplicity considerations in clinical trials. *Statistics in Medicine*, 36, 4423–4426.

Dmitrienko, A., Tamhane, A., & Bretz, F. (2009). *Multiple Testing Problems in Pharmaceutical Statistics*. Boca Raton, Florida: CRC Press. ISBN 9781584889847.

Efron, B. (1979). Bootstrap methods: Another look at the jacknife. *The Annals of Statistics*, 7(1), 1–26.

EMA (2017). Ema guidline on multiplicity issues in clinical trials. Online. `http://www.http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2017/03/WC500224998.pdf`.

Emunah, R. (2015). Self-revelatory performance: A form of drama therapy and theatre. *Drama Therapy Review*, 1(1), 71–85.

FDA (2017). Fda draft guidance for industry. multiplicity endpoints in clinical trials. `http://www.fda.gove/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM536750.pdf`.

Furnham, A. (1986). Response bias, social desirability and dissimulation. *Personality and Individual Differences*, 7(3), 385–400.

Holms, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2), 65–70.

Hotelling, H. (1931). The generalization of student's ratio. *Annals of Mathematical Statistics*, 2(3), 360–378.

Joe, H. (2006). Generating random correlation matrices based on partial correlations. *Journal of Multivariate Analysis*, 97(10), 2177–2189.

Killgore, W. (1998). The affect grid: a moderately valid, nonspecific measure of pleasure and arousal. *Psychological reports*, 83(2), 639—642.

Kinsinger, E. (2004). Remembering emotional experiences: The contribution of valence and arousal. *Reviews in the Nuerosciences*, 15(4), 241–252.

Li, J. & Ji, L. (2005). Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity*, 95(3), 221–227.

Li, J., Tai, B. C., & Nott, D. J. (2009). Confidence interval for the bootstrap p-value and sample size calculation of the bootstrap test. *Journal of Nonparametric Statistics*, 21(5), 649–661.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, 22(140), 5–55.

Mardia, K. (1975). Assessment of multinormality and the robustness of hotelling's t^2 test. *Journal of the Royal Statistical Society, Series C*, 24(2), 163–171.

Maturi, T. & Elsayigh, A. (2010). A comparison of correlation coefficients via a three-step bootstrap approach. *Journal of mathematics research*, 2(2), 3–10.

Montgomery, R. & Mahnken, J. (2019). The prediction test: A go/no-go hypothesis test for early stage research. Manuscript submitted for publication.

Newcombe, R. (1998). Two-sided confidence intervals for the single proportion: comparison of seven methods. *Statistics in Medicine*, 17(8), 857–872.

O'Briend, P. (1984). Procedures for comparing samples with multiple endpoints. *Biometrics*, 40(4), 1079–1087.

O'Donnell, L. & Westin, C. (2011). An introduction to diffusion tensor image analysis. *Neurosurgery Clinics of North America*, 22(2), 185–196.

O'Neill, R. (1997). Secondary endpoints cannot be validly analyzed if the primary endpoint does not demonstrate clear statistical significance. *Controlled Clinical Trials*, 18(9), 550–556.

O'Neill, R. (1999). End-point interpretation in clinical trials: the case for discipline. *Controlled Clinical Trials*, 20(1), 40–51.

Protter, M. & Protter, P. (2009). *Calculus with Analytic Geometry*. Jones and Bartlett.

Qui, W. & Joe, H. (2015). clustergeneration: Random cluster generation (with specified degree of separation). R Package. `https://CRAN.R-project.org/package=clusterGeneration`.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Russell, J., Weiss, A., & Mendelsohn, G. (1989). Affect grid: A single-item scale of pleasure and arousal. *Journal of Personality and Social Psychology*, 57(3), 493–502.

Russell, Y. & Gobet, F. (2012). Sinuosity and the affect grid: A method for adjusting repeated mood scores. *Perceptual and Motor Skills*, 114(1), 125–136.

Schonbrodt, F. & Perugini, M. (2013). At what sample size do correlations stabilize. *Journal of Research in Personality*, 47(5), 609–612.

Sidak, Z. (1967). Rectangluar confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*, 62(318), 626–633.

Soares, J., Marques, P., Alves, V., & Sousa, N. (2013). A hitchhiker's guide to diffusion tensor imaging. *Frontiers in Neuroscience*, 7.

Team, R. C. (2017). R: A language and environment for statistical computing. Online. https://www.R-project.org/.

Townsend, M. (2011). When will alzheimer's disease be cured? a pharmaceutical perspective. *Journal of Alzheimer's disease : JAD*, 24 Suppl 2, 43 – 52.

Wilson, E. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158), 209–212.

