STATISTICAL APPROACHES IN THE RDOC PARADIGM FOR POST-MORTEM BRAIN TISSUE STUDIES

by

Hong Gu

B.S., Tsinghua University, 2006

M.S., University of Cincinnati, 2009

Submitted to the Graduate Faculty of

the Kenneth P. Dietrich School of Arts and Sciences in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2015

UNIVERSITY OF PITTSBURGH

KENNETH P. DIETRICH SCHOOL OF ARTS AND SCIENCES

This dissertation was presented

by

Hong Gu

It was defended on

December 2, 2015

and approved by

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

Henry W. Block, Ph.D., Department of Statistics, Professor

Kehui Chen, Ph.D., Department of Statistics, Assistant Professor

Robert A. Sweet, M.D., Department of Psychiatry, Professor

Dissertation Director: Allan R. Sampson, Ph.D., Department of Statistics, Professor

STATISTICAL APPROACHES IN THE RDOC PARADIGM FOR POST-MORTEM BRAIN TISSUE STUDIES

Hong Gu, PhD

University of Pittsburgh, 2015

Mental disorders are diagnosed by the way a person perceives and behaves. No neurobiological measures are involved in current diagnostic systems such as the Diagnostic and Statistical Manual of Mental Disorder (DSM), which is the mostly widely used. Because of the lack of neurobiological etiological information, the DSM diagnosis is ambiguous in two ways: first, patients within the same diagnosis could be different in both symptoms and neurobiological measures; second, patients with different DSM diagnoses could be similar in both symptoms and neurobiological measures. As a result, treatments for mental illnesses have not been accurate and successful.

In order to better understand and treat mental illnesses, the National Institute of Mental Health has launched a Strategic Plan for Research in 2008. Part of the plan is the Research Domain Criteria (RDoC) Initiative, which is a framework to link basic neurobiology with mental functions. Under the RDoC framework, a study focuses on a particular mental function, which is called construct, and would ignore the DSM diagnosis of a patient. The RDoC intends to guide studies to find neurobiological characteristics such as genes that are significantly associated with a construct of interest and also the mechanism how defects in these neurobiological characteristics lead to illness in the construct. Although without symptom measures, existing post-mortem brain tissue databases are still useful to facilitate RDoC research.

In this dissertation, we develop statistical approaches to utilize existing post-mortem brain tissue databases following the RDoC spirit. We first propose a method to identify the neurobiological characteristics that are significantly associated with a construct of interest. This is achieved through identifying the neurobiological characteristics that are significantly associated with all the DSM diagnoses relevant to the construct. We then propose a matched subject study design to compare the distribution of the identified neurobiological characteristics in the means and the quantiles between the population with dysfunction in the construct and the healthy population. Finally we develop an algorithm to optimally determine the sample size for each DSM diagnosis in the matched subject study subject to a sample availability constraint as well as a budget constraint.

Keywords: RDoC, Post-Mortem Brain Tissue Database, FDR Control, Quantile Regression, Optimal Design, Mixture Population.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1.0 INTRODUCTION

As the National Institute of Mental Health (NIMH) is calling for new strategies in mental health research, new statistical approaches need to be developed accordingly to push forward the progress. In the mean time, data that were collected previously are still useful and should not be simply discarded. In this dissertation, we develop several statistical methods that use the existing post-mortem tissue databases to facilitate future studies in mental health research.

1.1 BACKGROUND

1.1.1 Definition of Mental Disorders

Mental disorders are abnormalities in one's perception and behavior that can lead to significant disability or emotional instability such that normal daily life is impacted. There are many different types of mental disorders. For example, anxiety disorder is characterized by the aberrant anxiety or worry about the future and is severe enough to impact normal functioning. Mood disorders involve abnormal emotional difficulties such as severe and sustained sadness. The description of mental disorders in human history dates back to ancient times when people in different regions started to record different disorders. In the past, mental disorders were considered diseases of the mind, whereas today, due to developments in modern neuroscience, they are understood as dysfunctions of the physical components in the brain. In particular, the recent advancements in neuroscience and basic biotechnology such as molecular biology, genetics and imaging technology, have equipped researchers in psychiatry with powerful tools. With these modern techniques, they have been able to not only gain more accurate knowledge about how the brain works but also seek the fundamental biological mechanisms of mental disorders. The rapid developments in different aspects of biomedical research provide valuable new insights to permit rethinking mental disorders.

In general, mental disorders are defined by the way a person perceives and behaves. Though variations exist in the identification and evaluation of mental heath conditions, standard guidelines in diagnostics are widely used. One of the most commonly accepted classification systems is the Diagnostic and Statistical Manual of Mental Disorders (abbreviated as DSM) produced by the American Psychiatric Association since 1952.

1.1.2 Problems with the DSM

The DSM provides standardized diagnostic criteria, and is widely used by clinicians, researchers, pharmaceutical companies and regulatory agencies. The latest version is the DSM-5 issued in May 2013. Even with its prevailing use, the DSM is a source of controversy and criticism. One of the major concerns about the DSM is its reliance on categorization of observed symptoms and the use of artificial thresholds to distinguish different disorders and distinguish abnormal from normal. Two examples, schizophrenia and bipolar disorder, are used below to illustrate the DSM approach. Schizophrenia is a psychotic disorder marked by severely impaired cognition, emotions, and behaviors and has a worldwide lifetime prevalence of 0.3% - 0.66% (van Os & Kapur (2009)). A person is diagnosed with schizophrenia under the DSM-5 if the following criteria are all satisfied: 1)two or more characteristic symptoms from delusions, hallucinations, disorganized speech, grossly disorganized behavior and negative symptoms, at least one must be from the first three and each should present for a significant portion of time during one month period or less if successfully treated; 2)social or occupational dysfunction; 3)significant duration for at least six months, which includes at least one month of symptoms that meets Criterion 1); and 4)the disturbance is not attributable to the physiological effects of a substance or another medical condition. Bipolar disorder is a mood disorder in which a patient suffers mood alterations from mania to depression. It is estimated to have a lifetime prevalence of 3.9% in the US (Kessler et al. (2005)).

Most bipolar disorder patients experience alternating episodes of mood swings and the transition between mania and depression can be very quick. There are more than one sub-type of bipolar disorders under the DSM-5. People are identified with bipolar disorder I if they show at least three symptoms of manic episodes during most of the days in at least one week and the mood disturbance is severe enough to cause marked impairment in social or occupational functioning and is not due to physiological effects of a substance or other medical conditions. People are diagnosed to have bipolar disorder II if they have no manic episodes, but at least one hypomanic episode and at least one major depressive episodes. Symptoms of manic episodes include inflated self-esteem or grandiosity, decreased need for sleep, more talkative than usual or pressure to keep talking, flight of ideas or subjective experience that thoughts are racing, distractibility, increase in goal-directed activity or psychomotor agitation, and excessive involvement in activities that have a high potential for painful consequences. Patients with bipolar disorder I are also common to have major depressive episodes, which include symptoms such as depressed mood most of the day, nearly every day, diminished ability to think or concentrate, or indecisiveness, nearly every day and recurrent thoughts of death, recurrent suicidal ideation without a specific plan, or a suicide attempt or a specific plan for committing suicide. In diagnosing bipolar disorder I, the occurrence of the manic and major depressive episodes is not better explained by schizoaffective disorder, schizophrenia, schizophreniform disorder, delusional disorder, or other specified or unspecified schizophrenia spectrum and other psychotic disorder. From the above descriptions, we can see that the DSM definitions of mental disorders are only some typical symptoms. Also the diagnostic criteria set by the DSM rely on self-reported behaviors from the subject or evaluation of some functional tasks. If the behavior from a subject is close to the typical symptoms for a specific disorder, then the subject is categorized to that disorder. The closer one's behavior is to the typical symptoms, the more precise and reliable the diagnosis is. However, there are dramatic variations in people's behaviors in the real world and those diagnosed with the bipolar disorder can show similar psychotic symptoms to individuals diagnosed with schizophrenia, as reported in Khan & Akella (2009). In such cases, the DSM criteria are not unique to one disorder.

In addition to the ambiguity in symptoms or phenotypes, the explosive developments in neurosciences and basic biotechnology mentioned previously have started to suggest that mental disorders diagnosed by the DSM, which we call DSM diagnoses in this dissertation, may not be homogeneous diseases even in terms of genotypes. For instance, schizophrenia is now viewed as caused by multiple genes whose individual effects were insignificant but whose collective effects can be substantial (Bray et al. (2010)). Some schizophrenic patients may have deficits in these genes and other patients may have deficits in other genes. Thus among the patients with the same DSM diagnosis, there might be sub-groups according to their genotypes. On the other hand, patients classified into different DSM diagnoses may share common biological characteristics. For example, quite a few genes have been found associated with both schizophrenia and bipolar disorder based on multiple studies (Owen, Craddock & Jablensky (2007)). Lichtenstein et al. (2009) conducted a very large family study and found that relatives of the study subjects diagnosed with either schizophrenia or bipolar disorder were at higher risks for both DSM diagnoses. This study suggested that there might be some underlying biological characteristics that are responsible for both DSM diagnoses. Even though the exploration of the relationship between the biological characteristics and clinical features for mental illness is still at the early stage, the molecular findings reported so far show that the classification using the DSM, which is based mostly on clinical observations, may not match well with the underlying biology. In other words, the etiology and psychopathology of mental disorders are not considered in the diagnosis with the DSM. In light of all these, the DSM might not be the optimal classification guideline given the exploding information on genes, proteins, neural circuits, etc.

The NIMH launched a project in 2009 called the Research Domain Criteria Initiative (RDoC) as part of its longterm strategic plan for research. The RDoC's goal is to move to an enhanced understanding of the mechanism of mental illness so that new methods to classify mental illness can be developed. Using the new methods genetic and neurobiological information such as gene expression levels, neural circuit measurements and brain imaging data will be integrated with clinical observations. As the RDoC is targeting more on the underlying dysfunctions in the brain rather than just symptoms in behavior to recruit subjects into research studies, it is hoped that more successful treatments can be obtained. In addition to the breakthroughs in the biological sciences and technologies, new statistical methodology is also required to support the implementation of the RDoC.

1.2 INTRODUCTION TO THE RDOC

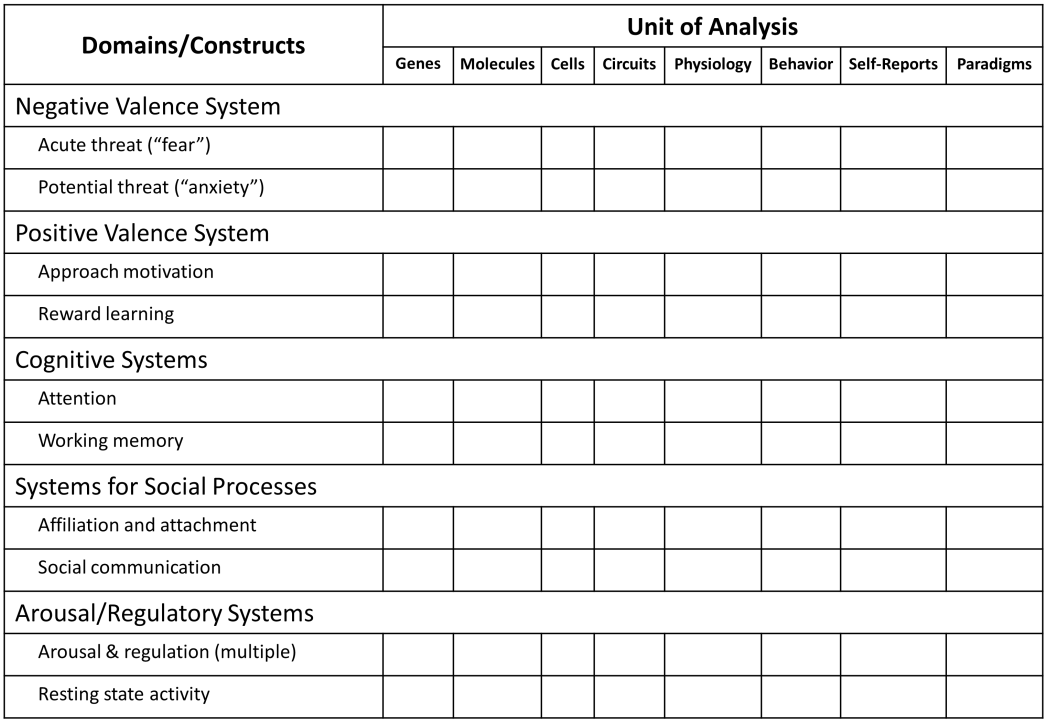
The NIMH Strategic Plan for Research is a comprehensive plan which incorporates different aspects of mental health research to fulfill NIMH's mission to "transform the understanding and treatment of mental illnesses through basic and clinical research, paving the way for prevention, recovery, and cure" (NIMH Strategic Plan for Research). As described in the plan, there are four strategic objectives to be achieved concentrating on the cause, development, treatment and public health impact of mental illnesses. The RDoC Initiative, which was originally part of Strategic Objective 1 in the 2008 NIMH Strategic Plan, has now evolved to be one of the cross-cutting themes that are relevant to each of the strategic objectives. The RDoC calls for transforming the diagnostics of mental illness. In particular, it is "a research framework for new ways of studying mental disorders. It integrates many levels of information (from genomics to self-report) to better understand basic dimensions of functioning underlying the full range of human behavior from normal to abnormal " (NIMH RDoC).

1.2.1 The RDoC

The RDoC Initiative contributes to the NIMH Strategic Plan towards the new classification method for mental illness, which will integrate reliable and valid measures from both basic biological and clinical resources. The RDoC is a framework that directly brings in the information from neurobiological studies and sets up the research foundations for new methods. It can be viewed as a bridge linking the more basic biological components in the brain with mental functional dimensions such as fear and working memory while ignoring the current diagnostic structure.

As described in the NIMH RDoC, the essence of the RDoC can be thought of as a matrix (see Figure 1.1) in which each row is a specific mental functional dimension, for example, acute threat and attention. Each row is called a “construct”, which is a concept with integrated information about the specified mental functional dimension. The constructs are the basic entities to be studied under the RDoC framework and NIMH suggests that most RDoC studies should focus on one construct. Related constructs are grouped together into a bigger concept called “Domain”, which reflects the understanding of major aspects of psychiatry such as emotion or behavior. For the time being, five domains are used in the matrix: Negative Valence Systems, Positive Valence Systems, Cognitive Systems, Systems for Social Processes, and Arousal/Regulatory Systems (NIMH RDoC). The columns in the matrix are defined to be different variables such as genes, molecules and circuits, etc and are called units of analysis in the RDoC. They can be either dependent or independent variables in the research study. The goal is to establish the relationship between each row and column, i.e., to fill in each matrix cell with specific details connecting each row and column. For example, to fill the cell in the column “Gene” and the construct “Acute Threat”, it is critical to identify the individual genes that are potentially significantly associated with the acute threat function, and also to find out how the structural variations or changes in expression levels of these genes relate to changes in acute threat function or what types of mutations in these genes would cause dysfunctions in acute threat. Even more specific constructs have been developed, which are called “sub-constructs”. For example, the construct “Working Memory” has sub-constructs “Active Maintenance”, “Flexible Updating”, “Limited Capacity” and so on (NIMH RDoC), because working memory is a large concept including more than one aspect. So far, some of the cells in the RDoC matrix have been filled with currently known knowledge. However, the NIMH emphasized that the current RDoC matrix is just a draft and the rows and columns are not definitive. It is expected that researchers will revise it as the knowledge about the brain accumulates so that more columns and rows may be added.

Three basic principles are suggested in implementing the RDoC. The first is that the RDoC is a dimensional system ranging from normal to abnormal. People have known that mental disorders under the DSM criteria span more than one dimension, in terms of both symptoms and neurobiological measures. For example, schizophrenia patients may suffer from disrupted working memory, hallucination and so on. Each of these symptoms relates



Domains/Constructs	Unit of Analysis							
	Genes	Molecules	Cells	Circuits	Physiology	Behavior	Self-Reports	Paradigms
Negative Valence System								
Acute threat (“fear”)								
Potential threat (“anxiety”)								
Positive Valence System								
Approach motivation								
Reward learning								
Cognitive Systems								
Attention								
Working memory								
Systems for Social Processes								
Affiliation and attachment								
Social communication								
Arousal/Regulatory Systems								
Arousal & regulation (multiple)								
Resting state activity								

Figure 1.1: Illustration of the RDoC Matrix (Excerpt from the NIMH RDoC)

to a different construct in the RDoC matrix. Also schizophrenia patients may have anomalies in some gene expression levels or protein structures and thus more than one columns in the RDoC matrix are involved in schizophrenia. Each row or column can be thought of as one dimension to describe mental illness. In each dimension, the measurements vary over the spectrum from normal to abnormal among the general population which consists of the healthy people, as well as those with mental problems of different degrees. Currently, the DSM criteria diagnose patients if the symptoms are severe enough. Therefore, typical research studies comparing the population with a certain DSM diagnosis with the healthy population include only two types of subjects: those who are already severe in some symptoms and those who are unaffected by any mental disorder. However, measurements from these two types of subjects are usually at the two opposite ends in a spectrum of some dimension of interest. And there is still a large population who are neither healthy nor totally severe in the symptoms, so that information is missing for this in between population. With the old research paradigm, it has been very difficult to establish the relationship between measures on the neurobiological characteristics and measures on the mental functions in a continuous fashion. For instance, it has been hard to examine how working memory changes according to variation in a particular gene expression level, because research subjects are grouped to only two categories, either healthy or severely affected by some mental disorders that may have abnormal working memory as a symptom. Under the RDoC, the entire spectrum rather than just the extremes is supposed to be studied. This population-wide research provides us with better insight to understand the mechanism of mental illnesses and to define what is pathological, and to more accurately classify mental disorders.

The second principle is that in implementing the RDoC, one in theory needs to ignore the current classification of mental disorders. The goal of the RDoC is not to map the neurobiological characteristics onto the currently defined disorders, but onto the mental functional dimensions, which are the constructs in the RDoC matrix. It can be viewed that the spirit of the RDoC is to break down the currently defined mental disorders into their more basic components. For instance, both schizophrenic or schizoaffective patients may have cognitive problems, so in a study to find out the neural circuits associated with cognition, patients with both DSM diagnoses can be enrolled. The purpose of such a study would be to focus on cognition rather than schizophrenia or schizoaffective disorder. If enough research can be done following this second principle, ultimately patients can be treated more accurately for the specific construct that is not normally functioning, instead of being given a vaguely defined diagnosis only.

The last principle is that different units of analysis can be used. As stated in NIMH RDoC and illustrated in Figure 1.1, units of analysis can include but are not limited to genes, molecules, cells, neural circuits, physiology, behaviors or self-reports. Researchers can choose different columns in the RDoC matrix as independent variables, depending on the study goal. In the NIMH RDoC, an example is given where fear circuitry is studied. The independent variable in the example study would be the extent of response to fearful stimuli using a measure such as amygdala response and the dependent variable could be some symptom measures on fear and stress.

1.2.2 The RDoC and the DSM

In summary, even though the RDoC also refers to clinical symptoms as the DSM does, it is quite different from the DSM. The diagnostic criteria in DSM use symptom clusters and have been agreed by experts from different fields on a pragmatic perspective, although the subjects in a DSM study can be very heterogeneous in symptoms. The impact of the DSM is still substantial as it drives virtually all of the clinical psychiatry. For example, some treatments have been developed to target on a particular DSM diagnosis. The RDoC, on the other hand, focuses on a particular mental functional dimension called construct and intends to cover the gap between neurobiological findings and clinical symptoms around this particular construct from the etiological point of view. Because the RDoC tries to understand the mechanisms how brain disruptions lead to illness in a specific construct on a dimensional basis, the unhealthy subjects in an RDoC study should all have some symptoms relevant to the construct of interest, regardless of the DSM diagnoses they are categorized into. For example, if an RDoc study intends to understand the role a protein plays in working memory, the unhealthy subjects should have varying degrees of working memory problems, no matter what DSM diagnoses they have. Following the spirit of the RDoC, people with mental illness will be treated in the future precisely for a particular symptom rather than a DSM diagnosis. It is to be emphasized that we are still at the very beginning, and RDoC is introduced by the NIMH at the present time as a framework for research purposes only. The RDoC is not expected to replace the DSM any time soon.

1.3 MOTIVATION FOR THIS DISSERTATION

In spite of the promising future the RDoC portrays, it is difficult, if not impossible, to carry out all the research studies in the RDoC spirit for the time being. This is particularly applicable for post-mortem tissue studies. One of the difficulties is due to the lack of representativeness of the existing post-mortem tissue databases. Recall that the RDoC advocates research to examine the underlying biological mechanism for mental illness over the general population and ignores the current DSM diagnosis. However, in the post-mortem tissue databases, subjects either are healthy or have symptoms severe enough to manifest a DSM diagnosis. In other words, subjects in the post-mortem tissue databases are on the two extremes over the spectrum of some mental function. Subjects from the in between population are still missing. Another difficulty of using the post-mortem tissue databases is that although neurobiological characteristics are measured for the subjects in these databases, there are no measures of mental functions on these subjects.

The question arises how to use existing post-mortem tissue data to do studies that have an RDoC flavor. This is what motivates this dissertation. In fact the NIMH provides grant opportunities that "seeks applications which propose secondary analyses of existing clinical research datasets to investigate constructs identified in the NIMH's Research Domain Criteria (RDoC) initiative and to test novel hypotheses using the RDoC framework". Therefore using existing post-mortem tissue databases aligns well with what the RDoC Initiative is looking for.

Post-mortem brain tissues are those brain tissues available for neuroscience studies after the death of the subject. In each post-mortem tissue database, each subject is determined either to be healthy or to have a DSM diagnosis based on the behavior when he or she was still alive. Their brain tissue samples are then studied and provide information about the mental illness that cannot be studied in living patients. For example, brain lesions can be inspected to identify the brain regions connected with a certain mental disorder. In particular, neurobiological characteristics such as gene expression levels and protein levels can be measured using these tissue samples. Through comparing the post-mortem tissue samples from healthy people and from subjects with mental disorders, researchers have the opportunities to discover the neurobiological characteristics associated with a mental illness and to explore the underlying biological mechanism for this mental illness, which is exactly what the RDoC advocates.

Even though the post-mortem tissue databases have been available to researchers for decades and allowed investigation of some biological mechanism of mental illness, they have not been used in the direction as the RDoC guides. Currently post-mortem tissue studies are carried out focusing on a particular DSM diagnosis. For example, researchers have been trying to find out the pathogenic genes for schizophrenia using post-mortem tissues. However, due to the limitations of the DSM diagnostic system, subjects in current post-mortem tissue studies are heterogeneous in symptoms. Therefore, it is difficult for current studies to lead to accurate treatments for all the subjects with this particular DSM diagnosis.

In this dissertation, we use the post-mortem tissue databases from the perspective that the RDoC expects to achieve the most from these databases. We would like to learn what and how the neurobiological characteristics function in a particular construct. Therefore, instead of looking at just one DSM diagnosis, more than one post-mortem tissue database spanning multiple DSM diagnoses can be used, as long as the DSM diagnosis involves symptoms related to the construct of interest. Although these post-mortem tissue databases label each subject with a DSM diagnosis such as schizophrenia or bipolar disorder, when combined together properly, these databases can be viewed as representing a random sample from the population with dysfunction in the construct of interest. In other words, in this dissertation, the original population with dysfunction in the construct of interest is replaced by the combination of several populations each with a different DSM diagnosis related to the construct. Subjects from the post-mortem tissue databases for these DSM diagnoses, if constructed appropriately, form a random sample of the original population. Through combining these post-mortem tissue databases, interesting scientific and statistical questions can be raised and answered in the RDoC spirit. By using the collective sample, we are able to draw inferences about the original population. For example, researchers might want to study some neurobiological characteristics that are possibly involved in psychosis. Here the population with psychosis is our original population with dysfunction and the populations with DSM diagnoses such as schizophrenia and bipolar disorder are used together to represent it because these DSM diagnoses involve psychotic symptoms. While there are numerous neurobiological characteristics measured in each subject in this psychosis study, methods are to be developed on how to select those that are significantly associated with psychosis by using databases of schizophrenia and bipolar disorder. Furthermore, error rates need to be protected in the selection process. After these significant neurobiological characteristics are identified, researchers might be interested in the mechanism through which these neurobiological characteristics are involved with psychosis and it is natural to ask in what aspects of the neurobiological characteristics the healthy and the psychotic populations differ from each other. Also investigators might be interested in how to design an experiment to study these neurobiological characteristics with limited number of post-mortem subjects and with limited study budget.

1.4 OVERVIEW OF DISSERTATION

The rest of the dissertation is organized as follows. In Chapter 2, we propose a method to identify the common significant neurobiological characteristics across DSM diagnoses while controlling the false discovery rate (FDR). Motivation for the identification problem is elaborated in Section 2.1. The identification is formulated statistically as a multiple hypothesis testing problem in Section 2.2. In the formulation, we first talk about some assumptions for the problem (Section 2.2.1) and the data structure (Section 2.2.2), and then discuss how to perform hypothesis testing on each individual neurobiological characteristic across the multiple DSM diagnoses (Section 2.2.3). Because there are more than one neurobiological characteristic being tested, multiplicity adjustments must be made and we choose to control the FDR using the BH procedure introduced in Benjamini & Hochberg (1995) (Section 2.2.4). The BH procedure is shown in Benjamini & Yekutieli (2001) to work for dependent test statistics if they follow a specific positive dependence structure called PRDS. We give a brief literature review on positive dependence structures in Section 2.2.5 and state the proposed method to identify the common significant neurobiological characteristics across multiple DSM diagnoses in Section 2.2.6. As the m neurobiological characteristics that are measured on the same subject could be correlated, the dependence structure of the m test statistics need to be examined to see if PRDS is satisfied. We are able to show that the test statistics do not follow PRDS. However, a careful examination of Benjamini & Yekutieli (2001) reveals dependence structure weaker than PRDS can also lead to the protection of

FDR by the BH procedure. These results are shown in Section 2.3.1. Being unable to prove the dependence structure of the test statistics leads to FDR protection, we show through simulation studies in Section 2.3.2 that in our setting, using the proposed method to identify the common significant neurobiological characteristics, the FDR is protected under different configuration of parameters. Finally a case study is presented in Section 2.3.3 using the proposed method and the result is compared with that obtained through another alternative method.

After the common significant neurobiological characteristics are identified, we would like to compare them between the population with dysfunction in the construct of interest and the healthy population. Specifically, we would like to design a study that estimates the difference in the mean as well as the quantiles of the neurobiological characteristic between the two populations. In Chapter 3, we propose the comparisons through the mean and the quantile between the two populations using a matched sample study. The motivation for the comparisons is discussed in Section 3.1. We talk about the comparison through the means of the neurobiological characteristics in Section 3.2 and the comparison through the quantiles in Section 3.3. For the comparison through the means, we first lay out the problem (Section 3.2.1) and then discuss some assumptions about the study design we are going to use (Section 3.2.2). We then present the form of the study design which we call triangular design under our assumptions (Section 3.2.3). Finally we state the model for the comparison through the means and provide an estimator for the difference in means between the two populations (Section 3.2.4). For the comparison through the quantiles, we also discuss some assumptions (Section 3.3.1) and the layout of the problem (Section 3.3.2) and then propose a model to estimate the difference in quantiles between the population with dysfunction and the healthy population (Section 3.3.3). A simulation study is run in Section 3.3.4 to show the performance of the proposed model in quantile estimation. Because the estimator of the mean difference between the two populations can be derived with a closed form, simulation studies for the mean estimation are unnecessary. We close the chapter with a summary of the two comparisons in Section 3.4.

The difference in the means or the quantiles between the population with dysfunction and the healthy population can be estimated through the modeling in Chapter 3 when the sample size for each DSM diagnosis is given. However, the sample size for each DSM diagnosis should be determined prior to any experiments. Because the post-mortem tissue databases are used, there are two constraints in the sample size determination: the number of available subjects for each DSM diagnosis and the total number of subjects that the budget allows. In Chapter 4, an algorithm is proposed to determine the optimal sample size for each DSM diagnosis under the above two constraints in order to compare the means between the two populations under the triangular design. The criteria we use for optimality is to have minimum variance of the estimator for the difference in the means between the two populations. In Section 4.1, the motivation for the optimal design problem is reviewed. The layout of the optimal design problem is carefully described in Section 4.2. We provide an illustration of the optimal design problem with a post-mortem database from the Stanley Brain Consortium (Section 4.2.1). We then define the notation we need (Section 4.2.2) and give a hypothetical numerical example to explain the optimal design problem and illustrate the algorithm (Section 4.2.3). Assumptions for the optimal design problem are also discussed in detail (Section 4.2.4). The triangular design with unknown sample sizes are presented in Section 4.3 and the variance formula of the estimator for the difference in means under unknown sample sizes is derived in Section 4.4 based on the model in Section 3.2. In Section 4.5, we elaborate the difficulty of minimizing the variance formula analytically and proposed a method to do the minimization. We also illustrate the minimization with numerical examples. In Section 4.6, we state the proposed algorithm to obtain the optimal sample size for each DSM diagnosis and illustrate it with a numerical example.

In Chapter 5, we summarize the conclusions for this dissertation and discuss some possible work for future research.

Note that in this dissertation, the appendices for each chapter are provided at the end of that chapter so that the contents for each chapter are organized together.

2.0 IDENTIFICATION OF COMMON SIGNIFICANT NEUROBIOLOGICAL CHARACTERISTICS ACROSS DSM DIAGNOSES

This chapter deals with the identification of common significant neurobiological characteristics across multiple DSM diagnoses.

2.1 MOTIVATION FOR THE IDENTIFICATION

As briefly introduced in Chapter 1, the RDoC calls for research to establish the one to one correspondence between brain and behavior, i.e., to link the neurobiological characteristics and a specific psychiatric construct in the RDoC matrix. To advance the research in the RDoC spirit, a first step would be to identify some neurobiological characteristics that are significantly associated with the construct of interest. Once identified, they can be studied further to understand the mechanism of dysfunction in this particular construct and thus provide targets for more accurate treatment of mental illness. What we would like to do is to develop a method to identify the significant neurobiological characteristics for a psychiatric construct.

As discussed in Chapter 1, due to the unavailability of neurobiological measures in living people, existing post-mortem tissue databases are used as an alternative to identify the neurobiological characteristics that are significantly associated with a construct. Because post-mortem tissue databases label subjects with DSM diagnoses, any DSM diagnosis, as long as it has symptoms related to the construct of interest, can be included in the study. And any post-mortem tissue databases for the included DSM diagnoses can be used. For example, if we would like to identify the genes significantly associated with psychosis, any post-mortem tissue database for schizophrenia or bipolar disorder can be used because patients in both DSM diagnoses may have psychosis. The neurobiological characteristics that are significant in all the included DSM diagnoses are the ones we want to identify as significant in the construct of interest. In other words, these DSM diagnoses are included in the study because they may share some common symptoms in the construct of interest, and the shared symptoms are driven by the neurobiological overlaps, which we think are the neurobiological characteristics significant in the construct. In this dissertation, we call these neurobiological overlaps "common significant neurobiological characteristics" across multiple DSM diagnoses.

Actually researchers have already been trying to explore the neurobiological overlaps underlying the shared clinical symptoms across different DSM diagnoses. And there is already emerging genetic evidence from various types of studies that some neurobiological characteristics are shared by more than one DSM diagnosis. Take schizophrenia and bipolar disorder as an example again. In addition to Owen, Craddock & Jablensky (2007) and Lichtenstein et al. (2009) mentioned in the explanation of the disadvantages of the DSM in Chapter 1, O'Donovan et al. (2008) identified a single nucleotide polymorphism (SNP) within ZNF804A, the gene for zinc finger binding protein 804A, associated with schizophrenia through a genome wide association study. They also found that the evidence of association increased when they added a group of bipolar disorder subjects into the study. Green et al. (2010) reported that a SNP within CACNA1C, the gene encoding the α-1C subunit of the L-type voltage-gated calcium channel, which was found to be associated with bipolar disorder, also conferred risk to schizophrenia and recurrent major depression with similar effect sizes. Numerous other findings have been reported in the literature.

However, these studies were mostly done and analyzed in each DSM diagnosis separately. The significance was first assessed in one DSM diagnosis, and then examined in another DSM diagnosis to see if there is anything significant for both. Unlike most current studies, what we would like to do is to identify the common significant characteristics by considering multiple DSM diagnoses simultaneously, not step by step. The neurobiological measurements from the post-mortem subjects with different DSM diagnoses are jointly analyzed to see which ones significantly differ from the healthy population. From the statistical point of view, we consider the simultaneous method easier than the step by step one, because the error rate for the identification process should be protected and the step by step method needs to address the partition of error rate in each step.

Similar problems have been considered in other medical fields, for instance, cancer studies. For example, Rhodes et al. (2004) identified genes that are significant in more than one cancer type by conducting a large scale meta-analysis using microarray data from 15 different cancers. Ma, Huang & Moran (2009) applied an integrative analysis to data from 7 tumor types and analyzed all the genes together with a logistic regression model. They successfully identified 60 genes that are associated with one or more cancers. However, most of these cancer studies defined a gene as shared by multiple cancers if it is significantly differentially expressed in more than one cancer type. This is different from the problem we would like to address, which is to identify the neurobiological characteristic that is significant in each of the DSM diagnoses considered. The reason we need the significance in each of the DSM diagnoses is that the post-mortem tissue databases with different DSM diagnoses are used collectively to represent the original population with dysfunction in the construct of interest. In order that all the identified neurobiological characteristics are significant in the original construct, significance in each DSM diagnosis should be required. If the "more than one" situation is used here, then it could be that neurobiological characteristic A is significant in DSM diagnoses 1 and 2, and neurobiological characteristic B is significant in DSM diagnoses 3 and 4 if four DSM diagnoses are considered. In such a case, neither of the neurobiological characteristics A and B is the original characteristics we would like to identify. Therefore the identified neurobiological characteristic is required to be significant over all the considered DSM diagnoses in this dissertation.

2.2 FORMULATION OF THE IDENTIFICATION STATISTICALLY

2.2.1 Assumptions for the Identification

As described previously, the underlying neurobiological characteristics extend very broadly. In general, a common significant neurobiological characteristic can be defined as consistently up-regulated, consistently down-regulated or differentially regulated across multiple DSM diagnoses, compared to the healthy population. The consistently up-regulated ones are those with significantly higher means in each DSM diagnosis compared to the healthy controls. Similarly, the consistently down-regulated ones are those with significantly lower means in each DSM diagnosis compared to the healthy controls and the differentially regulated ones are those with significantly different means from the healthy controls in each DSM diagnosis. Within the RDoC framework, it is more reasonable to identify the neurobiological characteristics that function in the same fashion among all the DSM diagnoses studied. In other words, the identified neurobiological characteristics should be either consistently up-regulated or consistently down-regulated. Those that are differentially regulated could have opposite effects in different DSM diagnoses, and thus it is difficult to interpret their significance in the construct of interest. Furthermore, from a technical perspective, identifying the differentially regulated neurobiological characteristics requires using the absolute values as test statistics due to the two-sided tests. The absolute values bring in more complexity because the absolute value of a normal random variable is not normally distributed again. As a result, the "common" here means the consistency of the significance direction of the neurobiological characteristic across all the considered DSM diagnoses. In this dissertation we only focus on the consistently up-regulated neurobiological characteristics as the down-regulated ones can be identified analogously. To simplify the discussion, throughout this chapter, schizophrenia and bipolar disorder are used to clarify ideas because it is already known that these two DSM diagnoses share some common neurobiological characteristics. We note that the method discussed here can be readily extended to more than two DSM diagnoses.

Most of the time, each post-mortem tissue database features in one DSM diagnosis and has its own healthy control group. When we use a schizophrenia database and a bipolar disorder database, there are four groups considered in total: schizophrenia subjects (S), bipolar subjects (B), schizophrenia control subjects (SC) and bipolar control subjects (BC). The subjects are assumed to be unpaired so there can be different number of subjects in each group. The number of subjects for the S, B, SC and BC group is denoted as n_s, n_b, n_{sc} and n_{bc} respectively. The same set of m neurobiological characteristics are measured on every subject in each of these four groups. Covariates such as age or baseline measurements are not considered in the unpaired model here. While measurements of any two characteristics within the same subject are correlated, it is assumed that subjects are independent of each other. In cases where the subject with a DSM diagnosis is paired with a healthy control subject using covariates like age and gender, the measurements on the subjects within each pair are not independent, and thus a paired test can be used accordingly, as shown in later sections.

2.2.2 Data Structure for the Identification

Throughout this dissertation Y is used to denote the measurement. Y_{ij}^d means the measurement of neurobiological characteristic i for subject j in group d, $i = 1, \cdots, m; j = 1, \cdots, n_d; d = s, b, sc, bc$. The data structure can be laid out as the following:

$$\begin{array}{c} 1 \\ 2 \\ \vdots \\ m \end{array}
\overbrace{\begin{array}{ccc} Y_{11}^s & \cdots & Y_{1n_s}^s \\ Y_{21}^s & \cdots & Y_{2n_s}^s \\ \vdots & \vdots & \vdots \\ Y_{m1}^s & \cdots & Y_{mn_s}^s \end{array}}^{\text{S},n_s\text{ subjects}}
\overbrace{\begin{array}{ccc} Y_{11}^{sc} & \cdots & Y_{1n_{sc}}^{sc} \\ Y_{21}^{sc} & \cdots & Y_{2n_{sc}}^{sc} \\ \vdots & \vdots & \vdots \\ Y_{m1}^{sc} & \cdots & Y_{mn_{sc}}^{sc} \end{array}}^{\text{SC},n_{sc}\text{ subjects}}
\overbrace{\begin{array}{ccc} Y_{11}^b & \cdots & Y_{1n_b}^b \\ Y_{21}^b & \cdots & Y_{2n_b}^b \\ \vdots & \vdots & \vdots \\ Y_{m1}^b & \cdots & Y_{mn_b}^b \end{array}}^{\text{B},n_b\text{ subjects}}
\overbrace{\begin{array}{ccc} Y_{11}^{bc} & \cdots & Y_{1n_{bc}}^{bc} \\ Y_{21}^{bc} & \cdots & Y_{2n_{bc}}^{bc} \\ \vdots & \vdots & \vdots \\ Y_{m1}^{bc} & \cdots & Y_{mn_{bc}}^{bc} \end{array}}^{\text{BC},n_{bc}\text{ subjects}}$$

We assume for subject j in group d, the measurements $(Y_{1j}^d, Y_{2j}^d, \cdots, Y_{mj}^d)$ follow a m-dimensional multivariate normal distribution which is denoted by $\mathbf{N_m}(\boldsymbol{\mu_y^d}, \boldsymbol{\Sigma_y^d})$, where

$$\boldsymbol{\mu_y^d} = (\mu_{y1}^d, \mu_{y2}^d, \cdots, \mu_{ym}^d).$$

Also all the subjects in the same group are assumed to share a common covariance matrix. Let ρ^d_{yih} denote the correlation coefficient between Y^d_{ij} and Y^d_{hj}, then the covariance matrix $\boldsymbol{\Sigma^d_y}$ can be written as:

$$\boldsymbol{\Sigma^d_y} = \begin{pmatrix} {\sigma^d_{y1}}^2 & \sigma^d_{y1}\sigma^d_{y2}\rho^d_{y12} & \cdots & \sigma^d_{y1}\sigma^d_{ym}\rho^d_{y1m} \\ \sigma^d_{y1}\sigma^d_{y2}\rho^d_{y12} & {\sigma^d_{y2}}^2 & \cdots & \sigma^d_{y2}\sigma^d_{ym}\rho^d_{y2m} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma^d_{y1}\sigma^d_{ym}\rho^d_{y1m} & \sigma^d_{y2}\sigma^d_{ym}\rho^d_{y2m} & \cdots & {\sigma^d_{ym}}^2 \end{pmatrix}, \tag{2.2.1}$$

where $\sigma^d_{yi}(i = 1, \cdots, m; d = s, b, sc, bc)$ is the standard deviation for Y^d_{ij}.

In a paired study, $n_s = n_{sc} = n_{sp}, n_b = n_{bc} = n_{bp}, Y^s_{ij}$ is paired with Y^{sc}_{ij} and Y^b_{ij} is paired with Y^{bc}_{ij}. The difference between Y^s_{ij} and Y^{sc}_{ij} and that between Y^b_{ij} and Y^{bc}_{ij} is used in this case. Let $X^s_{ij} = Y^s_{ij} - Y^{sc}_{ij}$ and $X^b_{ij} = Y^b_{ij} - Y^{bc}_{ij}$, so that for $f \in \{s, b\}$, the vector $(X^f_{1j}, X^f_{2j}, \cdots, X^f_{mj})$ follows a m-dimensional multivariate normal distribution denoted by $\mathbf{N_m}(\boldsymbol{\mu^f_x}, \boldsymbol{\Sigma^f_x})$. Here $\boldsymbol{\mu^f_x} = (\mu^f_{y1} - \mu^{fc}_{y1}, \mu^f_{y2} - \mu^{fc}_{y2}, \cdots, \mu^f_{ym} - \mu^{fc}_{ym})$ and $\boldsymbol{\Sigma^f_x}$ can be written as:

$$\boldsymbol{\Sigma^f_x} = \begin{pmatrix} {\sigma^f_{x1}}^2 & \sigma^f_{x1}\sigma^f_{x2}\rho^f_{x12} & \cdots & \sigma^f_{x1}\sigma^f_{xm}\rho^f_{x1m} \\ \sigma^f_{x1}\sigma^f_{x2}\rho^f_{x12} & {\sigma^f_{x2}}^2 & \cdots & \sigma^f_{x2}\sigma^f_{xm}\rho^f_{x2m} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma^f_{x1}\sigma^f_{xm}\rho^f_{x1m} & \sigma^f_{x2}\sigma^f_{xm}\rho^f_{x2m} & \cdots & {\sigma^f_{xm}}^2 \end{pmatrix}, \tag{2.2.2}$$

where $\sigma^f_{xi}(i = 1, \cdots, m; f = s, b)$ is the standard deviation for X^f_{ij} and $\rho^f_{xih}(i = 1, \cdots, m; h = 1, \cdots, m; f = s, b)$ is the correlation between X^f_{ij} and X^f_{hj}.

The process for identifying the neurobiological characteristics significantly up-regulated in both DSM diagnoses can be viewed as a multiple hypothesis testing problem. For each individual characteristic, the null hypothesis is: in at least one DSM diagnosis, the mean of the neurobiological characteristic is less than or equal to that in the corresponding healthy control group and the alternative hypothesis is: the mean of the neurobiological characteristic is greater than that in the corresponding healthy control group for both DSM diagnoses. The decision for a single neurobiological characteristic is based on two test statistics, one for each DSM diagnosis. Collectively, since there is more than one characteristic being tested, adjustments must be made for multiplicity.

2.2.3 Hypothesis Testing for each Single Neurobiological Characteristic

For characteristic i, in both the unpaired and paired cases, the null hypothesis and alternative hypothesis can be formulated as:

$$\begin{array}{l} H_{0i}: \mu_{yi}^{s} - \mu_{yi}^{sc} \leq 0 \text{ or } \mu_{yi}^{b} - \mu_{yi}^{bc} \leq 0 \\ H_{ai}: \mu_{yi}^{s} - \mu_{yi}^{sc} > 0 \text{ and } \mu_{yi}^{b} - \mu_{yi}^{bc} > 0 \end{array}, \quad i = 1, 2, \cdots, m. \tag{2.2.3}$$

It's easy to see that the null hypothesis for each characteristic is the union of two sub-null hypotheses: $H_{0i}^{s}: \mu_{yi}^{s} - \mu_{yi}^{sc} \leq 0$ and $H_{0i}^{b}: \mu_{yi}^{b} - \mu_{yi}^{bc} \leq 0$. The alternative hypothesis is the intersection of the complements of the two sub-null hypotheses: $H_{ai}^{s}: \mu_{yi}^{s} - \mu_{yi}^{sc} > 0$ and $H_{ai}^{b}: \mu_{yi}^{b} - \mu_{yi}^{bc} > 0$. Both sub-null hypotheses are one-sided. For the null hypothesis H_{0i} to be rejected at level α, both sub-null hypotheses need to be rejected at level α. In other words, characteristic i gets identified if and only if its mean measurements in both DSM diagnoses are significantly higher than that in the corresponding healthy control groups.

One method that can be applied here directly is the Laska's min test procedure. Laska & Meisner (1989) developed the test to detect if a treatment is the best among several candidates. Applying the method in our setting, when the covariance matrix is assumed to be known as in (2.2.1) for the unpaired case, we have that the test statistic W_i for characteristic i is:

$$W_i = \min(U_i, V_i), \quad \text{where } U_i = \frac{\bar{Y}_i^{s} - \bar{Y}_i^{sc}}{\sqrt{\frac{\sigma_{yi}^{s\;2}}{n_s} + \frac{\sigma_{yi}^{sc2}}{n_{sc}}}}, V_i = \frac{\bar{Y}_i^{b} - \bar{Y}_i^{bc}}{\sqrt{\frac{\sigma_{yi}^{b\;2}}{n_b} + \frac{\sigma_{yi}^{bc2}}{n_{bc}}}}. \tag{2.2.4}$$

The $\bar{Y}_i^{d}(d = s, sc, b, bc)$ in (2.2.4) is the sample mean of $Y_{i1}^{d}, Y_{i2}^{d}, \cdots, Y_{in_d}^{d}$. In a paired study, the difference X within a pair is used. Assume the covariance matrix is known as in (2.2.2) and let $\bar{X}_i^{f}(f = s, b)$ denote the sample mean of $X_{i1}^{f}, X_{i2}^{f}, \cdots, X_{in_{fp}}^{f}$, then the test statistic of the min test can be written as:

$$W_i = \min(U_i, V_i), \quad \text{where } U_i = \frac{\bar{X}_i^{s}}{\sqrt{\frac{\sigma_{xi}^{s\;2}}{n_{sp}}}}, V_i = \frac{\bar{X}_i^{b}}{\sqrt{\frac{\sigma_{xi}^{b\;2}}{n_{bp}}}}. \tag{2.2.5}$$

It is easy to see that U_i and V_i are the test statistics for testing the sub-null hypotheses H_{0i}^{s} and H_{0i}^{b}, respectively and they are independent normally distributed random variables. The joint distribution of $(U_1, U_2, \cdots, U_m)$ and that of $(V_1, V_2, \cdots, V_m)$ are derived in Appendix 2.A.1. In both the unpaired and paired cases, under the null hypothesis, U_i is distributed as $N(0,1)$ or V_i is distributed as $N(0,1)$. The null hypothesis H_{0i} is rejected at level α if $W_i \geq \Phi^{-1}(1-\alpha)$, where Φ^{-1} is the inverse cumulative distribution function for univariate standard normal distribution. The p-value of the test is $P_i = 1 - \Phi(W_i)$. Since the min test uses the minimum of the two test statistics U_i and V_i, it is equivalent as saying both sub-null hypotheses being rejected is the same as rejecting hypothesis H_{0i}. In the context of identifying the common significant neurobiological characteristics in schizophrenia and bipolar disorder, only when the mean of a neurobiological characteristic is significantly higher than corresponding control in both DSM diagnoses, will it be identified as the desired one.

Even though Laska & Meisner (1989) proved the result only in the case of two treatments, the min test, as these authors noted, can be used when there are more than two treatments. In order to identify the common significant neurobiological characteristics across k DSM diagnoses where $k \geq 2$, the min test still works by taking the minimum over the test statistics for each of the k sub-null hypotheses. Moreover, even if the DSM diagnoses have a common control group so that the test statistics for each of the sub-null hypotheses are not independent, the min test can still deal with it. For example, in one post-mortem tissue database there are both schizophrenia patients and bipolar disorder patients and both DSM diagnoses are compared with the same control group. In such a case, the $\bar{Y}_i^{sc}$ and $\bar{Y}_i^{bc}$ in (2.2.4) are the same random variable and $\frac{\sigma_i^{sc2}}{n_{sc}}$ and $\frac{\sigma_i^{bc2}}{n_{bc}}$ are equal, thus U_i and V_i have a common element in their definitions and become dependent. The min test still works for such cases as shown in Laska & Meisner (1989) since the test statistics for the sub-null hypotheses do not need to be independent. Also among all monotone test statisitcs, the min test is the UMP test even though it has low power sometimes, especially when the effect size is close to the origin.

2.2.4 FDR Control over m Neurobiological Characteristics

As noted previously, more than one characteristic is tested simultaneously, and therefore, multiplicity adjustments are needed. Usually when studying a number of neurobiological characteristics like what is described in this dissertation, the goal is to identify as many significant ones as possible. Thus, the usual control of familywise error rate (FWER) is considered to be too conservative, even though it is a standard method for multiplicity adjustments. One approach to handle the multiplicity issue that has been used more recently is to control the false discovery rate (abbreviated as FDR and formally introduced in Benjamini & Hochberg (1995)), which is the expected proportion of falsely rejected hypotheses.

When testing m hypotheses, the results can be summarized as in Table 2.1:

Table 2.1: Results of multiple testing

	not rejected	rejected	total
true null	u	v	m_0
false null	t	s	$m - m_0$
	$m - r$	r	m

Here in the table, m_0 is the number of true null hypothesis. u is the number of null hypothesis that are correctly not rejected, v is the number of false positives, i.e., true null hypothesis that is rejected, t is the false negatives, i.e., false null hypothesis that should have been rejected, and s is number of the correctly rejected null hypothesis. r is the sum of v and s, which is the number of rejections. It's worth noting that only r and m are observed when testing these m hypotheses. All the other values in Table 2.1 are unknown.

The FDR is defined as: FDR=$E(v/r)I_{r>0}$. From the definition, we can see that controlling FDR does not require v to be small, nor does it require the proportion v/r to be small. Only the expected value of v/r needs to be controlled at the desired level. This is very different from the control of FWER, which requires $P(v \geq 1)$ to be less than or equal to a certain level. FWER tends to be more stringent as attention is paid to whether any type I error is made or not, so making one type I error is viewed the same as making several type I errors. However, in FDR controlling, committing more than one type I error is allowed as long as the expected value of the false positive proportion is within a certain level. Because more rejections can be made with less worry about type I errors, FDR controlling tends to have a larger power than FWER controlling in hypothesis testing. Therefore in our case, FDR control is preferred since we would like to identify as many significant neurobiological characteristics as possible.