Winslow, B. T., Onysko, M. K., Stob, C. M., & Hazlewood, K. A. (2011). Treatment of alzheimer disease. *American family physician*, 83(12), 1403—1412.

Appendix A

Shiny app for the Prediction Test

A.1 Introduction

We've developed a user friendly tool for the Prediction Test. The tool is a shiny app (R Core Team, 2018) (Chang et al., 2018).This tool requires no software on the user-end, and the help file provides instructions on using the app. The focus of the prediction test is on early stage research, a stage when funding might not be available for a statistician or programmer to do an analysis, thus the tool allows a researcher to upload their data in several different formats and get the calculated results. Figure A.1 shows the prediction test as seen when opened.

A.2 Types of data and predictions

Researchers are currently only able to upload their data sets to the app in a Comma Separated Variable (csv) format. The two required data sets are one with the data on the relevant endpoints and another separate data set consisting of predictions for the endpoints. If both are successfully uploaded a progress bar displaying "Upload Complete" will display beneath the upload boxes. Both the data set and the predictions can take on two types of formats.

A.2.1 Raw data

The first format for the data set is the "Raw" format. This consists of a column for every endpoint, with the values for every endpoint being those needed to conduct the Prediction Test. For instance if the predictions are on a comparison to baseline measures, then the values in the Raw format would need to be the differences from baseline for every measure. An example data set is shown in on the left side of Figure A.2.