There have been a number of FDR controlling procedures developed in the literature. At first, methods for independent test statistics were developed. Benjamini & Hochberg (1995) proposed a linear step-up procedure (BH) to control the FDR at level α. Specifically, they showed the BH procedure controls the FDR at level $\frac{m_0}{m}\alpha$ under independent test statistics, where m_0 and m are defined in Table 2.1. Since $m_0 \leq m$, the FDR using the BH procedure is controlled at the α level. Later on, Benjamini & Liu (1999) proposed a step-down procedure which is shown to control the FDR at the desired level under independent test statistics. They showed through simulation that the step-down procedure is more powerful than the BH procedure when the number of tests is small and the effect sizes of the false null hypotheses are large. Benjamini & Hochberg (2000) improved the BH procedure in an adaptive way by combining the BH method with estimating the number of true null hypotheses. They showed that if the test statistics are independent, this adaptive method controls the FDR exactly at the α level even when the number of true null hypotheses is smaller than the number of tests, i.e., when $m_0 \leq m$. Ghosh (2011) generalized the BH procedure by using spacings of the p-values, which is the distance between neighboring ordered p-values. In what they termed as generalized BH procedure, a suitably selected monotone function was applied to the spacings of p-values and a step-up method was implemented on the dimension of the transformed spacings. He proved that if the p-values are independent, the generalized BH procedure controls the FDR at level α.

More recently, research has been done for FDR controlling methods for dependent test statistics. Benjamini & Yekutieli (2001) reexamined the BH procedure proposed in 1995. They showed if the test statistics follow a specific positive dependence structure termed PRDS conditioning on the true null hypotheses, then the step-up procedure still controls FDR at the desired level. Sarkar (2002) strengthened the work of Benjamini & Yekutieli (2001) and showed the FDR is controlled by a general step-up-step-down procedure of order r with the same set of critical values as in the BH procedure. He showed if the test statistics are independent or have PRDS conditioning on the true null hypotheses, then FDR is protected. Yekutieli (2008) modified the BH procedure to control FDR when the test statistics are not positively dependent. He proposed a seperate subset BH procedure (ssBH) and pointed out that ssBH is less powerful than the BH procedure in general, but if the p-values satisfy a certain condition, then the FDR using the ssBH procedure is controlled at level $\frac{m_0}{m}\alpha$. To apply the ssBH procedure, the p-values are divided into S sub-vectors $\mathbf{P}^s, s = 1, \cdots, S$. The special condition required is that for each $P_i \in \mathbf{P}_0$, where $\mathbf{P}_0$ is the set of p-values corresponding to the true null hypotheses, the union of sub-vectors containing P_i, i.e., $\cup\{\mathbf{P}^s : P_i \in \mathbf{P}^s\}$ is PRDS conditioning on P_i.

Furthermore, there are some Bayesian methods to control FDR. For instance, Storey (2002) studied FDR control from a Bayesian perspective. He introduced the concept of positive FDR (pFDR). Storey (2003) continued his work on pFDR and provided a Bayesian interpretation for FDR control. He investigated the advantages and disadvantages of pFDR and provided Bayesian interpretation of q-values, which are the pFDR analogues of p-values.

Even though there are lots of ways to control the FDR at level α, what is needed in our setting is a method that works for dependent non-normal test statistics since the minimum W_i for each single test is not normally distributed and W_i is correlated with W_j due to the correlation among the original observations within the same subject. The Bayesian methods are not considered because they fix the rejection region and estimate the FDR and involve other parameters. Among the frequentist's methods, BH is a commonly used simple method and it works for dependent non-normal statistics.

The BH procedure is based on the ordered p-values, $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$, from the individual tests. Let $H_{0(i)}$ be the null hypothesis corresponding to p-value $p_{(i)}$, then the BH procedure is defined as:

$$\text{Reject all } H_{0(i)}, 1 \leq i \leq l, \text{ where } l = max\{i : p_{(i)} \leq \frac{i}{m}\alpha\}. \tag{2.2.6}$$

For example, if we start from the largest p-value and find $p_{(5)}$ is the first p-value that satisfies $p_{(i)} \leq \frac{i}{m}\alpha$, then all of $H_{0(1)}, H_{0(2)}, \cdots, H_{0(5)}$ are rejected. If no such l is found, then no null hypothesis is rejected.

2.2.5 Positive Dependence Structures

The BH procedure was proved in Benjamini & Yekutieli (2001) to control the FDR at the α level when the test statistics follow a dependence structure called PRDS conditioning on the test statistics corresponding to the true null hypotheses. Benjamini & Yekutieli (2001) defined PRDS as the following.

Definition 2.1 (Increasing set). *A set $D \in \mathbb{R}^n$ is called an increasing set if for any $x \in D$, $y \geq x$ implies $y \in D$, where "$\geq$ " is the usual ordering on $\mathbb{R}^n$.*

Definition 2.2 (PRDS). *For any increasing set D and each $i \in I_0$, where $I_0 \subseteq \{1, 2, \cdots, n\}$ is a sub-index set, the random vector $\mathbf{X} = (X_1, X_2, \cdots, X_n)$ is said to follow PRDS conditioning on the test statistics corresponding to I_0 if $P(\mathbf{X} \in D | X_i = x_i)$ is a nondecreasing function of x_i.*

While PRDS is a special dependence structure used by Benjamini & Yekutieli (2001), in the literature, there are various notions of positive dependences proposed for random variables. For bivariate variables, Lehmann (1966) introduced several concepts of dependence. Among these concepts, PQD, which was defined as the inequality $P(X \leq x, Y \leq y) \geq P(X \leq x)P(Y \leq y)$ holding for all x, y, was the weakest condition. A stronger concept was positively regression dependence (PRD) where a random variable Y was PRD on X if $P(Y \leq y|X = x)$ was non-increasing in x for all y. Without formally introducing the concept of left tail decreasing (LTD), which requires $P(Y \leq y|X \leq x)$ to be non-increasing in x for all y, Lehmann (1966) mentioned that LTD was between PRD and PQD in strength. In 1972, Esary & Proschan (1972) introduced the concept of right tail increasing (RTI) and studied the relationships among a number of bivariate dependence concepts. It defined RTI as $P(Y > y|X > x)$ being non-decreasing in x for all y.

Multivariate dependence is more complicated than the bivariate case. For example, Joe (1997) gave two multivariate extensions of the bivariate dependence PRD. One is called positive dependent through stochastic ordering (PDS) and the other is called conditional increasing in sequence (CIS). PDS is satisfied if for any $i = 1, 2, \cdots, n$, the conditional joint distribution of $(X_1, \cdots, X_{i-1}, X_{i+1}, \cdots, X_n | X_i = x)$ is stochastically increasing as x increases. CIS is satisfied if the conditional distribution of X_i given $X_{i-1} = x_{i-1}, \cdots, X_1 = x_1$, i.e.,

$P(X_i > x_i | X_{i-1} = x_{i-1}, \cdots, X_1 = x_1)$ is increasing in $x_{i-1}, \cdots, x_1$ for $i = 2, 3, \cdots, n$. Dykstra, Hewett & Thompson (1973) generalized the PQD defined in Lehmann (1966) to the multivariate case. They defined $(X_1, X_2, \cdots, X_n)$ to be positively orthant dependent if $P(X_i \leq x_i, i = 1, 2, \cdots, n) \geq \prod_{i=1}^{n} P(X_i \leq x_i)$. Shaked (1982) introduced the concepts of positively upper orthant dependence (PUOD) and positively lower orthant dependence (PLOD). For a random vector $\mathbf{X} = (X_1, X_2, \cdots, X_n)$, if for every $\mathbf{x} = (x_1, x_2, \cdots, x_n), P(\mathbf{X} > \mathbf{x}) \geq (\leq) \prod_{i=1}^{n} P(X_i > x_i)$ then $\mathbf{X}$ is PUOD (PLOD). The PLOD defined here is the same as the positively orthant dependent defined in Dykstra, Hewett & Thompson (1973). Alzaid & Proschan (1994) extended the bivariate notion of LTD and RTI to the multivariate case.

For multivariate random variables, the notions PUOD and PLOD are not equivalent. However, if $n = 2$, they are the same and reduce to PQD. Also the notions of CIS, mPRD and PDS are different but they all reduce to the PRD in Lehmann (1966) when $n = 2$. The implications among those dependence concepts differ depending on whether one is looking at the bivariate case or the general multivariate case. For example, in the multivariate case, PDS implies both PUOD and PLOD and in the bivariate case, PRD implies LTD and RTI, both of which imply PQD.

2.2.6 Proposed Method for the Identification

As described previously, our goal is to identify the common significant neurobiological characteristics across several DSM diagnoses. By combining the techniques explained in the preceding subsections, the proposed method is to apply the BH method at level α on the set of p-values obtained from the m min tests. That is to say, a p-value is obtained after applying the min test on each of the m hypotheses and then the BH procedure is implemented on the m p-values to control the FDR at level α. Decisions (to reject a null hypothesis or not) are made for each hypothesis.

2.3 RESULTS

2.3.1 Conditions Needed to Control FDR by BH Procedure on the Min Test

An important issue of the proposed method is to assess whether or not the FDR can be controlled when applying the BH procedure to the p-values obtained from the min tests. If all the assumptions for the BH procedure are satisfied, for example, the test statistics W_i and W_j are independent, Benjamini & Hochberg (1995) proved that the FDR would be protected automatically. However, due to the correlation among multiple measurements on the same individual, $\bar{Y}_i^s$ and $\bar{Y}_h^s$ as well as $\bar{Y}_i^{sc}$ and $\bar{Y}_h^{sc}$ are correlated in (2.2.1). For the same reason, in the paired case, $\bar{X}_i^s$ and $\bar{X}_h^s$ are correlated in (2.2.2). So U_i and U_h are correlated. Similarly, V_i and V_h are correlated. As a result, the test statistics W_i and W_h are not independent. It is even unclear what the particulars of the dependence structure for $(W_1, W_2, \cdots, W_m)$ is. Because Benjamini & Yekutieli (2001) proved that if the test statistics follow the PRDS dependence conditioning on the test statistics corresponding to the true null hypotheses (PRDS on the true nulls), then the BH procedure controls FDR at the desired level, it is reasonable to first check if $(W_1, W_2, \cdots, W_m)$ follows PRDS on the true nulls.

2.3.1.1 PRDS and Min Statistics Suppose m hypotheses are tested, among which m_0 are true. Also the FDR needs to be controlled at the q level. Theorem 1.2 in Benjamini & Yekutieli (2001) states the following.

Result 2.1. *If the joint distribution of the test statistics for the m hypotheses follows PRDS on the true nulls, the FDR using the BH procedure can be controlled at level less than or equal to $\frac{m_0}{m}q$.*

Let $\mathbf{W}$ be the resulting vector for testing m hypotheses using the min test, i.e., $\mathbf{W} = (W_1, W_2, \cdots, W_m)$, where $W_i (i = 1, 2, \cdots, m)$ is defined in (2.2.4). Replacing the test statistics in Result 2.1 with $\mathbf{W}$, we can see that if $\mathbf{W}$ were to satisfy the PRDS on the true nulls in our setting, then the FDR is protected by the BH procedure. However, we show that the PRDS dependence is not satisfied by $\mathbf{W}$ on the true nulls.

Result 2.2. *The test statistics for the min test $\mathbf{W}$ does not follow PRDS on the true nulls.*

Proof. If **W** follows PRDS on the true nulls, it is known from Esary & Proschan (1972) that it would follow weaker dependence structures LTD and RTI conditioning on the true nulls when there are only two tests, i.e., m=2. The converse is, if LTD or RTI on the true nulls is not satisfied for the two tests case, PRDS on the true nulls cannot be satisfied in general.

In order to show that (W_1, W_2) is not LTD on the true null, the following numerical example is used. By definition, if (W_1, W_2) were LTD on the true null, $\mathrm{P}(W_1 \leq t|W_2 \leq s)$ is a non-increasing function of s for any t, where W_2 corresponds to the true null hypothesis. Recall from Section 2.2.3 that in our setting, if the null hypothesis is true, then at least one of the sub-null hypotheses is true. Using the formulated hypothesis in (2.2.3), it means that $\mu^s_{y2} - \mu^{sc}_{y2}$, which is the mean of U_2, is less than or equal to 0 or $\mu^b_{y2} - \mu^{bc}_{y2}$ which is the mean of V_2, is less than or equal to 0.

Now suppose (U_1, U_2) follows a bivariate normal distribution with mean $(0, 0, 1, 1, 0.01)$, where $(0, 0)$ is the mean vector of $(U_1, U_2), (1, 1)$ is the variance vector and 0.01 is the correlation between U_1 and U_2. Also suppose (V_1, V_2) follows a bivariate normal distribution $(0, 1, 1, 1, 0.4)$ and (U_1, U_2) and (V_1, V_2) are independent. Because the mean of U_2 is 0, the null hypothesis that W_2 is used to test is true. The probability $P(W_1 \leq t|W_2 \leq s)$ is computed according to the following formula for $t = 0$.

$$\begin{aligned}
P(W_1 \leq t|W_2 \leq s) &= \frac{P(W_1 \leq t, W_2 \leq s)}{P(W_2 \leq s)} \\
&= \frac{1 - P(W_1 > t) - P(W_2 > s) + P(W_1 > t, W_2 > s)}{1 - P(W_2 > s)} \\
&= 1 - \frac{P(U_1 > t, V_1 > t) - P(U_1 > t, V_1 > t, U_2 > s, V_2 > s)}{1 - P(U_2 > s, V_2 > s)} \\
&= 1 - \frac{P(U_1 > t)P(V_1 > t) - P(U_1 > t, U_2 > s)P(V_1 > t, V_2 > s)}{1 - P(U_2 > s)P(V_2 > s)}.
\end{aligned}$$

The values of the probability for selected s's are computed through the functions "pnorm " and "pmnorm " in the R package "mnormt ". Also double precision is used with the R package "Rmpfr ". They are given in Table 2.2 below:

From Table 2.2, we can see that as s increases, the probability $P(W_1 \leq 0|W_2 \leq s)$ first increases and then decreases, so that it is not a monotone function of s. As double precision is used in the computation, the observed non-monotonicity of $P(W_1 \leq 0|W_2 \leq s)$ in s is

Table 2.2: Counter example to show **W** is not LTD

s	$P(W_1 \leq 0 \| W_2 \leq s)$
-2.0	0.766058
-1.8	0.767262
-1.6	0.768487
-1.4	0.769653
-1.2	0.770658
-1.0	0.771389
-0.8	0.771733
-0.6	0.771595
-0.4	0.770910
-0.2	0.769665

unlikely due to numerical errors. Therefore, **W** does not follow PRDS on the true nulls in general. □

2.3.1.2 Condition Weaker than PRDS Even though PRDS is not satisfied by the test statistic **W**, FDR may still be controlled since PRDS is a sufficient but not necessary condition for the BH procedure to control FDR. We carefully examined the proof of Theorem 1.2 in Benjamini & Yekutieli (2001) to understand precisely why PRDS on the true nulls allows FDR to be controlled. Our goal was to see if other weaker dependence conditions would achieve the same result and that we would show our situation satisfies these less stringent conditions. Examination of the proof shows that there is control of FDR due to two factors:

1. the p-values corresponding to true null are stochastically larger than Uniform(0,1).
2. the p-values follow an LTD-like dependence structure which is actually weaker than the required PRDS conditioning on p-values from the true null hypotheses.

We note that PRDS of the test statistic $\mathbf{W}$ on the true nulls is sufficient to imply the LTD-like dependence of the p-values on the true nulls if the p-values are one-sided, as in our setting. If the above two factors were satisfied by the p-values from the min test, then FDR would be controlled without the specific need of $\mathbf{W}$ being PRDS on the true null. The p-values being stochastically larger than the Uniform(0,1) is easy to show and stated in Result 2.3.

Result 2.3. *The P-value from the min test corresponding to the true null hypothesis is stochastically larger than the Uniform(0,1).*

Proof. Let (U, V) be the vector for the two test statistics in the min test corresponding to a true null hypothesis, and assume it follows the following bivariate distribution:

$$\begin{pmatrix} U \\ V \end{pmatrix} \sim \mathbf{N_2}(\begin{pmatrix} \mu_U \\ \mu_V \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}), \quad \mu_U \leq 0 \text{ or } \mu_V \leq 0.$$

Also let $Z_1 = U - \mu_U, Z_2 = V - \mu_V$, so that Z_1 and Z_2 are both standard normal random variables. And let $P_U = 1 - \Phi(U), P_V = 1 - \Phi(V)$, where Φ is the cdf of the standard normal distribution. The p-value of the min test is $P = \max(P_U, P_V)$.

$$\begin{aligned}
Pr(P \leq t) &= Pr(P_U \leq t, P_V \leq t) \\
&= Pr(1 - \Phi(U) \leq t, 1 - \Phi(V) \leq t) \\
&= Pr(U \geq \Phi^{-1}(1-t), V \geq \Phi^{-1}(1-t)) \\
&= Pr(U - \mu_U \geq \Phi^{-1}(1-t) - \mu_U, V - \mu_V \geq \Phi^{-1}(1-t) - \mu_V) \\
&= Pr(Z_1 \geq \Phi^{-1}(1-t) - \mu_U, Z_2 \geq \Phi^{-1}(1-t) - \mu_V) \qquad (2.3.1) \\
&\leq Pr_{\mu_U=0,\mu_V=\infty}(U - 0 \geq \Phi^{-1}(1-t), V - \mu_V \geq -\infty) \qquad (2.3.2) \\
&= t.
\end{aligned}$$

where the inequality (2.3.2) follows because the probability in equation (2.3.1) is an increasing function of μ_U and μ_V. □

The proof of Theorem 1.2 in Benjamini & Yekutieli (2001) says the LTD-like condition of the p-values conditioning on the true nulls requires a series of $(m-1)$ functions based on the ordered p-values to be monotone. These conditions can be written as: for $k = 1, 2, \cdots, m-1$ and any $i \in I_0$, where I_0 is the index set for the true null hypotheses,

$$P(D_k^{(i)}|p_i \leq p) = P(p_{(k)}^{(i)} > q_{k+1}, p_{(k+1)}^{(i)} > q_{k+2}, \cdots, p_{(m-1)}^{(i)} > q_m|p_i \leq p) \quad \uparrow p, \tag{2.3.3}$$

where $q_k = \frac{k}{m}q$ and $p_{(1)}^{(i)} \leq \cdots \leq p_{(m-1)}^{(i)}$ are the ordered p-values without considering p_i.

We term the above condition in (2.3.3) as a "LTD-like condition" because $D_k^{(i)}$ is not required to be any upper set. Instead, it is relaxed to the upper orthant.

The remaining work is to check if this "LTD-like condition" is satisfied by the p-values after conditioning on the true null hypotheses. Since test statistics are easier to work with, the "LTD-like condition" on the p-values can be translated as an "RTI-like condition" on the test statistics for the one-sided p-values from the right tail. Similar to the "LTD-like condition", the "RTI-like condition" requires a series of (m-1) functions based on the ordered test statistics to be monotone, i.e., for $k = 1, 2, \cdots, m-1$ and any $i \in I_0$,

$$P(T_{(k)}^{(i)} < t_{k+1}, T_{(k+1)}^{(i)} > t_{k+2}, \cdots, T_{(m-1)}^{(i)} > t_m|T_i \leq t) \quad \downarrow t, \tag{2.3.4}$$

where $T_{(1)}^{(i)} \geq \cdots \geq T_{(m-1)}^{(i)}$ are the ordered test statistics without considering T_i.

To ensure working with test statistics is valid, the following shows that the "LTD-like condition" on p-values is equivalent to the "RTI-like condition" on test statisics for one-sided p-values from the right tail.

Result 2.4. *If the p-values are one-sided from the right tail, the p-values are "LTD-like" iff the test statistics are "RTI-like".*

Proof. If the p-value is one-sided from the right tail, then it is a decreasing function of the test statistic, i.e., $P_k = 1 - F(T_k)$, where P_k is the p-value corresponding to test statistic T_k and F is the cdf for T_k. Let $T_{(1)}^{(i)} \geq \cdots \geq T_{(m-1)}^{(i)}$ be the ordered test statistics without considering T_i and $t_k = F^{-1}(1-q_k)$ for $k = 1, 2, \cdots, m-1$. Also let $\mathbf{P}$ denote the vector for the p-values and $\mathbf{T}$ denote the vector for the test statistics, then:

$$
\begin{aligned}
D^{(i)}_k &= \{\mathbf{P} : P^{(i)}_{(k)} > q_{k+1}, P^{(i)}_{(k+1)} > q_{k+2}, \cdots, P^{(i)}_{(m-1)} > q_m\} \\
&= \{\mathbf{T} : 1 - F(T^{(i)}_{(k)}) > q_{k+1}, 1 - F(T^{(i)}_{(k+1)}) > q_{k+2}, \cdots, 1 - F(T^{(i)}_{(m-1)}) > q_m\} \\
&= \{\mathbf{T} : F(T^{(i)}_{(k)}) < 1 - q_{k+1}, F(T^{(i)}_{(k+1)}) < 1 - q_{k+2}, \cdots, F(T^{(i)}_{(m-1)}) < 1 - q_m\} \\
&= \{\mathbf{T} : T^{(i)}_{(k)} < F^{-1}(1 - q_{k+1}), T^{(i)}_{(k+1)} < F^{-1}(1 - q_{k+2}), \cdots, T^{(i)}_{(m-1)} < F^{-1}(1 - q_m)\} \\
&= \{\mathbf{T} : T^{(i)}_{(k)} < t_{k+1}, T^{(i)}_{(k+1)} < t_{k+1}, \cdots, T^{(i)}_{(m-1)} < t_{k+1}\}.
\end{aligned}
$$

It's noteworthy that $D^{(i)}_k$ is an upper set in terms of $\mathbf{P}$ but a lower set in terms of $\mathbf{T}$. Also let $t = F^{-1}(1-p)$, then $\{P_i \leq p\} = \{1 - F(T_i) \leq p\} = \{T_i \geq t\}$. Therefore:

$$
\begin{aligned}
&P(P^{(i)}_{(k)} > q_{k+1}, P^{(i)}_{(k+1)} > q_{k+2}, \cdots, P^{(i)}_{(m-1)} > q_m | P_i \leq p) \\
=&P(T^{(i)}_{(k)} < t_{k+1}, T^{(i)}_{(k+1)} < t_{k+2}, \cdots, T^{(i)}_{(m-1)} < t_m | T_i \geq t).
\end{aligned}
\tag{2.3.5}
$$

If the p-values are "LTD-like", then when p increases, i.e.,t decreases, the above probability in (2.3.5) increases, so the test statistics are "RTI-like". On the other hand, if the test statistics are "RTI-like", then when t increases, i.e.,p decreases, the above probability in (2.3.5) decreases, so the p-values are "LTD-like", thus completing the proof. □

Similarly, it can be shown that if the p-values are one-sided from the right tail, the p-values are "RTI-like" iff the test statistics are "LTD-like". The relationship of the dependence structures on the p-values ($\mathbf{P}$) and test statistics ($\mathbf{T}$) when the p-values are one-sided from the right tail can be summarized as in Figure 2.1:

Because the p-values from the min tests are one-sided from the right tail in our case, by replacing $\mathbf{T}$ with $\mathbf{W}$, Result 2.4 above allows us to restate the weaker condition needed in Benjamini & Yekutieli (2001) for the FDR to be controlled in terms of $\mathbf{W}$.

Result 2.5. *If the min test statistics* $\mathbf{W}$ *follow the "RTI-like condition" on the true nulls, then the FDR using the proposed method is controlled by the BH procedure.*

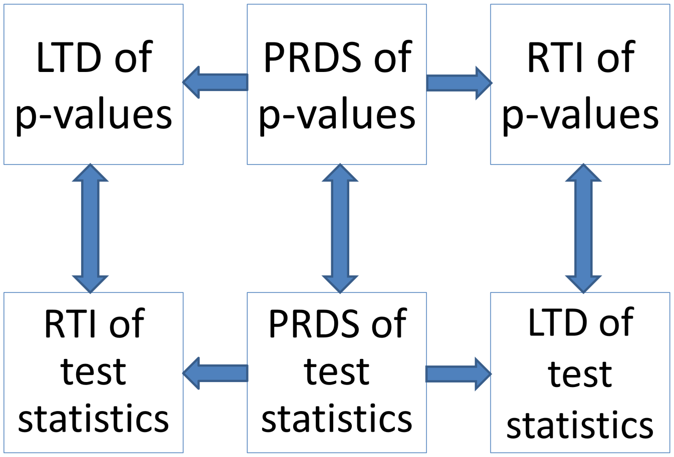


Figure 2.1: Relationship of dependence between test statistics and p-values

Note that the "RTI-like condition" of $\mathbf{W}$ involves a series of $(m-1)$ functions to be non-increasing in t. Replacing the statistic $\mathbf{T}$ in function (2.3.4) with $\mathbf{W}$, the "RTI-like condition" of $\mathbf{W}$ can be written as:

$$\begin{aligned}
k = m-1: \quad & P(W^{(i)}_{(k)} < t_{k+1}, W^{(i)}_{(k+1)} < t_{k+2}, \cdots, W^{(i)}_{(m-1)} < t_m | W_i \geq t) \\
& = P(W^{(i)}_{(m-1)} < t_m | W_i \geq t) \downarrow t. \\
k = m-2: \quad & P(W^{(i)}_{(k)} < t_{k+1}, W^{(i)}_{(k+1)} < t_{k+2}, \cdots, W^{(i)}_{(m-1)} < t_m | W_i \geq t) \\
& = P(W^{(i)}_{(m-2)} < t_{m-1}, W^{(i)}_{(m-1)} < t_m | W_i \geq t) \downarrow t. \\
\cdots: \quad & \cdots \\
k = 1: \quad & P(W^{(i)}_{(k)} < t_{k+1}, W^{(i)}_{(k+1)} < t_{k+2}, \cdots, W^{(i)}_{(m-1)} < t_m | W_i \geq t) \\
& = P(W^{(i)}_{(1)} < t_2, W^{(i)}_{(2)} < t_3, \cdots, W^{(i)}_{(m-1)} < t_m | W_i \geq t) \downarrow t.
\end{aligned} \tag{2.3.6}$$

All of the above $(m-1)$ sub-conditions need to be satisfied for the "RTI-like condition" of $\mathbf{W}$ to hold so that the resulting FDR to be controlled. However, since more than one order statistic is involved, it turns out difficult to show all these sub-conditions hold. The easiest one to show is the $k = m-1$ condition in (2.3.6), whose proof is stated next.