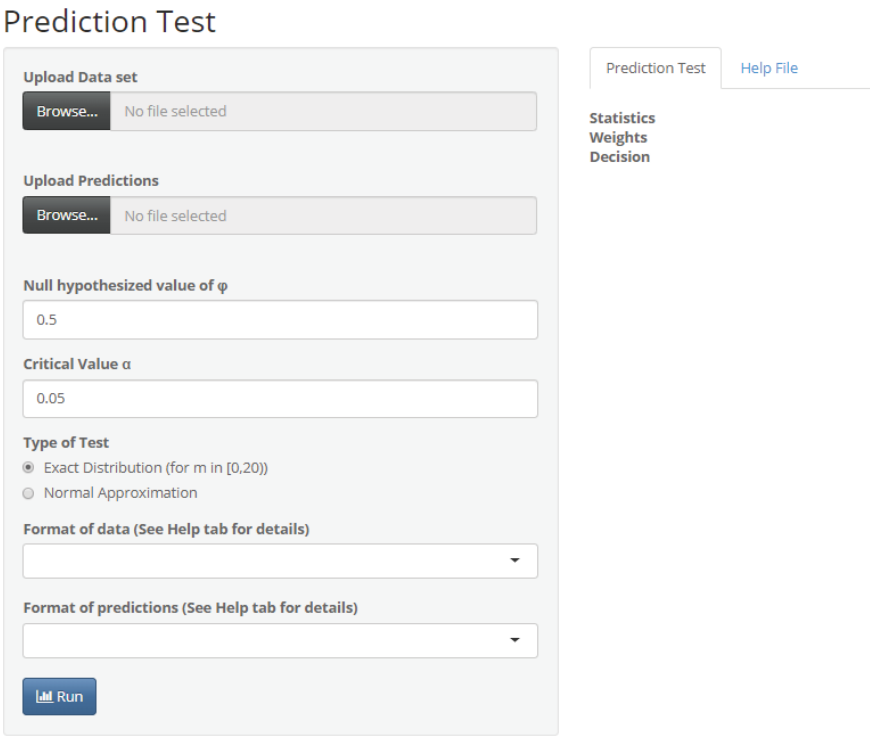


Figure A.1: App as displayed when opened

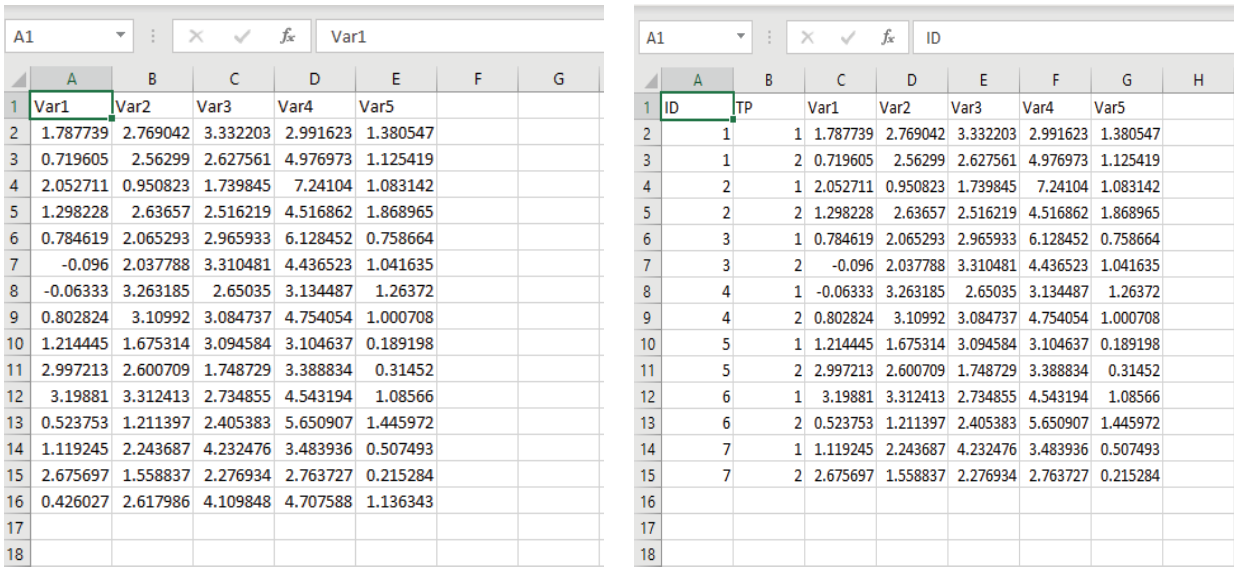
A.2.2 Pre-post data

Data in the "Pre-post" format are required to be input with the following variables only: an ID variable, named ID, a time point variable, named TP, and all the observed values for the endpoints of interest. By default the prediction test will do the pre (TP 1) vs post (TP 2) difference. If there are more than 2 time points for each participant only the two on which the prediction is being based should be included, and example is shown on the right side of Figure A.2 .

A.2.3 "Results" predictions

The simplest way to upload the predictions is the upload the "Results" of the predictions, a 1 for a successful prediction and a 0 for an unsuccessful prediction for each endpoint. This method is the most flexible in that predictions of many types can be calculated however it may be more time consuming on the front-end to determine the result of each individual prediction.

The predictions need to be formatted as a column in the CSV named Predictions, with row 1 in



	A	B	C	D	E	F	G
1	Var1	Var2	Var3	Var4	Var5		
2	1.787739	2.769042	3.332203	2.991623	1.380547		
3	0.719605	2.56299	2.627561	4.976973	1.125419		
4	2.052711	0.950823	1.739845	7.24104	1.083142		
5	1.298228	2.63657	2.516219	4.516862	1.868965		
6	0.784619	2.065293	2.965933	6.128452	0.758664		
7	-0.096	2.037788	3.310481	4.436523	1.041635		
8	-0.06333	3.263185	2.65035	3.134487	1.26372		
9	0.802824	3.10992	3.084737	4.754054	1.000708		
10	1.214445	1.675314	3.094584	3.104637	0.189198		
11	2.997213	2.600709	1.748729	3.388834	0.31452		
12	3.19881	3.312413	2.734855	4.543194	1.08566		
13	0.523753	1.211397	2.405383	5.650907	1.445972		
14	1.119245	2.243687	4.232476	3.483936	0.507493		
15	2.675697	1.558837	2.276934	2.763727	0.215284		
16	0.426027	2.617986	4.109848	4.707588	1.136343		
17							
18							

	A	B	C	D	E	F	G	H
1	ID	TP	Var1	Var2	Var3	Var4	Var5	
2	1	1	1.787739	2.769042	3.332203	2.991623	1.380547	
3	1	2	0.719605	2.56299	2.627561	4.976973	1.125419	
4	2	1	2.052711	0.950823	1.739845	7.24104	1.083142	
5	2	2	1.298228	2.63657	2.516219	4.516862	1.868965	
6	3	1	0.784619	2.065293	2.965933	6.128452	0.758664	
7	3	2	-0.096	2.037788	3.310481	4.436523	1.041635	
8	4	1	-0.06333	3.263185	2.65035	3.134487	1.26372	
9	4	2	0.802824	3.10992	3.084737	4.754054	1.000708	
10	5	1	1.214445	1.675314	3.094584	3.104637	0.189198	
11	5	2	2.997213	2.600709	1.748729	3.388834	0.31452	
12	6	1	3.19881	3.312413	2.734855	4.543194	1.08566	
13	6	2	0.523753	1.211397	2.405383	5.650907	1.445972	
14	7	1	1.119245	2.243687	4.232476	3.483936	0.507493	
15	7	2	2.675697	1.558837	2.276934	2.763727	0.215284	
16								
17								
18								

Figure A.2: Example Data Sets

column 1 corresponding to prediction for the first endpoint column in the data set, row 2 in column 1 corresponding to the prediction for the second endpoint column in the data set and so on. An example is shown on the left of Figure A.3.

A.2.4 "Type" predictions

The "Type" prediction allows for one-sided and two sided predictions to be entered. Similarly to the "Results" prediction the format needs to be a single column with row 1 corresponding to the first endpoint column in the uploaded data set. Predictions in this format must have three columns, "Predictions", "Null" and "SD". The Predictions column refers to the type, a "U" stands for up, a prediction of a directional increase, "D" stands for down, a prediction of a decrease, and "Diff" is for a two sided prediction of being different from some value. The Null column will default to 0 if left blank, for directional predictions it is the value the prediction is being over or under. For predictions of a difference, it is the value that the prediction states the endpoint is different from. The SD column denotes how many standard deviations away from the Null value the observed value needs to be for a two sided prediction to be deemed correct. A discussion of the choice of this value is in Chapter 4. An example is show in Figure A.3

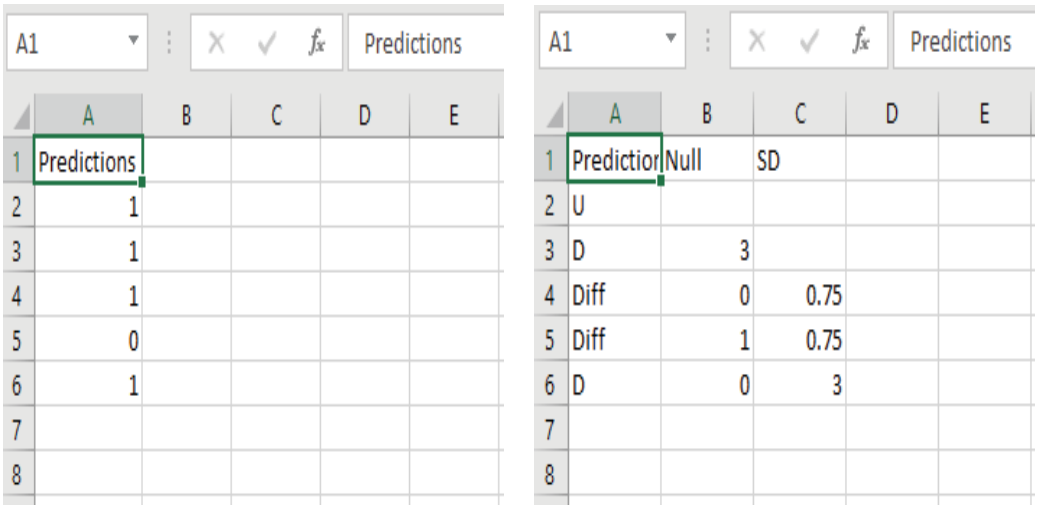


Figure A.3: Example Predictions

As an example third row the predictions would correspond to a prediction of a decrease below a baseline of 3, the fourth row would indicate a prediction of a difference from 0 with 0.75 standard deviations as the threshold.

A.3 Parameters

The parameters of the app are the hypothesized value ϕ which can take on any value between 0 and 1. The default is set to 0.5. The critical value α is the threshold for "significance", it's default is 0.05. The type of test can either be exact or using a normal approximation. The exact test can only be used for $m < 20$.

A.4 Output

When the data has been uploaded correctly, selecting the Run button will run the Prediction Test according to the provided predictions. The output consists of the observed test statistic value, the p-value, weights for all the variables of interest and the decision based on the critical value. Figure A.4 displays the results with an example data set.