Result 2.6. *If* $\mathbf{U} = (U_1, U_2, \cdots, U_m), \mathbf{V} = (V_1, V_2, \cdots, V_m)$ *are two independent normally distributed vectors and the correlation coefficients* $\rho_{Uij} \geq 0, \rho_{Vij} \geq 0$ *for* $j = 1, 2, \cdots, m, j \neq i$*, then the elementwise minimum vector* $\mathbf{W} = \min(\mathbf{U}, \mathbf{V})$ *satisfies the* $k = m - 1$ *condition* $P(W^{(i)}_{(m-1)} < t_m | W_i \geq t) \downarrow t$ *for any* $i \in I_0$*.*

Proof. $\mathbf{U}$ can be decomposed into two parts, $\mathbf{U}^{(i)}$ and U_i, where $\mathbf{U}^{(i)}$ is the vector left after dropping U_i. The mean vector and covariance matrix can be decomposed correspondingly as

$$\mathbf{U} = \begin{pmatrix} \mathbf{U}^{(i)} \\ U_i \end{pmatrix} \sim \mathbf{N}_m \left(\begin{pmatrix} \mu^{(i)} \\ \mu_i \end{pmatrix}, \begin{pmatrix} \mathbf{\Sigma_U}^{(i)} & \rho_{\mathbf{U}}{}^{(i)} \\ \rho_{\mathbf{U}}{}^{(i)\prime} & 1 \end{pmatrix} \right),$$

where $\rho_{\mathbf{U}}{}^{(i)} = (\rho_{U1i}, \cdots, \rho_{U(i-1)i}, \rho_{U(i+1)i}, \cdots, \rho_{Umi})'$. Let $\mu_{\mathbf{c}} = \mu^{(i)} + \rho_{\mathbf{U}}{}^{(i)}(u - \mu_i), \mathbf{\Sigma_c} = \mathbf{\Sigma_U}^{(i)} - \rho_{\mathbf{U}}{}^{(i)}\rho_{\mathbf{U}}{}^{(i)\prime}$, then the conditional distribution of $\mathbf{U}^{(i)}|U_i = u$ is:

$$\mathbf{U}^{(i)}|U_i = u \sim \mathbf{N}_{m-1}(\mu_{\mathbf{c}}, \mathbf{\Sigma_c}).$$

If $\rho_{Uij} \geq 0$ for $j = 1, 2, \cdots, m, j \neq i$, then $\rho_{\mathbf{U}}{}^{(i)} \geq 0$, so the conditional mean $\mu_{\mathbf{c}}$ is a nondecreasing function of u. Let $\mathbf{s} = (s_1, s_2, \cdots, s_{m-1})'$ and $(\mathbf{U}^{(i)} \geq \mathbf{s}) = (U_1^{(i)} \geq$

$s_1, \cdots, U_{m-1}^{(i)} \geq s_{m-1})$, then we have

$$\begin{aligned}
P(\mathbf{U}^{(i)} \geq \mathbf{s}|U_i = u) &= P({\mathbf{\Sigma_c}}^{-1/2}(\mathbf{U}^{(i)} - \mu_{\mathbf{c}}) \geq {\mathbf{\Sigma_c}}^{-1/2}(\mathbf{s} - \mu_{\mathbf{c}})) \\
&= P({\mathbf{\Sigma_c}}^{-1/2}(\mathbf{U}^{(i)} - \mu_{\mathbf{c}}) \geq {\mathbf{\Sigma_c}}^{-1/2}(\mathbf{s} - \mu^{(i)} - {\rho_{\mathbf{U}}}^{(i)}(u - \mu_i))) \\
&\uparrow u. \\
P(\mathbf{U}^{(i)} \geq \mathbf{s}|U_i \geq t) &= \frac{P(\mathbf{U}^{(i)} \geq \mathbf{s}, U_i \geq t)}{P(U_i \geq t)} \\
&= \frac{\int_t^\infty P(\mathbf{U}^{(i)} \geq \mathbf{s}|U_i = u) f_{U_i}(u)du}{\int_t^\infty f_{U_i}(u)du}. \\
\frac{\partial}{\partial t} P(\mathbf{U}^{(i)} \geq \mathbf{s}|U_i \geq t) &= \frac{1}{(\int_t^\infty f_{U_i}(u)du)^2} \\
&\quad * [-P(\mathbf{U}^{(i)} \geq \mathbf{s}|U_i = t) f_{U_i}(t) * \int_t^\infty f_{U_i}(u)du \\
&\quad + \int_t^\infty P(\mathbf{U}^{(i)} \geq \mathbf{s}|U_i = u) f_{U_i}(u)du * f_{U_i}(t)] \\
&\geq \frac{f_{U_i}(t)}{(\int_t^\infty f_{U_i}(u)du)^2} \\
&\quad * [-P(\mathbf{U}^{(i)} \geq \mathbf{s}|U_i = t) * \int_t^\infty f_{U_i}(u)du \\
&\quad + P(\mathbf{U}^{(i)} \geq \mathbf{s}|U_i = t) \int_t^\infty f_{U_i}(u)du] \\
&= 0.
\end{aligned}$$

Therefore, $P(\mathbf{U}^{(i)} \geq \mathbf{s}|U_i \geq t)$ is an increasing function in t, and similarly, we can show $P(\mathbf{V}^{(i)} \geq \mathbf{s}|V_i \geq t)$ also increases in t. Now let $s_1 = s_2 = \cdots = s_{m-1} = t_m$, then

$$P(\mathbf{U}^{(i)} \geq \mathbf{s}|U_i \geq t) = P(U_1^{(i)} \geq t_m, \cdots, U_{m-1}^{(i)} \geq t_m | U_i \geq t) \uparrow t,$$
$$P(\mathbf{V}^{(i)} \geq \mathbf{s}|V_i \geq t) = P(V_1^{(i)} \geq t_m, \cdots, V_{m-1}^{(i)} \geq t_m | V_i \geq t) \uparrow t.$$

Therefore, we have

$$\begin{aligned}
&P(W^{(i)}_{(m-1)} < t_m | W_i \geq t)\\
=&1 - P(W^{(i)}_{(m-1)} \geq t_m | W_i \geq t)\\
=&1 - P(W^{(i)}_1 \geq t_m, W^{(i)}_2 \geq t_m, \cdots, W^{(i)}_{m-1} \geq t_m | W_i \geq t)\\
=&1 - \frac{P(W^{(i)}_1 \geq t_m, \cdots, W^{(i)}_{m-1} \geq t_m, W_i \geq t)}{P(W_i \geq t)}\\
=&1 - \frac{P(U^{(i)}_1 \geq t_m, V^{(i)}_1 \geq t_m, \cdots, U^{(i)}_{m-1} \geq t_m, V^{(i)}_{m-1} \geq t_m, U_i \geq t, V_i \geq t)}{P(U_i \geq t, V_i \geq t)}\\
=&1 - \frac{P(U^{(i)}_1 \geq t_m, \cdots, U^{(i)}_{m-1} \geq t_m, U_i \geq t)}{P(U_i \geq t)} * \frac{P(V^{(i)}_1 \geq t_m, \cdots, V^{(i)}_{m-1} \geq t_m, V_i \geq t)}{P(V_i \geq t)}\\
=&1 - P(U^{(i)}_1 \geq t_m, \cdots, U^{(i)}_{m-1} \geq t_m | U_i \geq t) * P(V^{(i)}_1 \geq t_m, \cdots, V^{(i)}_{m-1} \geq t_m | V_i \geq t)\\
&\downarrow t.
\end{aligned}$$

□

While the $k = m - 1$ case of the RTI-like condition of $\mathbf{W}$ conditioning on the true nulls can be shown, the rest of the conditions require very complex proofs and are hard to show, because not only is the minimum of two variables involved but also the joint distribution of the ordered statistics on these minimums needs to be dealt with. We have been unsuccessful in developing a theoretical approach to show these conditions hold, but are somewhat optimistic given that the $k = m - 1$ condition holds. We next undertook a small simulation study to see if the FDR is controlled using the BH method in our setting.

2.3.2 Simulation Study

As it is hard to show $\mathbf{W}$ follows the RTI-like condition after conditioning on the true nulls, a simulation study is performed to get an idea whether FDR can be controlled at the 0.05 level using the proposed method to identify common significant neurobiological characteristics across two DSM diagnoses.

2.3.2.1 Simulation Method In each simulation 500 neurobiological characteristics were tested ($m = 500$). For the ith test, two independent and normally distributed test statistics U_i and V_i were simulated. Because U_i and U_h were correlated, all the U_i's were simulated together through a 500-dimension vector $\mathbf{U}$ from a multivariate normal distribution. Similarly, a 500-dimension vector $\mathbf{V}$ was simulated. Among the 500 hypotheses, there are four scenarios regarding to whether the two sub-null hypotheses are true or false. The proportions of each scenario are summarized in Table 2.3. For instance, p_{00}*100% of the 500 hypotheses were true in both null hypothese tested by U_i and V_i.

Table 2.3: Proportion of Hypotheses

	true null tested by U_i true	false null tested by U_i
true null tested by V_i	p_{00}	p_{10}
false null tested by V_i	p_{01}	$1 - p_{00} - p_{01} - p_{10}$

Because the null hypothesis is true when either of the two sub-null hypothesis is true, collectively, $(p_{00} + p_{01} + p_{10})$*100% of the 500 null hypotheses were true and $(1 - p_{00} - p_{01} - p_{10})$*100% were false. Equal correlation was assumed between any two elements within each of $\mathbf{U}$ and $\mathbf{V}$. The mean vectors and covariance matrices for $\mathbf{U}$ and $\mathbf{V}$ were as follows:

$$E\begin{pmatrix} U_1 \\ \vdots \\ U_{500*p_{00}} \\ U_{500*p_{00}+1} \\ \vdots \\ U_{500*(p_{00}+p_{01})} \\ U_{500*(p_{00}+p_{01})+1} \\ \vdots \\ U_{500*(p_{00}+p_{01}+p_{10})} \\ U_{500*(p_{00}+p_{01}+p_{10})+1} \\ \vdots \\ U_{500} \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 0 \\ \vdots \\ 0 \\ \delta \\ \vdots \\ \delta \\ \delta \\ \vdots \\ \delta \end{pmatrix}, \quad E\begin{pmatrix} V_1 \\ \vdots \\ V_{500*p_{00}} \\ V_{500*p_{00}+1} \\ \vdots \\ V_{500*(p_{00}+p_{01})} \\ V_{500*(p_{00}+p_{01})+1} \\ \vdots \\ V_{500*(p_{00}+p_{01}+p_{10})} \\ V_{500*(p_{00}+p_{01}+p_{10})+1} \\ \vdots \\ V_{500} \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \delta \\ \vdots \\ \delta \\ 0 \\ \vdots \\ 0 \\ \delta \\ \vdots \\ \delta \end{pmatrix};$$

$$Cov(\mathbf{U}) = \begin{pmatrix} 1 & \rho_U & \dots & \rho_U \\ \rho_U & 1 & \dots & \rho_U \\ \vdots & \vdots & \ddots & \vdots \\ \rho_U & \rho_U & \dots & 1 \end{pmatrix}, \quad Cov(\mathbf{V}) = \begin{pmatrix} 1 & \rho_V & \dots & \rho_V \\ \rho_V & 1 & \dots & \rho_V \\ \vdots & \vdots & \ddots & \vdots \\ \rho_V & \rho_V & \dots & 1 \end{pmatrix}.$$

Here δ was the effect size whenever the sub-null hypothesis corresponding to U_i or V_i was not true. After the proposed method was applied to the simulated data each time, the total number of rejections r_{w} and false rejections v_{w} were recorded. Here the subscript w indicates the min test is applied for each individual hypothesis. The false discovery proportion ($\mathrm{FDP}_{\mathrm{w}}$) was calculated as $v_{\mathrm{w}}/r_{\mathrm{w}}$. $\mathrm{Power}_{\mathrm{w}}$ was calculated as $(r_{\mathrm{w}} - v_{\mathrm{w}})/(500 * (1 - p_{00} - p_{01} - p_{10}))$. Different effect sizes δ, correlations ρ_U and ρ_V and proportions of true nulls p_{00}, p_{01} and p_{10} were used. For each parameter configuration, 2000 simulations were run and the average of the 2000 FDPs and Powers were obtained as the estimated FDR and Power.

The parameters used in the simulation are listed in Table 2.4. The combination of p_{00}, p_{01} and p_{10} used satisfied the condition $p_{00} + p_{01} + p_{10} \leq 1$.

Table 2.4: Simulation Parameter List

parameter	value
m	500
α	0.05
δ	0.5, 1, 2, 4
ρ_U	0, 0.01, 0.4, 0.8
ρ_V	0, 0.01, 0.4, 0.8
p_{00}	0.36, 0.49, 0.64
p_{01}	0.16, 0.21, 0.24
p_{10}	0.16, 0.21, 0.24

2.3.2.2 Simulation Result

The simulation results are summarized in Tables 2.5 - 2.8 with standard deviations in parentheses. For example, in Table 2.5, if in each simulation

49% of the 500 hypotheses are true in both sub-nulls ($p_{00} = 0.49$), 21% of the hypotheses are true only in the sub-null tested by U_i ($p_{01} = 0.21$), 21% of the hypotheses are true only in the sub-null tested by V_i ($p_{10} = 0.21$), the correlation between U_i and U_h is 0.4 ($\rho_U = 0.4$) and the correlation between V_i and V_h is 0.8 ($\rho_V = 0.8$), the calculated FDR of the proposed method, i.e., BH method being applied to p-values obtained from the min tests, is 0.0004 with a sample standard deviation 0.0169. The calculated power of the proposed method is 0.0001 with a sample standard deviation 0.0040. The columns of $\mathrm{FDR_I}$ and $\mathrm{Power_I}$ are explained in later sections.

As can be seen in Tables 2.5 - 2.8, even though we cannot prove the min test statistic $\mathbf{W}$ follows the "RTI-like condition" after conditioning on the true nulls, the FDRs under all the simulated parameter configurations have been controlled at the 0.05 level using the proposed method. The FDR and power increases as the effect size increases or the correlation among elements in $\mathbf{U}$ or among those in $\mathbf{V}$ decreases. When the proportion of the true null hypotheses ($p_{00} + p_{01} + p_{10}$) increases so that the actual level the FDR was controlled at, which is $(p_{00} + p_{01} + p_{10}) * 0.05$, increases, the FDR stayed roughly the same and the power decreases. Due to the limited space here, not all results are listed, however, similar patterns are indeed obtained for other parameter configurations.

2.3.2.3 Comparison with the Intersection Method To better understand the power and FDR control using the proposed method, its simulation result was compared with an alternative method which is called the intersection method by us. The intersection method is to mimic the way how the genes were found to be significant in both schizophrenia and bipolar disorder in some of the genome-wide association studies, i.e., identification in one DSM diagnosis at one time. It is implemented by testing if the mean of a neurobiological characteristic is significantly up-regulated in each DSM diagnosis separately, applying the BH procedure at level α on each set of the p-values and then finding the intersection of the decisions for each DSM diagnosis. So for characteristic i, there is no overall null hypothesis as the one in (2.2.3). If there are two DSM diagnoses considered, the two sub-null hypotheses are tested separately, one based on U_i and the other based on V_i. Two sets of p-values

Table 2.5: Simulation Results ($\delta = 0.5$)

p_{00}	p_{01}	p_{10}	ρ_U	ρ_V	FDR_w	FDR_I	Power_w	Power_I
0.36	0.16	0.16	0.00	0.00	0.0000	0.0000	0.0000	0.0000
					(0.0000)	(0.0000)	(0.0000)	(0.0000)
			0.00	0.01	0.0000	0.0000	0.0000	0.0000
					(0.0000)	(0.0000)	(0.0001)	(0.0001)
			0.01	0.01	0.0000	0.0000	0.0000	0.0000
					(0.0000)	(0.0000)	(0.0000)	(0.0000)
			0.40	0.01	0.0000	0.0010	0.0000	0.0000
					(0.0000)	(0.0316)	(0.0000)	(0.0000)
			0.40	0.40	0.0000	0.0013	0.0000	0.0002
					(0.0000)	(0.0292)	(0.0000)	(0.0061)
			0.40	0.80	0.0003	0.0017	0.0002	0.0003
					(0.0127)	(0.0365)	(0.0082)	(0.0096)
			0.80	0.80	0.0003	0.0006	0.0005	0.0009
					(0.0145)	(0.0185)	(0.0224)	(0.0278)
0.49	0.21	0.21	0.00	0.00	0.0000	0.0000	0.0000	0.0000
					(0.0000)	(0.0000)	(0.0000)	(0.0000)
			0.00	0.01	0.0000	0.0000	0.0000	0.0000
					(0.0000)	(0.0000)	(0.0000)	(0.0000)
			0.01	0.01	0.0000	0.0000	0.0000	0.0000
					(0.0000)	(0.0000)	(0.0000)	(0.0000)
			0.40	0.01	0.0000	0.0005	0.0000	0.0000
					(0.0000)	(0.0224)	(0.0000)	(0.0000)
			0.40	0.40	0.0000	0.0008	0.0000	0.0001
					(0.0000)	(0.0244)	(0.0000)	(0.0045)
			0.40	0.80	0.0004	0.0013	0.0001	0.0002
					(0.0169)	(0.0344)	(0.0040)	(0.0071)
			0.80	0.80	0.0004	0.0008	0.0005	0.0008
					(0.0201)	(0.0265)	(0.0224)	(0.0258)

Table 2.6: Simulation Results ($\delta = 1$)

p_{00}	p_{01}	p_{10}	ρ_U	ρ_V	$\mathrm{FDR_w}$	$\mathrm{FDR_I}$	$\mathrm{Power_w}$	$\mathrm{Power_I}$
0.36	0.16	0.16	0.00	0.00	0.0000	0.0005	0.0000	0.0001
					(0.0000)	(0.0224)	(0.0003)	(0.0007)
			0.00	0.01	0.0000	0.0005	0.0000	0.0001
					(0.0000)	(0.0224)	(0.0002)	(0.0008)
			0.01	0.01	0.0000	0.0000	0.0000	0.0001
					(0.0000)	(0.0000)	(0.0002)	(0.0007)
			0.40	0.01	0.0000	0.0060	0.0000	0.0004
					(0.0000)	(0.0667)	(0.0002)	(0.0027)
			0.40	0.40	0.0003	0.0030	0.0002	0.0019
					(0.0083)	(0.0354)	(0.0073)	(0.0236)
			0.40	0.80	0.0005	0.0034	0.0008	0.0035
					(0.0148)	(0.0379)	(0.0208)	(0.0374)
			0.80	0.80	0.0010	0.0022	0.0026	0.0059
					(0.0204)	(0.0283)	(0.0458)	(0.0669)
0.49	0.21	0.21	0.00	0.00	0.0000	0.0000	0.0000	0.0000
					(0.0000)	(0.0000)	(0.0005)	(0.0007)
			0.00	0.01	0.0005	0.0005	0.0000	0.0000
					(0.0224)	(0.0224)	(0.0005)	(0.0009)
			0.01	0.01	0.0005	0.0005	0.0000	0.0000
					(0.0224)	(0.0224)	(0.0007)	(0.0009)
			0.40	0.01	0.0000	0.0042	0.0000	0.0002
					(0.0000)	(0.0616)	(0.0000)	(0.0025)
			0.40	0.40	0.0000	0.0051	0.0000	0.0010
					(0.0000)	(0.0621)	(0.0005)	(0.0169)
			0.40	0.80	0.0004	0.0044	0.0004	0.0018
					(0.0185)	(0.0556)	(0.0150)	(0.0261)
			0.80	0.80	0.0008	0.0032	0.0009	0.0033
					(0.0252)	(0.0463)	(0.0283)	(0.0494)

Table 2.7: Simulation Results ($\delta = 2$)

p_{00}	p_{01}	p_{10}	ρ_U	ρ_V	FDR_w	FDR_I	Power_w	Power_I
0.36	0.16	0.16	0.00	0.00	0.0013	0.0246	0.0043	0.1247
					(0.0289)	(0.0348)	(0.0088)	(0.0332)
			0.00	0.01	0.0011	0.0243	0.0040	0.1252
					(0.0194)	(0.0345)	(0.0082)	(0.0402)
			0.01	0.01	0.0002	0.0254	0.0044	0.1248
					(0.0061)	(0.0364)	(0.0093)	(0.0449)
			0.40	0.01	0.0024	0.0205	0.0152	0.1208
					(0.0187)	(0.0505)	(0.0468)	(0.1233)
			0.40	0.40	0.0033	0.0178	0.0330	0.1138
					(0.0177)	(0.0538)	(0.1154)	(0.1910)
			0.40	0.80	0.0044	0.0163	0.0526	0.1197
					(0.0284)	(0.0630)	(0.1695)	(0.2362)
			0.80	0.80	0.0048	0.0104	0.0727	0.1236
					(0.0332)	(0.0481)	(0.2335)	(0.2911)
0.49	0.21	0.21	0.00	0.00	0.0004	0.0633	0.0025	0.0669
					(0.0134)	(0.1537)	(0.0088)	(0.0420)
			0.00	0.01	0.0009	0.0652	0.0022	0.0678
					(0.0256)	(0.1517)	(0.0080)	(0.0459)
			0.01	0.01	0.0021	0.0626	0.0023	0.0669
					(0.0391)	(0.1578)	(0.0086)	(0.0473)
			0.40	0.01	0.0003	0.0458	0.0034	0.0711
					(0.0066)	(0.1128)	(0.0162)	(0.0934)
			0.40	0.40	0.0023	0.0350	0.0075	0.0722
					(0.0305)	(0.1093)	(0.0487)	(0.1510)
			0.40	0.80	0.0033	0.0303	0.0134	0.0755
					(0.0382)	(0.1172)	(0.0808)	(0.1874)
			0.80	0.80	0.0035	0.0178	0.0200	0.0814
					(0.0437)	(0.0916)	(0.1247)	(0.2368)

Table 2.8: Simulation Results ($\delta = 4$)

p_{00}	p_{01}	p_{10}	ρ_U	ρ_V	$\mathrm{FDR}_{\mathrm{w}}$	$\mathrm{FDR}_{\mathrm{I}}$	$\mathrm{Power}_{\mathrm{w}}$	$\mathrm{Power}_{\mathrm{I}}$
0.36	0.16	0.16	0.00	0.00	0.0161	0.0247	0.9346	0.9576
					(0.0103)	(0.0122)	(0.0199)	(0.0160)
			0.00	0.01	0.0160	0.0248	0.9342	0.9576
					(0.0103)	(0.0125)	(0.0227)	(0.0172)
			0.01	0.01	0.0156	0.0245	0.9339	0.9575
					(0.0105)	(0.0128)	(0.0235)	(0.0179)
			0.40	0.01	0.0154	0.0248	0.9294	0.9538
					(0.0201)	(0.0282)	(0.0911)	(0.0687)
			0.40	0.40	0.0159	0.0257	0.9273	0.9521
					(0.0283)	(0.0392)	(0.1172)	(0.0878)
			0.40	0.80	0.0155	0.0253	0.9177	0.9463
					(0.0418)	(0.0562)	(0.1903)	(0.1498)
			0.80	0.80	0.0149	0.0247	0.9070	0.9379
					(0.0497)	(0.0669)	(0.2458)	(0.2002)
0.49	0.21	0.21	0.00	0.00	0.0189	0.0688	0.8218	0.9332
					(0.0219)	(0.0357)	(0.0641)	(0.0379)
			0.00	0.01	0.0195	0.0694	0.8212	0.9330
					(0.0232)	(0.0380)	(0.0688)	(0.0390)
			0.01	0.01	0.0188	0.0677	0.8230	0.9330
					(0.0227)	(0.0386)	(0.0684)	(0.0383)
			0.40	0.01	0.0188	0.0665	0.8142	0.9279
					(0.0363)	(0.0750)	(0.1593)	(0.0932)
			0.40	0.40	0.0183	0.0648	0.8054	0.9255
					(0.0478)	(0.1011)	(0.2136)	(0.1188)
			0.40	0.80	0.0169	0.0587	0.7913	0.9147
					(0.0681)	(0.1255)	(0.2951)	(0.1909)
			0.80	0.80	0.0150	0.0509	0.7792	0.9034
					(0.0774)	(0.1427)	(0.3544)	(0.2490)

are obtained when using the intersection method, one for each DSM diagnosis. The BH procedure is then applied to the two sets of p-values, each at level α. Final decisions are made according to the intersection of the results on the two sets of p-values.

In the simulation, the intersection method was applied to the same dataset used by the proposed method. In each simulation, the total number of rejections r_{I} and that of false rejections v_{I} were recorded. Here the subscript I indicates the intersection method is used. $\mathrm{FDP_I}$ was calculated as $v_{\mathrm{I}}/r_{\mathrm{I}}$. $\mathrm{Power_I}$ was calculated as $(r_{\mathrm{I}} - v_{\mathrm{I}})/(500 * (1 - p_{00} - p_{01} - p_{10}))$. The results using the intersection method are compared with that using the proposed method and are also shown in Tables 2.5 - 2.8. For example, in Table 2.5 again, if in each simulation 49% of the 500 hypotheses are true in both sub-nulls ($p_{00} = 0.49$), 21% of the hypotheses are true only in the sub-null tested by U_i ($p_{01} = 0.21$), 21% of the hypotheses are true only in the sub-null tested by V_i ($p_{10} = 0.21$), the correlation between U_i and U_h is 0.4 ($\rho_U = 0.4$) and the correlation between V_i and V_h is 0.8 ($\rho_V = 0.8$), the calculated FDR of the intersection method is 0.0013 with a sample standard deviation 0.0344. The calculated power of the intersection method is 0.0002 with a sample standard deviation 0.0071.

When the effect size δ is large or the proportion of true null hypothesis is high, FDR using the intersection method is inflated even though it tends to have larger power. For example as in Table 2.7, when $p_{00} = 0.49, p_{01} = 0.21, p_{10} = 0.21, \rho_U = 0$ and $\rho_V = 0$, the calculated FDR using the intersection method is 0.0633 with sample standard deviation 0.1537. Similar as using the proposed method, the FDR using the intersection method tends to increase when the effect size increases or when the proportion of true null hypotheses increases or when the correlation among elements in $\mathbf{U}$ or $\mathbf{V}$ decreases. Power of the intersection method increases when the effect size increases or when the proportion of true null hypotheses decreases or when the correlation among elements in $\mathbf{U}$ or $\mathbf{V}$ decreases.

Comparing the calculated FDR values and power values between the proposed method and the intersection method in Tables 2.5 - 2.8, we can see that when the effect size δ is small as in Table 2.5, the calculated FDRs and powers from both methods are very low and are very close to each other, not matter how the proportion of true null hypothesis ($p_{00} + p_{01} + p_{10}$) or the correlation among elements in $\mathbf{U}$ or $\mathbf{V}$ changes. When the effect size δ is large as in Table 2.8, the calculated FDR as well as power with the intersection method are always higher than those with the proposed method. However, in this case, the FDR with the intersection method can be inflated and its variability is larger than that with the intersection method. If the proportion of true nulls and the correlation among elements in $\mathbf{U}$ or $\mathbf{V}$ hold constant, the FDR and power using both methods increase when δ increases and both FDR and power with the intersection method increase faster than those with the proposed method.

The proposed method compared favorably to the intersection method in terms of FDR protection, but at a cost of losing power. The seemingly larger power using the intersection method might be due to the fact that the BH procedure was applied at level α twice, thus the overall targeted FDR was already elevated. However, it is still unknown how to partition the α to the two applications so that the overall FDR can be controlled at the desired level and what the power would be like in that case.

2.3.3 A Case Study

The proposed method was applied to a dataset obtained from a post-mortem tissue study about mRNA levels in the brains of schizophrenia patients.

2.3.3.1 Brief Description of the Study and Data Cognitive deficit and dorsolateral prefrontal cortex (DLPFC) dysfunction are likely to be associated with altered mRNA levels in schizophrenia patients. In this post-mortem tissue study, brain tissue samples from 42 pairs of subjects were processed and the expression levels of 26 mRNAs were measured on each subject. Among these 42 pairs, 14 were schizoaffective (SA) patients matched with healthy controls on gender and as closely as possible on age and PMI, and the other 28 were schizophrenia (SZ) subjects matched with healthy controls in the same way. [1]

[1] As the mRNA levels from multiple brain regions were assessed, three patients were matched to two different healthy control subjects for different mRNAs. For each of the three patients, we combined the data from the two control subjects as if they were from a single control subject because dealing with the effects of differing controls was complicated, as shown in Wu & Sampson (2012) and there were only three such cases in our data.

2.3.3.2 Application Result The proposed method was applied twice to the dataset, one to identify mRNAs whose levels are significantly up-regulated in both SA and SZ and the other to identify those that are significantly down-regulated in both DSM diagnoses. The FDR was controlled at the 0.05 level for both applications. In order to find those that are significantly down-regulated, the negative values of the measurements were used so that the p-values were still one-sided from the right tail and the conditions for the proofs in previous sections hold. The measurements in the data were assumed to be normally distributed. However, the variances of the distribution were unknown, thus the two test statistics U_i and V_i followed t distributions. Because the patients and the healthy controls were matched, paired t tests were applied and the degrees of freedom were 13 and 27 for SA and SZ, respectively. The p-values from the t distribution and those from the standard normal distribution using the same test statistic value were compared and they were fairly close. Therefore, the p-values from the t distribution were used even though the preceding results and proofs were all based on normal distributions.

While controlling the FDR at the 0.05 level, two mRNA levels were found to be significantly up-regulated in both SZ and SA using the proposed method and none was found to be significantly down-regulated in both diagnoses after the multiplicity adjustments. The p-values based on the t distributions and final conclusion for each mRNA were tabulated in Tables 2.9 - 2.10. The intersection method was also applied with the same result achieved.

Table 2.9: mRNA levels significantly elevated in both SA and SZ

Gene	SA p-value	SZ p-value	max(SA p-value, SZ p-value)	Reject
CRIP1a	0.93241	0.96485	0.96485	0
DAGLa	0.15061	0.59606	0.59606	0
DAGLb	0.69489	0.39986	0.69489	0
FAAH	0.20300	0.16724	0.20300	0
GABA Receptor 1	0.95428	0.98195	0.98195	0
GAD67	0.91706	0.99929	0.99929	0
GAT1	0.18642	0.34657	0.34657	0
IFITM1	0.05481	0.00018	0.05481	0
IFITM23	0.02390	0.00002	0.02390	0
KCC2	0.39211	0.01201	0.39211	0
LHX6	0.89617	0.97673	0.97673	0
MGL	0.29445	0.33280	0.33280	0
Mu Opioid Receptor	0.08640	0.00004	0.08640	0
NKCC1	0.95251	0.94291	0.95251	0
OXSR1	0.00095	0.00000	0.00095	1
Parvalbumin	0.99887	0.98308	0.99887	0
RGS4	0.99265	1.00000	1.00000	0
STK39	0.67471	0.98696	0.98696	0
Somatostatin	0.93241	0.99937	0.99937	0
TRPV1	0.08923	0.00000	0.08923	0
WNK1	0.44529	0.06957	0.44529	0
WNK3	0.00037	0.00000	0.00037	1
WNK4	0.43069	0.04117	0.43069	0
mGluR1a	0.15880	0.00026	0.15880	0
mGluR5	0.15018	0.01928	0.15018	0
vGAT	0.93579	0.92328	0.93579	0

Table 2.10: mRNA levels significantly lowered in both SA and SZ

Gene	SA p-value	SZ p-value	max(SA p-value, SZ p-value)	Reject
CRIP1a	0.0675857	0.0351521	0.0675857	0
DAGLa	0.8493919	0.4039423	0.8493919	0
DAGLb	0.3051075	0.6001404	0.6001404	0
FAAH	0.7970013	0.8327632	0.8327632	0
GABA Receptor 1	0.0457178	0.018048	0.0457178	0
GAD67	0.0829396	0.0007101	0.0829396	0
GAT1	0.8135795	0.6534292	0.8135795	0
IFITM1	0.9451934	0.9998188	0.9998188	0
IFITM23	0.9760984	0.999981	0.999981	0
KCC2	0.6078946	0.9879924	0.9879924	0
Lhx6	0.1038305	0.0232733	0.1038305	0
MGL	0.705548	0.6672014	0.705548	0
Mu Opioid Receptor	0.9136047	0.9999644	0.9999644	0
NKCC1	0.0474918	0.0570873	0.0570873	0
OXSR1	0.9990466	1	1	0
Parvalbumin	0.0011287	0.0169198	0.0169198	0
RGS4	0.0073523	0.0000019	0.0073523	0
STK39	0.3252933	0.0130381	0.3252933	0
Somatostatin	0.0675898	0.0006275	0.0675898	0
TRPV1	0.9107743	0.9999965	0.9999965	0
WNK1	0.5547129	0.9304279	0.9304279	0
WNK3	0.999633	0.9999997	0.9999997	0
WNK4	0.5693099	0.9588297	0.9588297	0
mGluR1a	0.8411959	0.9997351	0.9997351	0
mGluR5	0.8498222	0.9807218	0.9807218	0
vGAT	0.0642093	0.0767167	0.0767167	0

2.A APPENDIX

2.A.1 Distribution of U and V

Let $\mathbf{U} = (U_1, U_2, \cdots, U_m)$ and $\mathbf{V} = (V_1, V_2, \cdots, V_m)$, where U_i and V_i are the test statistics defined in (2.2.4) for the unpaired case and in (2.2.5) for the paired case. Based on our assumptions, U_i and V_i are independent and thus $\mathbf{U}$ and $\mathbf{V}$ are independent.

1. In the unpaired case,

$$\mathbf{U}=\begin{pmatrix}U_1\\U_2\\\vdots\\U_m\end{pmatrix}=\begin{pmatrix}\frac{\bar{Y}_1^s-\bar{Y}_1^{sc}}{\sqrt{\frac{{\sigma_{y1}^s}^2}{n_s}+\frac{{\sigma_{y1}^{sc}}^2}{n_{sc}}}}\\\frac{\bar{Y}_2^s-\bar{Y}_2^{sc}}{\sqrt{\frac{{\sigma_{y2}^s}^2}{n_s}+\frac{{\sigma_{y2}^{sc}}^2}{n_{sc}}}}\\\vdots\\\frac{\bar{Y}_m^s-\bar{Y}_m^{sc}}{\sqrt{\frac{{\sigma_{ym}^s}^2}{n_s}+\frac{{\sigma_{ym}^{sc}}^2}{n_{sc}}}}\end{pmatrix},\quad \mathbf{V}=\begin{pmatrix}V_1\\V_2\\\vdots\\V_m\end{pmatrix}=\begin{pmatrix}\frac{\bar{Y}_1^b-\bar{Y}_1^{bc}}{\sqrt{\frac{{\sigma_{y1}^b}^2}{n_b}+\frac{{\sigma_{y1}^{bc}}^2}{n_{bc}}}}\\\frac{\bar{Y}_2^b-\bar{Y}_2^{bc}}{\sqrt{\frac{{\sigma_{y2}^b}^2}{n_b}+\frac{{\sigma_{y2}^{bc}}^2}{n_{bc}}}}\\\vdots\\\frac{\bar{Y}_m^b-\bar{Y}_m^{bc}}{\sqrt{\frac{{\sigma_{ym}^b}^2}{n_b}+\frac{{\sigma_{ym}^{bc}}^2}{n_{bc}}}}\end{pmatrix}.$$

For $d=s,b,sc,bc$, we have

$$(Y_{1j}^d,Y_{2j}^d,\cdots,Y_{mj}^d)\sim i.i.d.\mathbf{N_m}(\boldsymbol{\mu}_{\boldsymbol{y}}^{\boldsymbol{d}},\boldsymbol{\Sigma}_{\boldsymbol{y}}^{\boldsymbol{d}}).$$

So,

$$(\bar{Y}_1^d,\bar{Y}_2^d,\cdots,\bar{Y}_m^d)\sim i.i.d.\mathbf{N_m}(\boldsymbol{\mu}_{\boldsymbol{y}}^{\boldsymbol{d}},\frac{1}{n_d}\boldsymbol{\Sigma}_{\boldsymbol{y}}^{\boldsymbol{d}}).$$

And thus

$$\begin{aligned}(\bar{Y}_1^s-\bar{Y}_1^{sc},\bar{Y}_2^s-\bar{Y}_2^{sc},\cdots,\bar{Y}_m^s-\bar{Y}_m^{sc})&\sim\mathbf{N_m}(\boldsymbol{\mu}_{\boldsymbol{y}}^{\boldsymbol{s}}-\boldsymbol{\mu}_{\boldsymbol{y}}^{\boldsymbol{sc}},\frac{1}{n_s}\boldsymbol{\Sigma}_{\boldsymbol{y}}^{\boldsymbol{s}}+\frac{1}{n_{sc}}\boldsymbol{\Sigma}_{\boldsymbol{y}}^{\boldsymbol{sc}}),\\(\bar{Y}_1^b-\bar{Y}_1^{bc},\bar{Y}_2^b-\bar{Y}_2^{bc},\cdots,\bar{Y}_m^b-\bar{Y}_m^{bc})&\sim\mathbf{N_m}(\boldsymbol{\mu}_{\boldsymbol{y}}^{\boldsymbol{b}}-\boldsymbol{\mu}_{\boldsymbol{y}}^{\boldsymbol{bc}},\frac{1}{n_b}\boldsymbol{\Sigma}_{\boldsymbol{y}}^{\boldsymbol{b}}+\frac{1}{n_{bc}}\boldsymbol{\Sigma}_{\boldsymbol{y}}^{\boldsymbol{bc}}).\end{aligned}$$

Therefore,

$$\mathbf{U}\sim\mathbf{N_m}(\begin{pmatrix}\frac{\mu_{y1}^s-\mu_{y1}^{sc}}{\sqrt{\frac{{\sigma_{y1}^s}^2}{n_s}+\frac{{\sigma_{y1}^{sc}}^2}{n_{sc}}}}\\\frac{\mu_{y2}^s-\mu_{y2}^{sc}}{\sqrt{\frac{{\sigma_{y2}^s}^2}{n_s}+\frac{{\sigma_{y2}^{sc}}^2}{n_{sc}}}}\\\vdots\\\frac{\mu_{ym}^s-\mu_{ym}^{sc}}{\sqrt{\frac{{\sigma_{ym}^s}^2}{n_s}+\frac{{\sigma_{ym}^{sc}}^2}{n_{sc}}}}\end{pmatrix},Cov(\mathbf{U})),\quad \mathbf{V}\sim\mathbf{N_m}(\begin{pmatrix}\frac{\mu_{y1}^b-\mu_{y1}^{bc}}{\sqrt{\frac{{\sigma_{y1}^b}^2}{n_b}+\frac{{\sigma_{y1}^{bc}}^2}{n_{bc}}}}\\\frac{\mu_{y2}^b-\mu_{y2}^{bc}}{\sqrt{\frac{{\sigma_{y2}^b}^2}{n_b}+\frac{{\sigma_{y2}^{bc}}^2}{n_{bc}}}}\\\vdots\\\frac{\mu_{ym}^b-\mu_{ym}^{bc}}{\sqrt{\frac{{\sigma_{ym}^b}^2}{n_b}+\frac{{\sigma_{ym}^{bc}}^2}{n_{bc}}}}\end{pmatrix},Cov(\mathbf{V})).$$

where

$$
\begin{aligned}
&Cov(\mathbf{U})\\
&=\begin{pmatrix}
1 & \frac{\frac{\sigma_{y1}^{s}\sigma_{y2}^{s}\rho_{y12}^{s}}{n_s}+\frac{\sigma_{y1}^{sc}\sigma_{y2}^{sc}\rho_{y12}^{sc}}{n_{sc}}}{\sqrt{\frac{{\sigma_{y1}^{s}}^2}{n_s}+\frac{{\sigma_{y1}^{sc}}^2}{n_{sc}}}\sqrt{\frac{{\sigma_{y2}^{s}}^2}{n_s}+\frac{{\sigma_{y2}^{sc}}^2}{n_{sc}}}} & \cdots & \frac{\frac{\sigma_{y1}^{s}\sigma_{ym}^{s}\rho_{y1m}^{s}}{n_s}+\frac{\sigma_{y1}^{sc}\sigma_{ym}^{sc}\rho_{y1m}^{sc}}{n_{sc}}}{\sqrt{\frac{{\sigma_{y1}^{s}}^2}{n_s}+\frac{{\sigma_{y1}^{sc}}^2}{n_{sc}}}\sqrt{\frac{{\sigma_{ym}^{s}}^2}{n_s}+\frac{{\sigma_{ym}^{sc}}^2}{n_{sc}}}}\\
\frac{\frac{\sigma_{y1}^{s}\sigma_{y2}^{s}\rho_{y12}^{s}}{n_s}+\frac{\sigma_{y1}^{sc}\sigma_{y2}^{sc}\rho_{y12}^{sc}}{n_{sc}}}{\sqrt{\frac{{\sigma_{y1}^{s}}^2}{n_s}+\frac{{\sigma_{y1}^{sc}}^2}{n_{sc}}}\sqrt{\frac{{\sigma_{y2}^{s}}^2}{n_s}+\frac{{\sigma_{y2}^{sc}}^2}{n_{sc}}}} & 1 & \cdots & \frac{\frac{\sigma_{y2}^{s}\sigma_{ym}^{s}\rho_{y2m}^{s}}{n_s}+\frac{\sigma_{y2}^{sc}\sigma_{ym}^{sc}\rho_{y2m}^{sc}}{n_{sc}}}{\sqrt{\frac{{\sigma_{y2}^{s}}^2}{n_s}+\frac{{\sigma_{y2}^{sc}}^2}{n_{sc}}}\sqrt{\frac{{\sigma_{ym}^{s}}^2}{n_s}+\frac{{\sigma_{ym}^{sc}}^2}{n_{sc}}}}\\
\vdots & \vdots & \ddots & \vdots\\
\frac{\frac{\sigma_{y1}^{s}\sigma_{ym}^{s}\rho_{y1m}^{s}}{n_s}+\frac{\sigma_{y1}^{sc}\sigma_{ym}^{sc}\rho_{y1m}^{sc}}{n_{sc}}}{\sqrt{\frac{{\sigma_{y1}^{s}}^2}{n_s}+\frac{{\sigma_{y1}^{sc}}^2}{n_{sc}}}\sqrt{\frac{{\sigma_{ym}^{s}}^2}{n_s}+\frac{{\sigma_{ym}^{sc}}^2}{n_{sc}}}} & \frac{\frac{\sigma_{y2}^{s}\sigma_{ym}^{s}\rho_{y2m}^{s}}{n_s}+\frac{\sigma_{y2}^{sc}\sigma_{ym}^{sc}\rho_{y2m}^{sc}}{n_{sc}}}{\sqrt{\frac{{\sigma_{y2}^{s}}^2}{n_s}+\frac{{\sigma_{y2}^{sc}}^2}{n_{sc}}}\sqrt{\frac{{\sigma_{ym}^{s}}^2}{n_s}+\frac{{\sigma_{ym}^{sc}}^2}{n_{sc}}}} & \cdots & 1
\end{pmatrix},\\
&Cov(\mathbf{V})\\
&=\begin{pmatrix}
1 & \frac{\frac{\sigma_{y1}^{b}\sigma_{y2}^{b}\rho_{y12}^{b}}{n_b}+\frac{\sigma_{y1}^{bc}\sigma_{y2}^{bc}\rho_{y12}^{bc}}{n_{bc}}}{\sqrt{\frac{{\sigma_{y1}^{b}}^2}{n_b}+\frac{{\sigma_{y1}^{bc}}^2}{n_{bc}}}\sqrt{\frac{{\sigma_{y2}^{b}}^2}{n_b}+\frac{{\sigma_{y2}^{bc}}^2}{n_{bc}}}} & \cdots & \frac{\frac{\sigma_{y1}^{b}\sigma_{ym}^{b}\rho_{y1m}^{b}}{n_b}+\frac{\sigma_{y1}^{bc}\sigma_{ym}^{bc}\rho_{y1m}^{bc}}{n_{bc}}}{\sqrt{\frac{{\sigma_{y1}^{b}}^2}{n_b}+\frac{{\sigma_{y1}^{bc}}^2}{n_{bc}}}\sqrt{\frac{{\sigma_{ym}^{b}}^2}{n_b}+\frac{{\sigma_{ym}^{bc}}^2}{n_{bc}}}}\\
\frac{\frac{\sigma_{y1}^{b}\sigma_{y2}^{b}\rho_{y12}^{b}}{n_b}+\frac{\sigma_{y1}^{bc}\sigma_{y2}^{bc}\rho_{y12}^{bc}}{n_{bc}}}{\sqrt{\frac{{\sigma_{y1}^{b}}^2}{n_b}+\frac{{\sigma_{y1}^{bc}}^2}{n_{bc}}}\sqrt{\frac{{\sigma_{y2}^{b}}^2}{n_b}+\frac{{\sigma_{y2}^{bc}}^2}{n_{bc}}}} & 1 & \cdots & \frac{\frac{\sigma_{y2}^{b}\sigma_{ym}^{b}\rho_{y2m}^{b}}{n_b}+\frac{\sigma_{y2}^{bc}\sigma_{ym}^{bc}\rho_{y2m}^{bc}}{n_{bc}}}{\sqrt{\frac{{\sigma_{y2}^{b}}^2}{n_b}+\frac{{\sigma_{y2}^{bc}}^2}{n_{bc}}}\sqrt{\frac{{\sigma_{ym}^{b}}^2}{n_b}+\frac{{\sigma_{ym}^{bc}}^2}{n_{bc}}}}\\
\vdots & \vdots & \ddots & \vdots\\
\frac{\frac{\sigma_{y1}^{b}\sigma_{ym}^{b}\rho_{y1m}^{b}}{n_b}+\frac{\sigma_{y1}^{bc}\sigma_{ym}^{bc}\rho_{y1m}^{bc}}{n_{bc}}}{\sqrt{\frac{{\sigma_{y1}^{b}}^2}{n_b}+\frac{{\sigma_{y1}^{bc}}^2}{n_{bc}}}\sqrt{\frac{{\sigma_{ym}^{b}}^2}{n_b}+\frac{{\sigma_{ym}^{bc}}^2}{n_{bc}}}} & \frac{\frac{\sigma_{y2}^{b}\sigma_{ym}^{b}\rho_{y2m}^{b}}{n_b}+\frac{\sigma_{y2}^{bc}\sigma_{ym}^{bc}\rho_{y2m}^{bc}}{n_{bc}}}{\sqrt{\frac{{\sigma_{y2}^{b}}^2}{n_b}+\frac{{\sigma_{y2}^{bc}}^2}{n_{bc}}}\sqrt{\frac{{\sigma_{ym}^{b}}^2}{n_b}+\frac{{\sigma_{ym}^{bc}}^2}{n_{bc}}}} & \cdots & 1
\end{pmatrix}.
\end{aligned}
$$

2. In the paired case,

$$
\mathbf{U}=\begin{pmatrix}U_1\\U_2\\\vdots\\U_m\end{pmatrix}=\begin{pmatrix}\frac{\bar{X}_1^s}{\sqrt{\frac{{\sigma_{x1}^{s}}^2}{n_{sp}}}}\\\frac{\bar{X}_2^s}{\sqrt{\frac{{\sigma_{x2}^{s}}^2}{n_{sp}}}}\\\vdots\\\frac{\bar{X}_m^s}{\sqrt{\frac{{\sigma_{xm}^{s}}^2}{n_{sp}}}}\end{pmatrix},\quad \mathbf{V}=\begin{pmatrix}V_1\\V_2\\\vdots\\V_m\end{pmatrix}=\begin{pmatrix}\frac{\bar{X}_1^b}{\sqrt{\frac{{\sigma_{x1}^{b}}^2}{n_{bp}}}}\\\frac{\bar{X}_2^b}{\sqrt{\frac{{\sigma_{x2}^{b}}^2}{n_{bp}}}}\\\vdots\\\frac{\bar{X}_m^b}{\sqrt{\frac{{\sigma_{xm}^{b}}^2}{n_{bp}}}}\end{pmatrix}.
$$

For $f=s,b$, we have

$$(X_{1j}^f, X_{2j}^f, \cdots, X_{mj}^f)\sim i.i.d.\mathbf{N_m}(\boldsymbol{\mu_x^f},\boldsymbol{\Sigma_x^f}).$$

So,

$$(\bar{X}_1^f,\bar{X}_2^f,\cdots,\bar{X}_m^f)\sim i.i.d.\mathbf{N_m}(\boldsymbol{\mu_x^f},\frac{1}{n_{fp}}\boldsymbol{\Sigma_x^f}).$$

Therefore,

$$\mathbf{U} \sim \mathbf{N_m}(\begin{pmatrix} \frac{\mu^s_{y1}-\mu^{sc}_{y1}}{\sqrt{\frac{{\sigma^s_{x1}}^2}{n_{sp}}}} \\ \frac{\mu^s_{y2}-\mu^{sc}_{y2}}{\sqrt{\frac{{\sigma^s_{x2}}^2}{n_{sp}}}} \\ \vdots \\ \frac{\mu^s_{ym}-\mu^{sc}_{ym}}{\sqrt{\frac{{\sigma^s_{xm}}^2}{n_{sp}}}} \end{pmatrix}, Cov(\mathbf{U})), \quad \mathbf{V} \sim \mathbf{N_m}(\begin{pmatrix} \frac{\mu^b_{y1}-\mu^{bc}_{y1}}{\sqrt{\frac{{\sigma^b_{x1}}^2}{n_{bp}}}} \\ \frac{\mu^b_{y2}-\mu^{bc}_{y2}}{\sqrt{\frac{{\sigma^b_{x2}}^2}{n_{bp}}}} \\ \vdots \\ \frac{\mu^b_{ym}-\mu^{bc}_{ym}}{\sqrt{\frac{{\sigma^b_{xm}}^2}{n_{bp}}}} \end{pmatrix}, Cov(\mathbf{V})),$$

where

$$Cov(\mathbf{U}) = \begin{pmatrix} 1 & \rho^s_{x12} & \cdots & \rho^s_{x1m} \\ \rho^s_{x12} & 1 & \cdots & \rho^s_{x2m} \\ \vdots & \vdots & \ddots & \vdots \\ \rho^s_{x1m} & \rho^s_{x2m} & \cdots & 1 \end{pmatrix}, \quad Cov(\mathbf{V}) = \begin{pmatrix} 1 & \rho^b_{x12} & \cdots & \rho^b_{x1m} \\ \rho^b_{x12} & 1 & \cdots & \rho^b_{x2m} \\ \vdots & \vdots & \ddots & \vdots \\ \rho^b_{x1m} & \rho^b_{x2m} & \cdots & 1 \end{pmatrix}.$$

3.0 COMPARING THE POPULATION WITH DYSFUNCTION AND THE HEALTHY POPULATION

In this chapter we compare the identified neurobiological characteristics between the population with dysfunction in the construct of interest and the healthy population through mean differences and quantile differences.