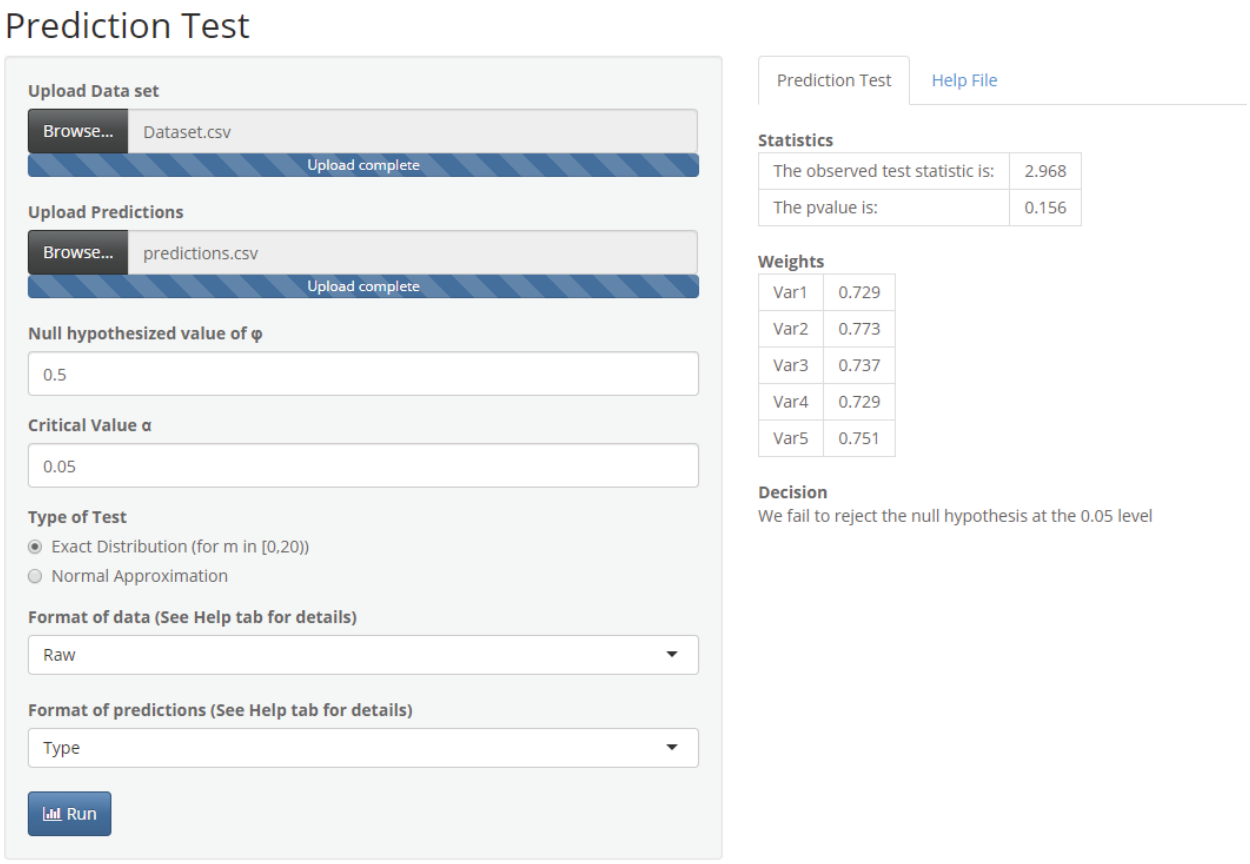


Figure A.4: Results of the Prediction Test

Appendix B

Code for to calculate the prediction test

```
#Required values:
#C: Sample correlation matrix
#observed: The Observed test statistic
#null_phi: Hypothesized value of phi
#alpha: alpha level

#By default the test will be calculated using the Exact
   distribution when m < 20,
#and the Normal
#approximation for m >=20, this can be changed within the function


prediction_test <- function(C, observed, null_phi, alpha){

  n_tests <- ncol(C)

  if (n_tests <20){ #Use exact test for m <20
    options(digits = 10)
    n_perm <- 2^n_tests
    m <- n_tests
    W <- as.matrix(1/rowSums(C^2))
    SW <- sum(W)

    perms <- as.matrix(expand.grid(rep(list(0:1), n_tests)))
    correct <- rowSums(perms)
```

```
  perms <- cbind(perms, correct)
  perms <- perms[order(perms[,m+1]),]

  out <-as.numeric(perms[,1:m] %*% W)
  perms <- cbind(perms,out)
  perms <- cbind(perms, as.numeric(dbinom(sort(correct), m,
      null_phi))/as.numeric(lapply(sort(correct), function(x)
      choose(m,x) )))
  perms <- perms[order(perms[,m+2]),]
  perms <- cbind(perms,cumsum(perms[,m+3]))

  pval <- (1-perms[which(round(perms[,m+2],10) ==
      as.numeric(round(observed,10)))  ,m+4]) +
      perms[which(round(perms[,m+2],10) ==
      as.numeric(round(observed,10)))  ,m+3]

  if (pval < alpha)
  {
    decision <- 1
  } else if (pval >= alpha){
    decision <- 0
  }
} else if (n_tests >=20){ #Use Normal Approximation for m >=20

  W <- as.matrix(1/rowSums(C^2))
  SW <- sum(W)

  z <- (observed -
      null_phi*SW)/sqrt(null_phi*(1-null_phi)*sum(W^2)  )
```

```
    pval <- 1-pnorm((z))

    if (pval < alpha)
    {
      decision <- 1
    } else if (pval >= alpha){
      decision <- 0
    }
  }
  return(list(pval,decision))
}
```