3.1 MOTIVATION FOR THE COMPARISON

What has been done in the last chapter is identifying some neurobiological characteristics that are significantly up-regulated or down-regulated across all the DSM diagnoses relevant to a psychiatric construct of interest. The goal of the identification is to determine which neurobiological characteristics are significantly involved in the dysfunction of the construct, for instance, working memory. These neurobiological characteristics provide great opportunities to understand the mechanism of the illnesses in the construct. More precise and effective treatments for mental illness may be developable in the future based on our knowledge about the relationship between these identified neurobiological characteristics and the construct.

There are various ways to further investigate the neurobiological characteristics to understand their roles in the mechanism they lead to dysfunction in the construct. Ideally, if there are behavior measures in those post-mortem tissue databases, an RDoC study can be used to establish the brain-behavior relationship, which is the relationship between the neurobiological characteristics and behavior symptoms around the particular construct over the general population. However, no behavior measures are available for the subjects in the post-mortem tissue databases. All we can study are the neurobiological measures themselves.

In spite of the unavailability of behavior measures, we can still carry out some studies in the RDoC spirit by focusing on a specific construct. The comparisons of the neurobiological characteristics between the healthy population and the population with dysfunction are worthwhile because they provide information about where and how the two populations differ. In this dissertation, we compare the means and quantiles in the neurobiological characteristics between the two populations. In these comparisons, as discussed previously, the random sample for the population with dysfunction is substituted with a mixture of random samples from relevant DSM diagnoses. The relative proportion for each DSM diagnosis among the entire population with dysfunction can be estimated from historical data. If the construct is about a mental function for which some particular clinic exists, then the historical composition of patients seen in the clinic can be used to estimate the relative proportions. These patients would all present the dysfunction in this particular construct and have been later given some DSM diagnosis in the clinic. For example, if researchers want to estimate the relative proportions of the DSM diagnoses among the psychotic population, the patients seen in a psychosis clinic can be used because they all have psychosis to some degree. Suppose among those patients seen in a psychosis clinic, three DSM diagnoses are later made. Also suppose among these patients, 20% have been later diagnosed with bipolar disorder, 10% have been diagnosed to have major depressive disorder and 70% have been diagnosed with schizophrenia, then the relative proportions of bipolar disorder, major depressive disorder and schizophrenia among the population with psychosis are 20%, 10% and 70%, respectively. If no psychiatric clinic exists for some construct, then the relative proportions can be calculated using Bayes theorem based on historical data from epidemiology studies. The prevalence for each DSM diagnosis and the probability of dysfunction in the construct of interest within each DSM diagnosis are needed to calculate the relative proportions. For example, there are no working memory clinics, and thus if we want to estimate the relative proportion of each DSM diagnosis among the population with low working memory, epidemiology data can be used. Suppose for illustration purpose, low working memory is reported only in schizophrenia and bipolar disorder. Also suppose the prevalences for schizophrenia and bipolar disorder are 4% and 6%, respectively. If the probabilities of low working memory among schizophrenia patients and bipolar disorder patients are 80% and 90%, respectively, then the relative proportion of schizophrenia among the population with low working memory is $80\% * 4\%/(80\% * 4\% + 90\% * 6\%) = 37\%$. Similarly, the relative proportion for bipolar disorder is 63%. In theory, the relative proportion of DSM diagnosis among the population with dysfunction in any particular construct can be calculated using epidemiological data, however, we think it would be easier to use the patients composition in a clinic if one exists. As noted in Assumption 3.3, we understand variability exists in estimating the relative proportions. However, they are considered to be known and fixed to simplify our problem.

3.2 COMPARISON OF THE POPULATIONS THROUGH THE MEANS OF NEUROBIOLOGICAL CHARACTERISTICS

In Chapter 2, we identify the neurobiological characteristics by doing hypothesis testing about their means. Although the neurobiological characteristics get identified if they are significantly up-regulated (or down-regulated) in each individual DSM diagnosis compared to the healthy population, it is still unknown how different they are collectively in the population with dysfunction compared to the healthy population. And it is this collective mean in the population with dysfunction in a construct that the researchers would like to know in the RDoC spirit. Therefore in this section, we want to estimate the differences in the means of the neurobiological characteristics between the healthy population and the population with dysfunction in the construct of interest, with adjustment for confounding covariates.

3.2.1 Layout of the Comparison through Means

Suppose there are k DSM diagnoses that are related to the construct of interest due to having symptoms relevant to the construct. As introduced in Chapter 1, the population with dysfunction in this construct is the original population we want to study and it is substituted by the mixture of k DSM diagnoses, each with a different relative proportion. Let D_i be the DSM diagnosis with relative proportion π_i among the population with dysfunction, where $\pi_1 < \pi_2 < \cdots < \pi_k$ and $\sum_{i=1}^k \pi_i = 1$. In other words, the DSM diagnosis D_i is defined by its relative proportion π_i. Among the population with dysfunction in the construct of interest, D_1 is the DSM diagnosis with the smallest relative portion and D_k is the one with the largest relative proportion. Once the relative proportion among the population with dysfunction is known, each D_i refers to a specific DSM diagnosis. Suppose there are n_i subjects in D_i and n_0 subjects from the healthy population.

Let Δ be the true difference in the neurobiological characteristic between the population with dysfunction in a particular construct and the healthy population. Our goal is to estimate Δ, i.e., obtain $\hat{\Delta}$. Because the population with dysfunction is represented by a mixture of the k DSM diagnoses, the difference in the neurobiological characteristic between the population with dysfunction and the healthy population can be written as a linear combination of the differences between each DSM diagnosis and the healthy population. In other words, if we use Δ_i to denote the true difference in the neurobiological characteristic between DSM diagnosis D_i and the healthy population, then

$$\Delta = \sum_{i=1}^{k} \pi_i \Delta_i. \tag{3.2.1}$$

If $\hat{\Delta}_i$ is the estimate of Δ_i, then $\sum_{i=1}^k \pi_i \hat{\Delta}_i$ can be used to estimate Δ, and we define

$$\hat{\Delta} = \sum_{i=1}^{k} \pi_i \hat{\Delta}_i. \tag{3.2.2}$$

The notation used in this chapter is summarized in Table 3.1. The notation in Table 3.1 that has not been already introduced is defined later in this chapter.

Table 3.1: Notations in the Comparison between the population with dysfunction and the healthy populations

Δ	true difference in the neurobiological characteristic between the population with dysfunction and healthy populations
$\hat{\Delta}$	estimated difference in the neurobiological characteristic between the population with dysfunction and healthy populations
k	number of DSM diagnoses
D_i	the ith DSM diagnosis defined by the relative proportion
$D_{(i)}$	the DSM diagnosis with sample size $n_{(i)}$
π_i	relative proportion of D_i in the population with dysfunction
$\pi_{(i)}$	relative proportion of $D_{(i)}$ in the population with dysfunction
Δ_i	true difference in the neurobiological characteristic between D_i and healthy population
$\hat{\Delta}_i$	estimated difference in the neurobiological characteristic between D_i and healthy population
n_i	number of subjects used for D_i
$n_{(i)}$	ith order statistic of $n_1, n_2, \cdots, n_k$
n_0	number of subjects used for healthy control group
$y_{(i)j}$	measurement of the neurobiological characteristic of the subject in $D_{(i)}$ and the jth match
$\epsilon_{(i)j}$	random error of observation $y_{(i)j}$
β_i	diagnosis effect for D_i
$\beta_{(i)}$	diagnosis effect for $D_{(i)}$
β_0	diagnosis effect for the healthy controls
γ_j	match effect for M_j
$\boldsymbol{\mathcal{X}}$	design matrix
$\boldsymbol{y}$	vector of all the observations
$\boldsymbol{\beta}$	vector of parameters
$\boldsymbol{\epsilon}$	vector of all the random errors

3.2.2 Assumptions for Comparison through Means

Before we are able to work out a formula to estimate Δ, we need some assumptions for the form of the design we use. Different designs give rise to different formulae. Because we want to adjust for covariates, a matched sample design is used where subjects are matched as closely as possible by some of the important covariates. For post-mortem studies, these covariates could be sex, age, post-mortem interval and so on. Matching not only reduces the confound effects from this covariates but also reduces the variation introduced by processing the tissue samples in different batches. In this dissertation some assumptions are needed for the matching so that a specific form of design can be obtained.

Assumption 3.1. *In each match, there is exactly one healthy control subject and at most one subject in each DSM diagnosis, and it is assumed at least one DSM diagnosis has a subject.*

Matches with no healthy control subject or more than one subject in any DSM diagnosis or the healthy control group are not considered here, even if the tissue processing capacity is large enough to process more than one subject in each DSM diagnosis or the control group. This assumption is to simplify the formulation of the problem. Otherwise, we need some other parameter for the sample processing capacity. Also we have to deal with differing controls if more than one healthy subject is allowed in each match, which is shown to be complicated in Wu & Sampson (2012). Based on this assumption, with k DSM diagnoses, the number of DSM diagnosed subjects in each match could range from 1 to k. For example when $k = 3$, without regard to the specific DSM diagnosis, a match could have one of the following three possibilities: one healthy control and three DSM diagnosed subjects, one from each diagnosis; one healthy control and two DSM diagnosed subjects, each from a different diagnosis; or one healthy control and one DSM diagnosed subject. Therefore in this example, in one match, at most four subjects and at least two subjects are matched together. Also according to this assumption, it follows immediately that the number of healthy control subjects is equal to the number of matches in the design. This means if the total number of subjects that can be processed is held fixed and we could have as few matches as possible, the number of healthy control subjects can be kept as small as possible. This is important in cases where the budget of a study is limited because we could then spend more resources on subjects with dysfunction. Actually Assumption 3.2 allows us to have the smallest number of matches with the sample size for each DSM diagnosis given.

One thing that needs to be mentioned is that for each match, each subject from the DSM diagnosis could have its own healthy control subject from the post-mortem tissue database it is sampled from. In cases where there are more than one candidate healthy subject, we can choose the one that has a closer mean distance in the matching variables from all the subjects with DSM diagnoses in that match. For example, suppose DSM diagnoses D_1 and D_2 are related to the construct of interest and we want to match a subject from a post-mortem tissue database for D_1 (denoted as Sub1) to one from a post-mortem tissue database for D_2 (denoted as Sub2) based on age and PMI. The ages and PMIs for the two subjects are listed in Table 3.2. Suppose each of Sub 1 and Sub 2 has its own healthy control in its own database (denoted as Con1 and Con2, respectively) and the age and PMI for the two healthy subjects are also listed. In order to determine which healthy control subject to use for this match, we can calculate for each of Con1 and Con2 the distance in age and PMI it has between each of Sub1 and Sub2. The control subject with a smaller mean distance from Sub1 and Sub2 would be chosen. For the hypothetical example in Table 3.2, because Con1 has a smaller mean distance, it is included into this match. Here the metric for distance is Euclidean distance, and clearly other metrics can be used.

Table 3.2: Illustration of how to choose among multiple healthy control subjects

Subject	Age	PMI	Distance to Sub1	Distance to Sub2	Mean Distance
Sub1	42	26.1			
Con1	40	29.1	3.61	9.14	6.37
Sub2	46	22.2			
Con2	47	15.3	11.90	6.97	9.44

Assumption 3.2. *Each match allows as many DSM diagnoses as possible.*

This assumption is a very critical one as it says when processing the tissue samples, we try to include as many DSM diagnosed subjects as possible into each match. In general, if nothing is assumed about the number of DSM diagnoses in one match, more than one type of design can be obtained even with the same configuration of $(n_1, n_2, \cdots, n_k)$, just by varying the number of matches, i.e., number of healthy controls. For example, consider a simple case where only two DSM diagnoses are involved and we have four subjects in D_1 and three subjects in D_2, i.e., $k = 2, n_1 = 4, n_2 = 3$. All of the four designs listed in Table 3.3 are possible. Here each row is a match.

Table 3.3: Possible designs with $k = 2, n_1 = 4, n_2 = 3$

(a) Desirable

D_1	D_2	Control
×	×	×
×	×	×
×	×	×
×		×

(b) Undesirable

D_1	D_2	Control
×	×	×
×	×	×
×		×
×		×
	×	×

(c) Undesirable

D_1	D_2	Control
×	×	×
×		×
×		×
×		×
	×	×
	×	×

(d) Undesirable

D_1	D_2	Control
×		×
×		×
×		×
×		×
	×	×
	×	×
	×	×

As shown in the above tables, with $n_1, n_2, \cdots, n_k$ holding constant, there are various possible designs. These possible designs are different from each other in the number of healthy control samples, and thus also different in the number of matches. This is problematic because there could be a large number of possible designs each with a different design matrix and thus it is difficult to write down the design matrix without knowing which design is used. However, according to Assumption 3.2, among all the designs shown in Table 3.3, only the one in (a) is considered in this dissertation. The ones in (b), (c) and (d) are not considered because they all have matches that could have included one more DSM diagnosis, thus contradicts Assumption 3.2. For instance, sample from D_2 in the last row of Table (b) could have been processed in the third or fourth match. In other words, with Assumption 3.2, designs represented in Tables (b), (c) and (d) can all be converted to that represented in Table (a) by combining some of the matches and dropping some healthy control subjects. In reality, we understand that the combination could possibly make the subjects within a match have very different covariates. However, it is still justifiable to process tissue samples from these subjects in one match because we think the batch effect brings in a larger variation than that introduced by the different covariates. Since in each match we try to include the most number of DSM diagnoses, Assumption 3.2 improves the efficiency of the tissue processing.

An immediate result of Assumption 3.2 is that it indicates the number of matches is as small as possible. Together with Assumption 3.1, the sample size for the healthy population is as small as possible. Furthermore, we can show that the sample size for the healthy population is equal to the maximum of the sample sizes in the DSM diagnoses. This result is stated and proved as the following.

Result 3.1. *Under Assumptions 3.1 and 3.2,* $n_0 = \max(n_1, n_2, \cdots, n_k)$.

Proof. The form of the design under the assumption can be easily stated if $n_{(1)}, n_{(2)}, \cdots, n_{(k)}$ are the order statistics of $n_1, n_2, \cdots, n_k$. Let $D_{(i)}$ be the DSM diagnosis with sample size $n_{(i)}$; $\pi_{(i)}$ be the relative proportion of $D_{(i)}$ among the population with dysfunction and $N_{(i)}$ be the number of subjects available for $D_{(i)}$ from all possible post-mortem tissue databases.

Based on Assumption 3.2, we want each match to contain as many DSM diagnoses as possible. Therefore, because there are at least $n_{(1)}$ subjects in each DSM diagnosis, the study always has $n_{(1)}$ matches each with k different DSM diagnosed subjects and one healthy control subject. Each of the k DSM diagnosed subjects within these $n_{(1)}$ matches comes from a different DSM diagnosis among $D_{(1)}, D_{(2)}, \cdots, D_{(k)}$. Similarly there are $(n_{(2)} - n_{(1)})$ matches each with $(k-1)$ DSM diagnosed subjects and one healthy control. Each of the $(k-1)$ DSM diagnosed subjects comes from a different DSM diagnosis among $D_{(2)}, \cdots, D_{(k)}$. There are $(n_{(3)} - n_{(2)})$ matches each with $(k-2)$ DSM diagnosed subjects and one healthy control. Each of the $(k-2)$ DSM diagnosed subjects within these matches comes from a different DSM diagnosis among $D_{(3)}, \cdots, D_{(k)}$. The list goes on until there are $(n_{(k)} - n_{(k-1)})$ matches each with only one DSM diagnosed subject from $D_{(k)}$ and one healthy control subject. So in every match, there is a healthy control subject and a subject from $D_{(k)}$. It immediately follows that the number of subjects in the healthy population is equal to $n_{(k)}$, which is the maximum of the sample sizes in the DSM diagnoses, i.e., $n_0 = \max(n_1, n_2, \cdots, n_k) = n_{(k)}$. □

To further simplify the problem, we also make the following assumption about the relative proportions.

Assumption 3.3. *The relative proportion π_i's are assumed to be known.*

Thus we assume π_i to be known and fixed, even though it is estimated from the patient composition in a clinic. The inherent variability in estimating π_i is not taken into account in our problem.

3.2.3 Study Design for Comparison through Means

In order that the study can be represented clearly and the design matrix be given in a general form for any design satisfying the assumptions, the order of the sample sizes for each DSM diagnosis needs to be known. Therefore, using the notation introduced in the proof of Result 3.1, the study design under consideration can be represented in a specific form called "triangular design" by us as illustrated in Table 3.3(a). In general, the triangular design is shown in Table 3.4.

As we can see in the above table, if the DSM diagnoses are listed with increasing sample sizes from the left to the right, the representation of the design takes the triangular shape. And this is where the name of the "triangular" design comes from. Under our assumptions, if the sample sizes $n_1, n_2, \cdots, n_k$ are fixed, then the triangular design is unique. For example, the desirable case in Table 3.3(a) is in the triangular form if we switch the columns D_1 and D_2.

Table 3.4: Triangular Design Under the Assumptions

$D_{(1)}$	$D_{(2)}$	$D_{(3)}$	$\cdots$	$D_{(k)}$	Control	
×	×	×	$\cdots$	×	×	$n_{(1)}$ matches
	×	×	$\cdots$	×	×	$(n_{(2)} - n_{(1)})$ matches
		×	$\cdots$	×	×	$(n_{(3)} - n_{(2)})$ matches
			$\vdots$	$\vdots$	$\vdots$	$\vdots$
				×	×	$(n_{(k)} - n_{(k-1)})$ matches
$n_{(1)}$	$n_{(2)}$	$n_{(3)}$	$\cdots$	$n_{(k)}$	$n_{(k)}$	sample size for each group

3.2.4 Modeling for Comparison through Means

In each of the subjects in the study, one or more neurobiological characteristics are measured and the differences in these neurobiological characteristics Δ are estimated between the healthy population and the population with dysfunction in the psychiatric construct of interest. Our estimate of the difference is obtained from an ANOVA model of the data. The design matrix and estimator of the difference are specified based on the ANOVA model.

We assume only one neurobiological characteristic is measured and focus on the univariate ANOVA model in this section. The multivariate case is shown in Appendix 3.A.2.

3.2.4.1 ANOVA Model Because the study design is represented using the notation $D_{(i)}$ and $n_{(i)}$, the ANOVA model is specified in the same way. Let $y_{(i)j}$ and y_{0j} be the measurements of the neurobiological characteristics in the jth match for the subjects in $D_{(i)}$ and the healthy control group, respectively. Let $\beta_{(i)}$ and β_0 denote the diagnosis effect for $D_{(i)}$ and the healthy control respectively. Let γ_j be the match effect for the jth match. Let $\epsilon_{(i)j}$ and ϵ_{0j} be the error terms for observations $y_{(i)j}$ and y_{0j}, respectively. It is assumed that the distributions of $\epsilon_{(i)j}$ and ϵ_{0j} are iid $N(0, \sigma^2)$. The ANOVA model can be written as:

$$\begin{aligned} y_{(i)j} &= \beta_{(i)} + \gamma_j + \epsilon_{(i)j}, \quad i = 1, 2, \cdots, k, \quad j = 1, 2, \cdots, n_{(i)}; \\ y_{0j} &= \beta_0 + \gamma_j + \epsilon_{0j}, \quad j = 1, 2, \cdots, n_{(k)}. \end{aligned} \tag{3.2.3}$$

Under the model in (3.2.3), if $\Delta_{(i)}$ denotes the true difference of the neurobiological characteristic between $D_{(i)}$ and the healthy population with adjustment for the $\gamma_j's$, then

$$\Delta_{(i)} = \beta_{(i)} - \beta_0, \quad i = 1, 2, \cdots, k.$$

Let $\widehat{\beta_{(i)} - \beta_0}$ denote the estimate of $\beta_{(i)} - \beta_0$, then the estimate of Δ is $\hat{\Delta}_{(i)} = \widehat{\beta_{(i)} - \beta_0}, (i = 1, 2, \cdots, k)$.

By definition, $\Delta_{(1)}, \Delta_{(2)}, \cdots, \Delta_{(k)}$ is just a permutation of $\Delta_1, \Delta_2, \cdots, \Delta_k$, so the sums $\sum_{i=1}^{k} \pi_i \Delta_i$ and $\sum_{i=1}^{k} \pi_{(i)} \Delta_{(i)}$ are equal. For the same reason, the sums $\sum_{i=1}^{k} \pi_i \hat{\Delta}_i$ and $\sum_{i=1}^{k} \pi_{(i)} \hat{\Delta}_{(i)}$ are equal. Therefore, the expressions in (3.2.1) and (3.2.2) become:

$$\Delta = \sum_{i=1}^{k} \pi_{(i)} \Delta_{(i)} = \sum_{i=1}^{k} \pi_{(i)} (\beta_{(i)} - \beta_0),$$
$$\hat{\Delta} = \sum_{i=1}^{k} \pi_{(i)} \hat{\Delta}_{(i)} = \sum_{i=1}^{k} \pi_{(i)} (\widehat{\beta_{(i)} - \beta_0}).$$

Otherwise stated, in order to estimate Δ, we only need to sort the DSM diagnoses by their sample sizes and then derive the estimator for $\beta_{(i)} - \beta_0$.

3.2.4.2 Estimator of Δ In this section, the design matrix $\mathcal{X}$ of the ANOVA model and the estimates $\widehat{\beta_{(i)} - \beta_0}$ and $\hat{\Delta}$ in terms of the raw data $y_{(i)j}$ are given. Before expressing $\widehat{\beta_{(i)} - \beta_0}$ in the raw data, the estimability of $\beta_{(i)} - \beta_0$ is proved with detailed proof in Appendix.

For the triangular design represented in Table 3.4, the stated ANOVA model in (3.2.3) can be rewritten in matrix form as:

$$\underbrace{\begin{bmatrix} y_{(1)1} \\ y_{(1)2} \\ \vdots \\ y_{(1)n_{(1)}} \\ y_{(2)1} \\ y_{(2)2} \\ \vdots \\ y_{(2)n_{(2)}} \\ \vdots \\ y_{(k)1} \\ y_{(k)2} \\ \vdots \\ y_{(k)n_{(k)}} \\ y_{01} \\ y_{02} \\ \vdots \\ y_{0n_{(k)}} \end{bmatrix}}_{\boldsymbol{y}} = \underbrace{\begin{bmatrix} 1 & \cdots & 0 & 0 & 1 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & \cdots & 0 \\ 1 & \cdots & 0 & 0 & 0 & 1 & \cdots & 0 & 0 & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & \cdots & 0 & 0 & 0 & 0 & \cdots & 1 & 0 & 0 & \cdots & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & 1 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & 0 & 1 & \cdots & 0 & 0 & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & & & \ddots & & & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 1 & 0 & 1 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & \cdots & 0 \\ 0 & \cdots & 1 & 0 & 0 & 1 & \cdots & 0 & 0 & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & & & & \ddots & & & & \vdots \\ 0 & \cdots & 1 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & \cdots & 1 \\ 0 & \cdots & 0 & 1 & 1 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 1 & 0 & 1 & \cdots & 0 & 0 & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & & & & \ddots & & & & \vdots \\ 0 & \cdots & 0 & 1 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & \cdots & 1 \end{bmatrix}}_{\mathcal{X}} \underbrace{\begin{bmatrix} \beta_{(1)} \\ \vdots \\ \beta_{(k)} \\ \beta_0 \\ \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_{n_{(1)}} \\ \gamma_{n_{(1)}+1} \\ \gamma_{n_{(1)}+2} \\ \vdots \\ \gamma_{n_{(2)}} \\ \vdots \\ \gamma_{n_{(k)}} \end{bmatrix}}_{\boldsymbol{\beta}} + \underbrace{\begin{bmatrix} \epsilon_{(1)1} \\ \epsilon_{(1)2} \\ \vdots \\ \epsilon_{(1)n_{(1)}} \\ \epsilon_{(2)1} \\ \epsilon_{(2)2} \\ \vdots \\ \epsilon_{(2)n_{(2)}} \\ \vdots \\ \epsilon_{(k)1} \\ \epsilon_{(k)2} \\ \vdots \\ \epsilon_{(k)n_{(k)}} \\ \epsilon_{01} \\ \epsilon_{02} \\ \vdots \\ \epsilon_{0n_{(k)}} \end{bmatrix}}_{\boldsymbol{\epsilon}} . \tag{3.2.4}$$

Here $\boldsymbol{y}$ is the data vector for all the observations, $\mathcal{X}$ is the design matrix, $\boldsymbol{\beta}$ is the vector of parameters and $\boldsymbol{\epsilon}$ is the random error vector for $\boldsymbol{y}$. Because both $\epsilon_{(i)j}(i = 1, 2, \cdots, k; j = 1, 2, \cdots, n_{(i)})$ and $\epsilon_{0j}(j = 1, 2, \cdots, n_{(k)})$ are assumed to follow iid $N(0, \sigma^2)$, the distribution of $\boldsymbol{y}$ is the multivariate normal distribution with mean $\mathcal{X}\boldsymbol{\beta}$ and variance matrix $\sigma^2 I_{n_{(k)}+\sum_{i=1}^{k} n_{(i)}}$, where $I_{n_{(k)}+\sum_{i=1}^{k} n_{(i)}}$ is the identity matrix with dimension $n_{(k)} + \sum_{i=1}^{k} n_{(i)}$.

Now we can show $\beta_{(i)} - \beta_0$ is estimable under the ANOVA model. Let $l_{(i)}$ be the column vector such that $l'_{(i)}\boldsymbol{\beta} = \beta_{(i)} - \beta_0(i = 1, 2, \cdots, k)$. Obviously $l_{(i)}$ is a vector with dimension $k+1+n_{(k)}$ and it has 0 everywhere except that the ith element is 1 and the $(k+1)$th element is -1.

$$\begin{array}{ccccccccccccc} l'_{(i)} & = & (0\cdots & 0 & 1 & 0 & \cdots & 0 & -1 & 0 & \cdots & 0) & \\ \uparrow & & & & \uparrow & & & & \uparrow & & & & . \\ (k+1+n_{(k)})\times 1 & & & & i\text{th} & & & & (k+1)\text{th} & & & & \end{array} \tag{3.2.5}$$

From linear model theory, we know that if $l_{(i)} = \mathcal{X}'\mathcal{X}\tau_{(i)}$ holds for some column vector $\tau_{(i)}$ with dimension $k+1+n_{(k)}$, then $l'_{(i)}\boldsymbol{\beta}$ is estimable and $\widehat{\beta_{(i)}-\beta_0}$ can be expressed as:

$$\widehat{\beta_{(i)}-\beta_0} = l'_{(i)}\hat{\boldsymbol{\beta}} = l'_{(i)}(\mathcal{X}'\mathcal{X})^{-}\mathcal{X}'\boldsymbol{y} = \tau'_{(i)}\mathcal{X}'\mathcal{X}(\mathcal{X}'\mathcal{X})^{-}\mathcal{X}'\boldsymbol{y} = \tau'_{(i)}\mathcal{X}'\boldsymbol{y}. \tag{3.2.6}$$

In other words, if we can find a $\tau_{(i)}$ such that $l_{(i)} = \mathcal{X}'\mathcal{X}\tau_{(i)}$ holds, $\beta_{(i)}-\beta_0$ is estimable. The existence of $\tau_{(i)}$ is shown in Appendix 3.A.1.

Because $\beta_{(i)}-\beta_0$ is estimable, $\widehat{\beta_{(i)}-\beta_0}$ can be expressed as $\tau'_{(i)}\mathcal{X}'\boldsymbol{y}$ as shown in (3.2.6). If vector $\boldsymbol{\pi}' = (\pi_{(1)}\,\pi_{(2)}\,\cdots\,\pi_{(k)})$ and matrix $T = (\tau_{(1)}\,\tau_{(2)}\,\cdots\,\tau_{(k)})$, then based on (3.2.2) we have

$$\begin{aligned} \hat{\Delta} &= \sum_{i=1}^{k}\pi_{(i)}(\widehat{\beta_{(i)}-\beta_0}) \\ &= \sum_{i=1}^{k}\pi_{(i)}(\tau'_{(i)}\mathcal{X}'\boldsymbol{y}) \\ &= (\pi_{(1)}\,\pi_{(2)}\,\cdots\,\pi_{(k)})\begin{pmatrix}\tau'_{(1)}\\ \tau'_{(2)}\\ \vdots\\ \tau'_{(k)}\end{pmatrix}\mathcal{X}'\boldsymbol{y} \\ &= \boldsymbol{\pi}'T'\mathcal{X}'\boldsymbol{y}. \end{aligned} \tag{3.2.7}$$

3.3 COMPARISON OF THE POPULATIONS THROUGH THE QUANTILES OF NEUROBIOLOGICAL CHARACTERISTICS

As discussed previously, the RDoC intends to learn the fundamental biological mechanism for mental illness and to incorporate the biological measures into a new diagnostic approach. The neurobiological characteristics identified in Chapter 2 are the potential targets to be studied in research because we already know they are significantly different in the population with dysfunction than those in the healthy population. This is because their means in each DSM diagnosis are significantly different from those in the healthy population. However, in order to understand the mechanism why these neurobiological characteristics could lead to mental illness and to possibly include them in the future diagnosis, research with dimensional approaches needs to be done under the RDoC framework. For example, researchers may want to compare the distributions of the neurobiological characteristics between the healthy population and the population with dysfunction so that they can establish the norm of the neurobiological characteristic and know what values of the measured neurobiological characteristic indicate illness for a patient.

The comparison through the means described in Section 3.2 would inform researchers about the difference in the distribution of the neurobiological characteristic only when the distribution in the population with dysfunction is unimodal and relatively symmetric. The reason is that what has been compared is just a central location and when the distribution is multi-modal or highly skewed, the comparison of a central location does not provide the picture of the full distribution. For example, suppose two DSM diagnoses, D_1 and D_2, are relevant to a construct of interest with relative proportions 0.3 and 0.7, respectively, then the population with dysfunction can be considered as a mixture of these two DSM diagnoses. If the neurobiological characteristic from D_1 follows a normal distribution with mean -4 and variance 1, i.e., $N(-4, 1)$ and that from D_2 follows a normal distribution $N(-1, 1)$, then the neurobiological characteristic follows a gaussian mixture distribution $0.3 * N(-4, 1) + 0.7 * N(-1, 1)$ over the population with dysfunction. Also suppose the neurobiological characteristic from the healthy population follows the standard normal distribution $N(0, 1)$. As can be shown in Figure 3.1, the mixture distribution is skewed and

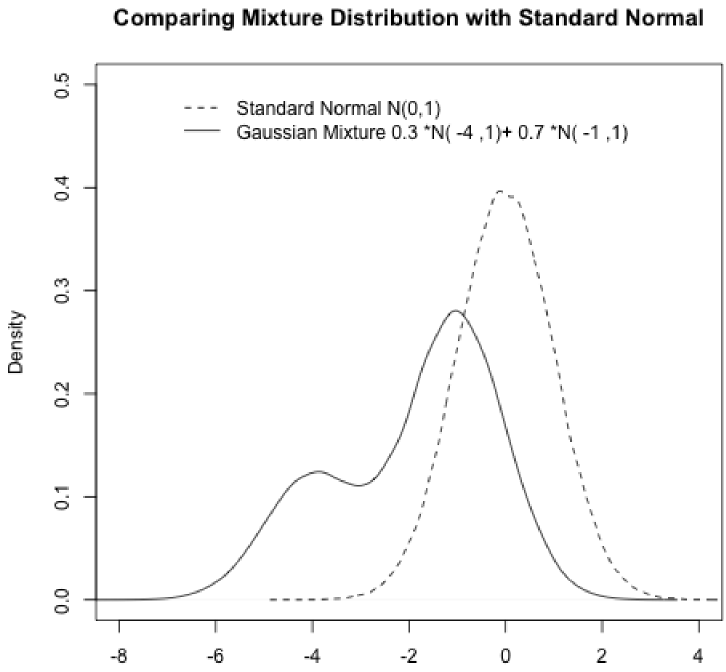


Figure 3.1: Illustration of limitation in the mean comparison

bimodal, and thus just comparing the mean of the mixture to that of the standard normal distribution does not indicate the bimodal shape of the mixture distribution, and thus does not give the full insight of how the two populations differ.

For the neurobiological characteristic identified with the proposed method in Chapter 2, there is for certain a significant mean difference in it between the population with dysfunction and the healthy population because a consistent significant mean difference exists between each DSM diagnosis and the healthy population. However, this significant mean difference could be driven by a very significant difference between the two populations in the 5% quantile but insignificant difference in the 95% quantile. The 5% quantile of a distribution is a measure in location such that there is 5% in chance a randomly selected observation from that distribution is smaller than or equal to it. In such a case of significant 5% quantile and insignificant 95% quantile, it suggests that the distance in the lower extremes of the distributions of the neurobiological characteristic between the two populations are more statistically significant than that in the upper extremes. It is possible that those subjects within the 5% quantile are from a special subpopulation among the population with dysfunction. Therefore, it is the subjects whose measured neurobiological characteristics are within the 5% quantile that deserve more investigation, rather than those with measured neurobiological characteristics above the 95% quantile. For the neurobiological characteristic that has not been identified as significant with the proposed method in Chapter 2, a significant difference in some quantile may still exist between the two populations although there is no universal significance in the mean comparison to the healthy population across all relevant DSM diagnoses. Therefore, a comparison through quantiles for these neurobiological characteristics between the two populations also makes sense. If only the means are compared, researchers would have not much of an idea at what quantiles the two populations start to differ from each other. In other words, in the case of multimodal or skewed mixture distribution, a comparison of the pth quantile in the distribution of the neurobiological characteristic rather than the means would be more helpful for researchers to understand where exactly the neurobiological characteristics differ between the two populations.

In order to compare the pth quantiles of the neurobiological characteristics between the population with dysfunction and the healthy population, brain tissues need to be collected from both populations. Again we use the available post-mortem tissue databases for these brain tissues. Two issues remain to be solved in the quantile comparison. First, as in the comparison of the means of the neurobiological characteristics, the comparison of the quantiles also needs adjustment for covariates such as age and PMI as well as the batch effect in tissue processing. The covariates and tissue processing are known to be likely related to the measurements of neurobiological characteristics. They need to be adjusted, so that a properly adjusted estimate of the difference in quantiles between the two populations can be obtained. To solve this problem, a matched subject design study can be used. The second issue in the quantile comparison comes from sampling from the mixture population. Because the population with dysfunction is a mixture of several DSM diagnoses, ideally the number of subjects from each DSM diagnosis in the sample reflects the relative proportion of that DSM diagnosis in the mixture population. However, in reality, we are limited in the number of subjects that are available for each DSM diagnosis because we are using post-mortem tissue databases. This limitation could lead to the fact that the sample sizes for the DSM diagnoses in the sample we use are disproportionate to those in the original population with dysfunction. Therefore, the sample size we use for each DSM diagnosis needs to be corrected in the comparison to reflect its true proportion. One technique that is available to compare the quantiles while taking care of the above two issues is weighted quantile regression. In this section, we will talk about the quantile comparison through this semi-parametric method.

3.3.1 Assumptions for Comparison through Quantiles

As discussed above, a matched subject design study is to be used and thus all the assumptions about subject matching in the comparison through the means in Section 3.2 still hold. The triangular design, which is a matched sample design, is still applicable here. In order to be matched, the subjects should have the same gender and have ages and PMI's as close as possible.

The measurements of the neurobiological characteristics from the subjects in the triangular design are to be used for the quantile comparison. However, unlike the comparison of the means, no assumptions need to be made about the distribution of the neurobiological characteristics in the comparison of quantiles thanks to the semi-parametric method we are using.

3.3.2 Layout of the Comparison through Quantiles

Most of the notation here follows that in the mean comparison. Again let D_i be the DSM diagnosis with relative proportion π_i among the population with dysfunction, where $\pi_1 < \pi_2 < \cdots < \pi_k$ and $\sum_{i=1}^{k} \pi_i = 1$. Let n_i be the sample size chosen in D_i and n_0 be the number of chosen subjects from the healthy population. Based on Result 3.1, $n_0 = \max(n_1, n_2, \cdots, n_k)$ for the triangular design.

In the mean comparison, because the design matrix is needed in a general form to compute the estimates, we need to know the order of the sample sizes and thus the problem is specified with ordered notation $D_{(i)}$ and $n_{(i)}$. For quantile comparison, because the design matrix is unnecessary, we can simply use D_i and n_i instead of the ordered notation.

Let $Y_{dij}(d = I(i \neq 0); i = 0, 1, \cdots, k; j = 1, 2, \cdots, n_i)$ be the measurement of the neurobiological characteristic for the subject in D_i and Match j. The index d indicates whether a subject is from the mixture population with dysfunction or the healthy population. The index i tells which DSM diagnosis the subject is from. If $i = 0$ which means the subject is from the healthy population, then $d = 0$. Otherwise the subject is from D_i and thus $d = 1$. The measurements with $d = 1$ form a sample from the mixture population with dysfunction and the measurements with $d = 0$ form a random sample from the healthy population. A model is built using the two samples to compare the pth quantile of a neurobiological characteristic in the mixture population with that in the single healthy population. Let Δ^p and $\hat{\Delta}^p$ be the true and estimated differences in the pth quantile in the distribution of the neurobiological characteristic between the population with dysfunction and the healthy population after adjustment for covariates, respectively, where p is given. Our goal is to estimate Δ^p, i.e., to obtain $\hat{\Delta}^p$. For example, when $p = 0.5$, the difference of the medians between the two populations is estimated.

3.3.3 Modeling for Comparison through Quantiles

As introduced previously, a weighted quantile regression of the data collected from the triangular design is used to estimate Δ^p. Before stating the model, we provide a brief review of quantile regression first.

3.3.3.1 Quantile Regression Quantile regression is a regression technique to estimate the conditional quantiles of a response variable. The idea of quantile regression dates back to 1951 when Brown & Mood (1951) proposed median regression. After that Hogg (1975) generalized it to percentile regression. In 1978, Koenker & Bassett (1978) first formally formulated it and called it regression quantiles. Quantile regression intends to examine the relationship between the pth quantile of the response variable and some independent variables. While p can be any value between 0 and 1, the relationship between any quantile of the response variable and the independent variables can be estimated, and thus the relationship

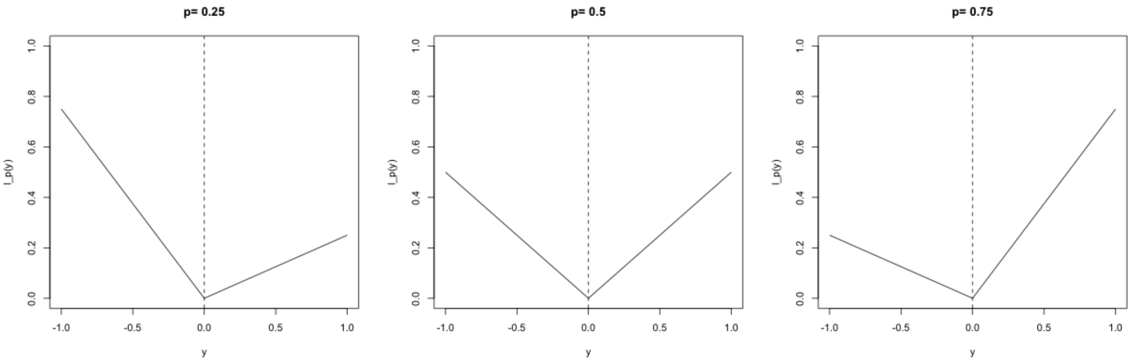


Figure 3.2: Illustration of $l_p(y)$ for $p = 0.25, 0.5, 0.75$

between the entire distribution of the response variable and the independent variables can be described functionally. To formally introduce quantile regression and how to estimate the parameters in quantile regression, we need to introduce quantiles first. Let Q_Y^p be the pth quantile of a random variable Y, then Q_Y^p is defined as:

$$Q_Y^p = F_Y^{-1}(p) = inf\{y : F(y) > p\},$$

where $F(y)$ is the cumulative distribution function of Y and $0 \leq p \leq 1$.

In contrast to the estimation of the mean of a random variable through the least square loss function, Q_Y^p can be obtained by minimizing the expectation of a particular asymmetric absolute loss function $l_p(y)$, where

$$l_p(y) = |(p - I(y < 0))y| = [p - I(y < 0)]y.$$

The expectation is taken with respect to the distribution of Y, as described in Davino et al. (2013). Figure 3.2 below illustrates $l_p(y)$ for three different p's. As can be seen from the figure, $l_p(y)$ specifies a different loss cost for $y > 0$ and $y \leq 0$ and hence the name asymmetric absolute loss function. When $y > 0$, the loss cost is p and when $y \leq 0$, the loss cost is $(p-1)$.

Using the loss function $l_p(y)$, Q_Y^p, the pth quantile of random variable Y, is just the solution that minimizes the expected loss function with respect to Y, according to Davino et al. (2013). That is to say, if the distribution function of Y is known, Q_Y^p can be obtained through the formula in (3.3.1):

$$Q_Y^p = \underset{c}{\operatorname{argmin}} E_Y[l_p(y-c)]. \tag{3.3.1}$$

If a random sample of Y is given rather than the distribution function is known, the expectation in (3.3.1) can be taken with respect to the empirical distribution of Y.

What the quantile regression does is to model the conditional quantile of Y given some independent variables $\mathbf{X}$. It assumes the pth quantile of Y is linearly dependent on $\mathbf{X}$ .Therefore, the pth conditional quantile is parametrized as a linear combination of the elements in $\mathbf{X}$ and then obtained through minimization of the above loss function. Equivalently, if $Q_{Y|\mathbf{X}}^p$ denotes the pth conditional quantile of Y on $\mathbf{X}$ and $\boldsymbol{\beta}^p$ is the coefficient for the linear combination, then

$$Q_{Y|\mathbf{X}}^p = \mathbf{X}\boldsymbol{\beta}^p,$$

and by Davino et al. (2013) $\boldsymbol{\beta}^p$ can be found by solving the following minimization in (3.3.2) with respect to the distribution of Y:

$$\hat{\boldsymbol{\beta}}^p = \underset{\boldsymbol{\beta}^p}{\operatorname{argmin}} E_Y[l_p(y-\mathbf{X}\boldsymbol{\beta}^p)]. \tag{3.3.2}$$

For a given sample $(y_1, \mathbf{X}_1), (y_2, \mathbf{X}_2), \cdots, (y_n, \mathbf{X}_n)$ where $\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_n$ are fixed values, in order to estimate the pth conditional quantile of Y on $\mathbf{X}$, the expectation in (3.3.2) would be taken with respect to the empirical distribution of Y and the objective function we minimize in (3.3.2) becomes

$$\begin{aligned}\hat{\boldsymbol{\beta}}^p &= \underset{\boldsymbol{\beta}^p}{\operatorname{argmin}} \sum_{i=1}^{n} [l_p(y_i - \mathbf{X}_i\boldsymbol{\beta}^p)] \\ &= \underset{\boldsymbol{\beta}^p}{\operatorname{argmin}} [\sum_{i \in \{i: y_i \geq \mathbf{X}_i\boldsymbol{\beta}^p\}} p(y_i - \mathbf{X}_i\boldsymbol{\beta}^p) + \sum_{i \in \{i: y_i < \mathbf{X}_i\boldsymbol{\beta}^p\}} (1-p)(\mathbf{X}_i\boldsymbol{\beta}^p - y_i)].\end{aligned}$$

If each sample $(y_i, \mathbf{X}_i)$ has a weight ω_i, then by Koenker (2005) the parameter $\boldsymbol{\beta}^p$ can be found through a weighted quantile regression by solving the following objective function in (3.3.3).

$$\begin{aligned}
\hat{\boldsymbol{\beta}}^p &= \underset{\boldsymbol{\beta}^p}{\operatorname{argmin}} \sum_{i=1}^{n} [\omega_i l_p(y_i - \mathbf{X}_i \boldsymbol{\beta}^p)] \\
&= \underset{\boldsymbol{\beta}^p}{\operatorname{argmin}} [\sum_{i \in \{i : y_i \geq \mathbf{X}_i \boldsymbol{\beta}^p\}} \omega_i p(y_i - \mathbf{X}_i \boldsymbol{\beta}^p) + \sum_{i \in \{i : y_i < \mathbf{X}_i \boldsymbol{\beta}^p\}} \omega_i (1-p)(\mathbf{X}_i \boldsymbol{\beta}^p - y_i)]. \qquad (3.3.3)
\end{aligned}$$

As can be seen from the above discussion, quantile regression is very different from the usual least square regression because the two regression techniques use different loss functions and estimate different things. The least square regression estimates how the mean of Y changes according to $\mathbf{X}$ and the quantile regression estimates how the pth quantile of Y changes according to $\mathbf{X}$.

The parameters in quantile regression are estimated through linear programming and thus there are no closed form expressions for the estimates. The parameters tell us for one unit increase in an independent variable while holding other independent variables constant, how much the expected pth quantile of the response variable changes.

3.3.3.2 Model for Quantile Comparison

Recall our goal is to model the data from the triangular design to estimate the difference in the pth quantile between the mixture population with dysfunction and the healthy population adjusting for covariates. The covariate adjustment issue can be addressed by adding a matching effect into the model. Now the question is what effect to use to distinguish subjects from the two different populations. In the mean comparison, the DSM diagnosis effect is used because the mean of the neurobiological characteristic in each DSM diagnosis can be estimated and then linearly combined to estimate the mean of the neurobiological characteristic in the population with dysfunction. However, the linear combination of the pth quantile in each DSM diagnosis is not the pth quantile of the population with dysfunction. Therefore specifying the DSM diagnosis in the quantile regression model does not lead to an estimate of the pth quantile of the neurobiological characteristic in the population with dysfunction. As a result, in the quantile regression model a population effect is used to describe whether a subject is from the healthy population or the population with dysfunction. For example, if we are interested in the quantile of a neurobiological characteristic involved in psychosis, then the population effect would have two levels, either psychotic or healthy.

If we use β_d^p to denote the population effect in the pth quantile of the neurobiological characteristic about whether a person has dysfunction in the construct or not ($d = 1$ if with dysfunction and $d = 0$ if healthy), then the quantile regression model can be written as:

$$Q_{dj}^p = \beta_d^p + M_j^p, \qquad d = 1, 0; j = 1, 2, \cdots, n_0; 0 < p < 1, \tag{3.3.4}$$

where Q_{dj}^p is pth quantile of the neurobiological characteristic for subjects in Match j and from population d. Here the index j goes from 1 to n_0 in both populations because both populations have n_0 matches. However, the number of subjects from population $d = 1$ is larger than that from population $d = 0$ because the population with $d = 1$ is a mixture population and have more than one DSM diagnosis. Together we have $n_0 + 2$ parameters to estimate, which are $(\beta_0^p, \beta_1^p, M_1^p, M_2^p, \cdots, M_{n_0}^p)$.

Let ϵ_{dij} be the error term for observation Y_{dij}. We prove in Appendix 3.A.3 that the above model in terms of the conditional quantile is equivalent to

$$Y_{dij} = \beta_d^p + M_j^p + \epsilon_{dij}, \qquad d = 1, 0; i = 0, 1, \cdots, k; j = 1, \cdots, n_i; 0 < p < 1, \tag{3.3.5}$$

provided that the pth quantile of F_0, the distribution of the error term in the healthy population and that of F_1, the distribution of the error term in the population with dysfunction are both 0. If the pth quantile of F_0 or that of F_1 is not 0, then the model in (3.3.4) does not provide an unbiased estimate of the pth conditional quantile of the neurobiological characteristic with given match and population.

The model in (3.3.4) estimates the pth quantile of the distribution of the neurobiological characteristic in each of the two populations while also adjusting for the matching effect. If $\hat{\beta}_d^p$ is the estimate of β_d^p, then $\hat{\Delta}^p = \hat{\beta}_1^p - \hat{\beta}_0^p$ is the estimate of the difference in the pth quantile of the neurobiological characteristic between the population with dysfunction and the healthy population. Although no DSM diagnosis is involved in the model statement, the above model allows the differences in the pth quantile of the distributions of the neurobiological characteristic between each DSM diagnosis and the healthy population to be different. The proof is in Appendix 3.A.4.

One thing that worth mentioning is that in the model the match effect M_j^p takes care of the variation introduced in the tissue processing as well as in the matching covariates such as age and PMI. If later on researchers are interested in the effect of some other covariates that are not used in the matching, such as PH, these covariates can be put into the model. However, we choose not to do it here in the dissertation because we want to keep the model for the quantile comparison as close as possible to that for the mean comparison.

3.3.3.3 Weighting the Subjects We have established a model to estimate the pth quantile of the neurobiological characteristic in the healthy population as well as the mixture population with dysfunction of interest. However, the problem is that the observed data are not a random sample from the population with dysfunction. Rather the observed data for each DSM diagnosis is a random sample from that diagnosis only. As mentioned earlier, the numbers of subjects in each DSM diagnosis collected in the sample do not necessarily have the same relative proportion as appeared in the population with dysfunction, i.e., $n_1 : n_2 : \cdots : n_k$ is not necessarily equal to $\pi_1 : \pi_2 : \cdots : \pi_k$. Thus, if we ignore that the data are not a random sample from the mixture population and treat it as such, because the n_i's can be quite different from the π_i's, the assumption of a random sample from the mixture population could be quite invalid. If no adjustment is made for the sample size in the data, what we input into the model does not reflect the true composition of the mixture population and thus we are unable to approach the true distribution of the neurobiological characteristic in the mixture population.

To obtain a valid estimate of Δ^p, the difference in the pth quantile between the two populations, we develop a heuristic weighted approach to find an estimator and study the properties of this estimator through simulation.

The weighted approach is to assign each observation y_{dij} a proper weight ω_{di}. For the population with dysfunction, i.e., $d = 1$, all the observations in D_i have the same weight ω_{1i}

in the modeling, where

$$\omega_{1i} = \pi_i * n_i^{-1} * \sum_{i=1}^{k} n_i, \quad i = 1, 2, \cdots, k.$$

Through weighting the subjects, it looks as if D_i has sample size $\omega_{1i} * n_i$. Instead of occurring for only once, the observation y_{1ij} would seem to occur for ω_{1i} times in the data. For the subjects from the healthy population, because they all come from a single population, no adjustments need to be made. Therefore, they all get weight 1, i.e., $\omega_{00} = 1$. The weighting of the subjects can be taken care of through weighted quantile regression, as described in Koenker (2005).

If we plug the model in (3.3.4) into the objective function in (3.3.3), the parameters in the model can be solved through minimization of the following specific objective function, which is

$$\begin{aligned} &(\hat{\beta}_0^p, \hat{\beta}_1^p, \hat{M}_1^p, \hat{M}_2^p, \cdots, \hat{M}_{n_0}^p) \\ = \underset{(\beta_0^p, \beta_1^p, M_1^p, M_2^p, \cdots, M_{n_0}^p)}{\operatorname{argmin}} &\Big[\sum_{(d,i,j) \in \{(d,i,j): y_{dij} \geq \beta_d^p + M_j^p\}} \omega_{di} p (y_{dij} - \beta_d^p - M_j^p) \\ &+ \sum_{(d,i,j) \in \{(d,i,j): y_{dij} < \beta_d^p + M_j^p\}} \omega_{di} (1-p)(\beta_d^p + M_j^p - y_{dij}) \Big]. \end{aligned}$$

3.3.4 Simulation Study for Comparison through Quantiles

A simulation study is run to illustrate the quantile regression and contrast the weighted and unweighted estimates of the quantiles for the mixture population.

3.3.4.1 Simulation Method In each simulation, two DSM diagnoses are used ($k = 2$), each with relative proportion π_i and sample size n_i. According to our assumption, $\pi_1 < \pi_2$ and $\pi_1 + \pi_2 = 1$. According to the assumptions made in Section 3.2, the sample size for the healthy population would be $\max(n_1, n_2)$. Suppose n_1, n_2 and $\max(n_1, n_2)$ observations are sampled from normal distributions $N(\mu_1, 1)$, $N(\mu_2, 1)$ and $N(0, 1)$, respectively. The observations from $N(\mu_1, 1)$ and $N(\mu_2, 1)$ collectively can be considered as a sample from the mixture population with dysfunction. This process is the same as what we do in the triangular design using the post-mortem tissue databases. Because we assume the identified neurobiological characteristics are either consistently up-regulated or consistently down-regulated across the multiple DSM diagnoses, μ_1 and μ_2 have the same sign. In the simulation we assume they are both negative.

A quantile regression model is built with the $n_1+n_2+\max(n_1, n_2)$ observations to estimate the pth quantiles of the distributions for the mixture population and the healthy population. Note that because we want to estimate the pth quantile with adjustments for the match effect, it means that the two estimates are based on the same match effect. Therefore, to simplify the simulation, we could just assume all the observations have the same covariates and eliminate the match effect from the model. By doing this, no match effect needs to be simulated. Therefore, only population effect exists in the model applied to the simulation data. In each simulation, the pth quantile in the distribution for both the mixture population and the healthy population are calculated, in both the weighted and unweighted method. These estimates are compared to the theoretical pth quantiles for both populations. Here the theoretical value of the quantiles in the mixture distribution $\pi_1 N(\mu_1, 1) + \pi_2 N(\mu_2, 1)$ are obtained analytically using a greedy search function.To further simplify the simulation, we fix $\mu_2 = -1$ and $n_2 = 100$ because it is the distance between μ_1 and μ_2 that determines the shape of the mixture distribution for fixed π_1 and π_2. Also it is the ratio of n_1 to n_2 that determines how far away the sample is from a random sample of the mixture distribution. In the simulation, n_1 and n_2 are intentionally kept small because it is known that the number of subjects for each DSM diagnosis from the post-mortem tissue databases is not large.

By changing the values of μ_1, π_1, n_1 and p, we could see the impact of weighting in the comparison of quantiles in the different parameter configuration. For each parameter configuration, 500 simulations are run. The parameters used in the simulation are listed in Table 3.5. The mean and standard deviation of the estimates over the 500 simulations under each parameter configuration are presented in Section 3.3.4.2.

3.3.4.2 Simulation Results The simulation results for the different p's are summarized in Tables 3.6 - 3.10.

As stated previously, these tables summarize the mean and standard deviation (in paren-

Table 3.5: Simulation Parameter List for Quantile Regression

parameter	value
μ_1	$-4, -2, -0.25$
μ_2	-1
π_1	$0.1, 0.2, 0.3$
n_1	$25, 100, 500$
n_2	100
p	$0.05, 0.25, 0.5, 0.75, 0.95$

thesis) of the quantile estimates over the 500 simulations under different simulation parameter configurations. For the mixture population, the theoretical quantile, unweighted estimate and weighted estimate are presented. For the healthy population, because subjects from the healthy population all get weight 1, the unweighted and weighted estimates are the same, and thus only a single column is kept here. For example, in Table 3.6, the 5% quantiles are estimated for both the mixture population and the healthy population. If in the mixture population DSM diagnosis D_1 accounts for 10%, i.e., $\pi_1 = 0.1$, and observations from D_1 follow a normal distribution $N(-4, 1)$, i.e., $\mu_1 = -4$, then the theoretical 5% quantile of the distribution for the mixture population is -4.028. If 25 subjects are sampled from D_1, i.e., $n_1 = 25$, and the unweighted quantile regression model is used, the average estimate of the 5% quantile of the distribution for the mixture population is -4.640 with a standard deviation of 0.281. If the weighted quantile regression model is used, the average estimate of the 5% quantile of the distribution for the mixture population is -4.020 with a standard deviation of 0.257. For the healthy population, because we assume all the subjects from it follow the standard normal distribution $N(0, 1)$, the theoretical 5% quantile is -1.645. The unweighted and weighted 5% quantile estimates are both -1.684 with a standard deviation of 0.223.

As can be seen from the tables, the weighted estimate of each quantile of the distribution for the mixture population is less biased than the unweighted estimate because on average

Table 3.6: Simulation Results for Lower 5% Quantile ($\mu_2 = -1, n_2 = 100, p = 0.05$)

π_1	μ_1	n_1	theoretical	Mixture Population unweighted	 weighted	Healthy Population theoretical=-1.645
0.1	-4	25	-4.028	-4.640 (0.281)	-4.020 (0.257)	-1.684 (0.223)
		100		-5.245 (0.167)	-3.997 (0.136)	-1.704 (0.209)
		500		-5.544 (0.083)	-4.014 (0.077)	-1.649 (0.096)
	-2	25	-2.836	-2.971 (0.204)	-2.834 (0.200)	-1.690 (0.207)
		100		-3.308 (0.152)	-2.835 (0.170)	-1.688 (0.201)
		500		-3.555 (0.090)	-2.832 (0.173)	-1.656 (0.093)
	-0.25	25	-2.603	-2.530 (0.184)	-2.592 (0.198)	-1.689 (0.214)
		100		-2.359 (0.160)	-2.593 (0.205)	-1.675 (0.213)
		500		-2.092 (0.088)	-2.595 (0.198)	-1.651 (0.092)
0.2	-4	25	-4.676	-4.638 (0.272)	-4.638 (0.272)	-1.680 (0.218)
		100		-5.254 (0.170)	-4.658 (0.129)	-1.680 (0.209)
		500		-5.542 (0.087)	-4.673 (0.062)	-1.656 (0.089)
	-2	25	-2.999	-2.984 (0.202)	-2.984 (0.202)	-1.681 (0.217)
		100		-3.311 (0.161)	-2.983 (0.149)	-1.695 (0.221)
		500		-3.546 (0.089)	-2.991 (0.133)	-1.648 (0.096)
	-0.25	25	-2.556	-2.526 (0.193)	-2.526 (0.193)	-1.694 (0.206)
		100		-2.356 (0.146)	-2.541 (0.186)	-1.691 (0.212)
		500		-2.087 (0.093)	-2.550 (0.170)	-1.656 (0.097)
0.3	-4	25	-4.968	-4.639 (0.259)	-4.911 (0.286)	-1.671 (0.220)
		100		-5.251 (0.168)	-4.974 (0.145)	-1.704 (0.225)
		500		-5.548 (0.087)	-4.970 (0.066)	-1.649 (0.097)
	-2	25	-3.134	-2.975 (0.208)	-3.122 (0.247)	-1.714 (0.225)
		100		-3.303 (0.162)	-3.119 (0.144)	-1.687 (0.209)
		500		-3.556 (0.085)	-3.140 (0.109)	-1.654 (0.099)
	-0.25	25	-2.504	-2.534 (0.192)	-2.488 (0.187)	-1.677 (0.203)
		100		-2.355 (0.157)	-2.497 (0.184)	-1.682 (0.224)
		500		-2.085 (0.088)	-2.503 (0.196)	-1.654 (0.095)

Table 3.7: Simulation Results for Lower 25% Quantile ($\mu_2 = -1, n_2 = 100, p = 0.25$)

				Mixture Population		Healthy Population
π_1	μ_1	n_1	theoretical	unweighted	weighted	theoretical=-0.675
0.1	-4	25	-1.958	-2.407 (0.166)	-1.946 (0.142)	-0.683 (0.142)
		100		-3.978 (0.129)	-1.965 (0.147)	-0.690 (0.142)
		500		-4.520 (0.060)	-1.964 (0.137)	-0.675 (0.064)
	-2	25	-1.795	-1.900 (0.130)	-1.785 (0.129)	-0.683 (0.141)
		100		-2.256 (0.102)	-1.791 (0.117)	-0.684 (0.131)
		500		-2.554 (0.058)	-1.804 (0.133)	-0.677 (0.062)
	-0.25	25	-1.618	-1.555 (0.116)	-1.622 (0.120)	-0.690 (0.135)
		100		-1.344 (0.098)	-1.622 (0.128)	-0.677 (0.131)
		500		-1.067 (0.056)	-1.622 (0.131)	-0.675 (0.060)
0.2	-4	25	-2.426	-2.400 (0.154)	-2.400 (0.154)	-0.691 (0.138)
		100		-3.988 (0.125)	-2.412 (0.152)	-0.691 (0.138)
		500		-4.524 (0.057)	-2.418 (0.152)	-0.679 (0.059)
	-2	25	-1.918	-1.907 (0.126)	-1.907 (0.126)	-0.679 (0.137)
		100		-2.244 (0.103)	-1.912 (0.115)	-0.694 (0.131)
		500		-2.556 (0.059)	-1.915 (0.107)	-0.678 (0.061)
	-0.25	25	-1.557	-1.547 (0.120)	-1.547 (0.120)	-0.688 (0.141)
		100		-1.343 (0.097)	-1.559 (0.121)	-0.685 (0.141)
		500		-1.068 (0.059)	-1.564 (0.121)	-0.681 (0.062)
0.3	-4	25	-3.166	-2.405 (0.174)	-3.158 (0.233)	-0.675 (0.137)
		100		-3.987 (0.123)	-3.162 (0.133)	-0.695 (0.141)
		500		-4.520 (0.061)	-3.159 (0.090)	-0.678 (0.064)
	-2	25	-2.039	-1.906 (0.126)	-2.036 (0.136)	-0.698 (0.138)
		100		-2.256 (0.101)	-2.035 (0.105)	-0.690 (0.136)
		500		-2.557 (0.054)	-2.041 (0.088)	-0.680 (0.063)
	-0.25	25	-1.492	-1.550 (0.124)	-1.489 (0.124)	-0.686 (0.133)
		100		-1.329 (0.100)	-1.479 (0.113)	-0.689 (0.141)
		500		-1.067 (0.053)	-1.484 (0.107)	-0.680 (0.059)

Table 3.8: Simulation Results for Lower 50% Quantile ($\mu_2 = -1, n_2 = 100, p = 0.50$)

π_1	μ_1	n_1	theoretical	Mixture Population		Healthy Population
				unweighted	weighted	theoretical=0
0.1	-4	25	-1.139	-1.316 (0.123)	-1.141 (0.122)	-0.010 (0.127)
		100		-2.487 (0.132)	-1.145 (0.120)	-0.010 (0.130)
		500		-3.747 (0.059)	-1.135 (0.122)	-0.006 (0.056)
	-2	25	-1.089	-1.178 (0.118)	-1.083 (0.119)	-0.014 (0.122)
		100		-1.487 (0.083)	-1.078 (0.114)	-0.007 (0.126)
		500		-1.843 (0.052)	-1.091 (0.117)	0.000 (0.055)
	-0.25	25	-0.930	-0.863 (0.110)	-0.931 (0.117)	-0.020 (0.127)
		100		-0.625 (0.090)	-0.933 (0.116)	-0.008 (0.125)
		500		-0.362 (0.052)	-0.927 (0.115)	0.002 (0.058)
0.2	-4	25	-1.316	-1.309 (0.128)	-1.309 (0.128)	-0.016 (0.119)
		100		-2.475 (0.132)	-1.317 (0.126)	-0.009 (0.119)
		500		-3.750 (0.054)	-1.311 (0.126)	-0.003 (0.056)
	-2	25	-1.184	-1.187 (0.121)	-1.187 (0.121)	-0.004 (0.125)
		100		-1.493 (0.097)	-1.186 (0.117)	-0.012 (0.120)
		500		-1.847 (0.053)	-1.186 (0.110)	-0.001 (0.056)
	-0.25	25	-0.857	-0.861 (0.112)	-0.861 (0.112)	-0.012 (0.127)
		100		-0.620 (0.090)	-0.854 (0.107)	-0.002 (0.132)
		500		-0.366 (0.053)	-0.859 (0.107)	-0.006 (0.054)
0.3	-4	25	-1.557	-1.325 (0.124)	-1.565 (0.132)	-0.004 (0.121)
		100		-2.487 (0.141)	-1.546 (0.128)	-0.009 (0.135)
		500		-3.744 (0.055)	-1.556 (0.120)	-0.001 (0.058)
	-2	25	-1.286	-1.176 (0.120)	-1.275 (0.119)	-0.020 (0.127)
		100		-1.497 (0.097)	-1.284 (0.103)	-0.011 (0.126)
		500		-1.848 (0.050)	-1.285 (0.095)	-0.004 (0.055)
	-0.25	25	-0.781	-0.857 (0.111)	-0.781 (0.117)	0.000 (0.124)
		100		-0.616 (0.090)	-0.775 (0.099)	-0.012 (0.126)
		500		-0.366 (0.052)	-0.775 (0.094)	-0.003 (0.055)

Table 3.9: Simulation Results for Lower 75% Quantile ($\mu_2 = -1, n_2 = 100, p = 0.75$)

π_1	μ_1	n_1		Mixture Population		Healthy Population
			theoretical	unweighted	weighted	theoretical=0.675
0.1	-4	25	-0.411	-0.513 (0.127)	-0.398 (0.133)	0.662 (0.129)
		100		-0.987 (0.119)	-0.401 (0.129)	0.655 (0.132)
		500		-2.759 (0.071)	-0.399 (0.127)	0.666 (0.060)
	-2	25	-0.393	-0.470 (0.127)	-0.387 (0.132)	0.650 (0.129)
		100		-0.731 (0.102)	-0.393 (0.127)	0.658 (0.142)
		500		-1.119 (0.058)	-0.402 (0.119)	0.673 (0.059)
	-0.25	25	-0.237	-0.153 (0.118)	-0.233 (0.119)	0.654 (0.137)
		100		0.101 (0.101)	-0.244 (0.124)	0.664 (0.138)
		500		0.332 (0.058)	-0.234 (0.117)	0.672 (0.061)
0.2	-4	25	-0.511	-0.523 (0.130)	-0.523 (0.130)	0.658 (0.132)
		100		-0.988 (0.127)	-0.522 (0.133)	0.663 (0.135)
		500		-2.763 (0.068)	-0.518 (0.128)	0.673 (0.062)
	-2	25	-0.467	-0.477 (0.124)	-0.477 (0.124)	0.663 (0.127)
		100		-0.736 (0.105)	-0.466 (0.125)	0.655 (0.130)
		500		-1.122 (0.060)	-0.468 (0.120)	0.675 (0.061)
	-0.25	25	-0.149	-0.153 (0.125)	-0.153 (0.125)	0.651 (0.129)
		100		0.099 (0.095)	-0.155 (0.109)	0.661 (0.138)
		500		0.332 (0.057)	-0.151 (0.112)	0.668 (0.062)
0.3	-4	25	-0.633	-0.521 (0.132)	-0.633 (0.132)	0.653 (0.123)
		100		-0.986 (0.122)	-0.635 (0.124)	0.662 (0.138)
		500		-2.756 (0.071)	-0.625 (0.128)	0.670 (0.060)
	-2	25	-0.548	-0.470 (0.130)	-0.542 (0.127)	0.653 (0.137)
		100		-0.728 (0.105)	-0.543 (0.118)	0.660 (0.137)
		500		-1.118 (0.057)	-0.555 (0.113)	0.669 (0.062)
	-0.25	25	-0.064	-0.153 (0.125)	-0.055 (0.138)	0.668 (0.135)
		100		0.108 (0.097)	-0.065 (0.101)	0.655 (0.130)
		500		0.328 (0.056)	-0.063 (0.087)	0.672 (0.062)

Table 3.10: Simulation Results for Lower 95% Quantile ($\mu_2 = -1, n_2 = 100, p = 0.95$)

π_1	μ_1	n_1		Mixture Population		Healthy Population
			theoretical	unweighted	weighted	theoretical=1.645
0.1	-4	25	0.593	0.504 (0.187)	0.590 (0.196)	1.613 (0.203)
		100		0.307 (0.172)	0.591 (0.207)	1.587 (0.209)
		500		-0.453 (0.125)	0.596 (0.210)	1.630 (0.094)
	-2	25	0.598	0.526 (0.184)	0.601 (0.190)	1.580 (0.197)
		100		0.361 (0.167)	0.587 (0.205)	1.602 (0.219)
		500		-0.050 (0.091)	0.593 (0.198)	1.633 (0.093)
	-0.25	25	0.769	0.867 (0.191)	0.776 (0.187)	1.598 (0.199)
		100		1.152 (0.145)	0.757 (0.179)	1.597 (0.200)
		500		1.330 (0.087)	0.767 (0.166)	1.639 (0.092)
0.2	-4	25	0.534	0.498 (0.193)	0.498 (0.193)	1.581 (0.195)
		100		0.315 (0.168)	0.510 (0.186)	1.603 (0.208)
		500		-0.459 (0.129)	0.514 (0.187)	1.631 (0.096)
	-2	25	0.545	0.505 (0.189)	0.505 (0.189)	1.587 (0.192)
		100		0.351 (0.157)	0.533 (0.188)	1.585 (0.206)
		500		-0.046 (0.089)	0.526 (0.178)	1.640 (0.096)
	-0.25	25	0.879	0.848 (0.198)	0.848 (0.198)	1.592 (0.199)
		100		1.145 (0.152)	0.879 (0.162)	1.600 (0.202)
		500		1.332 (0.086)	0.876 (0.150)	1.638 (0.097)
0.3	-4	25	0.465	0.510 (0.196)	0.434 (0.186)	1.590 (0.213)
		100		0.303 (0.166)	0.432 (0.182)	1.585 (0.204)
		500		-0.455 (0.131)	0.438 (0.183)	1.631 (0.091)
	-2	25	0.486	0.509 (0.202)	0.450 (0.194)	1.587 (0.205)
		100		0.363 (0.157)	0.476 (0.175)	1.603 (0.206)
		500		-0.053 (0.093)	0.485 (0.177)	1.627 (0.096)
	-0.25	25	0.975	0.866 (0.195)	0.992 (0.231)	1.578 (0.200)
		100		1.148 (0.145)	0.970 (0.138)	1.592 (0.204)
		500		1.323 (0.090)	0.977 (0.118)	1.630 (0.091)

the weighted estimate is closer to the theoretical quantile than the unweighted one is. For the unweighted estimate, the bias changes according to the parameter configuration. For fixed μ_1, as the ratio of n_1 to n_2 becomes farther away from that of π_1 to π_2, the unweighted estimate of the quantile of the distribution for the mixture population becomes more biased. For example, in Table 3.6, when $\pi_1 = 0.1$ and $\mu_1 = -4$, the ratio of π_1 to π_2 is 1/9. When n_1 goes from 25 to 500, the ratio of n_1 to n_2 goes farther away from 1/9, and thus the expected value of the unweighted estimate of the 5% quantile of the distribution for the mixture population goes from -4.640 to -5.544, which is more and more biased. Furthermore, for fixed π_1, n_1 and μ_1, as p increases, the bias of the unweighted estimate of the quantile for the mixture population increases first and then decreases. This means the bias is smaller for the more extreme quantiles. For example, when $\pi_1 = 0.1, n_1 = 500$ and $\mu_1 = -4$, the average unweighted estimate for the 5%, 50% and 95% quantile is -5.544, -3.747 and -0.453, respectively. The bias of the unweighted estimate for the 5%, 50% and 95% quantile would be -5.544-(-4.028)=-1.516, -3.747-(-1.139)=-2.608 and -0.453-0.593=-1.046, respectively. Also when π_i, n_i and p are fixed, as μ_1 gets farther away from μ_2, the bias of the unweighted estimate becomes larger. For example, in Table 3.6, when $\pi_1 = 0.1$ and $n_1 = 100$, as μ_1 goes from -4 to -2 and then to -0.25, the distance of μ_2 to μ_1 first decreases and then increases. And the bias of the unweighted estimate goes from -5.245-(-4.028)=-1.217 to -3.308-(-2.836)=-0.472 and then to -2.359-(-2.603)=0.244.

Bias also exists in the weighted estimates of the quantiles for the mixture population as well as in the estimates for the healthy population. This is because the assumption for the model error term is not satisfied. As discussed in Section 3.3.3, for the quantile regression model to be stated as in (3.3.4), the pth quantile for the distribution of the error term has to be 0. However, in the simulation, the error terms for the observations from the mixture population may not have pth quantile as 0.

One thing that is worthwhile pointing out is that when π_1/π_2 is equal to n_1/n_2, the weights for subjects in D_1 and D_2 are both 1. Therefore, the weighted and unweighted estimates of the quantiles for the mixture population are the same in this case. For example, in Table 3.6, when $\pi_1 = 0.2$ and $n_1 = 25$, $\pi_1/\pi_2 = 0.2/0.8 = 0.25$ and $n_1/n_2 = 25/100 = 0.25$, so the weighted and unweighted estimates are the same no matter what μ_1 is.

3.4 SUMMARY OF COMPARISON

In Chapter 3, we propose the triangular design to implement two comparisons in the neurobiological characteristics between the population with dysfunction and the healthy population with adjustments for covariates. The design is based on some assumptions about matching subjects in order to improve efficiency in tissue processing. In order for us to write down the design matrix in general, the triangular design is laid out with the order statistics of the sample sizes. The two comparisons provide different insights in terms of the distribution of the neurobiological characteristics in the two populations and can be used in different situations. When the distributions of the neurobiological characteristics in the DSM diagnoses are close to each other and thus make its distribution in the mixture population unimodal and symmetric, the mean comparison is applicable to inform researchers about the difference between the two populations. When the distribution of the neurobiological characteristic in the mixture population is multimodal or skewed, the quantile comparison is suitable.

For both comparisons, adjustment for the disproportionate sample sizes in the sample need to be made. However, they are done with different methods. In the comparison through the means, an ANOVA model is employed with a DSM diagnosis effect and a match effect in the model. The estimate of the mean neurobiological characteristic in the mixture population is written as the linear combination of the estimates in each DSM diagnosis, with the relative proportions as the combination coefficients. In the comparison through the quantiles, the quantile regression model is used with a population effect and a match effect. The disproportionate sample sizes are corrected heuristically by applying a proper weight for each DSM diagnosis.

In the comparisons, both models have only two effects. The goal here is to propose a simple solution or framework to think about using multiple post-mortem tissue databases in general. In later research, if investigators would want to use more effects in the model, these effects can be added accordingly.

3.A APPENDIX

3.A.1 Proof of estimability of $\beta_{(i)} - \beta_0$

Result 3.2. *$\beta_{(i)} - \beta_0$ is estimable under the ANOVA model in* (3.2.3) *for the triangular design in Table 3.4.*

Proof. Following the discussion in Section 3.2.4.2, if we can find a vector $\tau_{(i)}$ with dimension $k + 1 + n_{(k)}$ such that $l_{(i)} = \mathcal{X}'\mathcal{X}\tau_{(i)}$ holds, $\beta_{(i)} - \beta_0$ is estimable. Based on the model in (3.2.3), the matrix $\mathcal{X}'\mathcal{X}$ is:

$$\mathcal{X}'\mathcal{X} = \begin{bmatrix}
n_{(1)} & 0 & \cdots & 0 & 0 & 1 & 1 & \cdots & 1 & 0 & 0 & \cdots & 0 & \cdots & 0 \\
0 & n_{(2)} & \cdots & 0 & 0 & 1 & 1 & \cdots & 1 & 1 & 1 & \cdots & 1 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \cdots & \vdots & \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\
0 & 0 & \cdots & n_{(k)} & 0 & 1 & 1 & \cdots & 1 & 1 & 1 & \cdots & 1 & \cdots & 1 \\
0 & 0 & \cdots & 0 & n_{(k)} & 1 & 1 & \cdots & 1 & 1 & 1 & \cdots & 1 & \cdots & 1 \\
1 & 1 & \cdots & 1 & 1 & k+1 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & \cdots & 0 \\
1 & 1 & \cdots & 1 & 1 & 0 & k+1 & \cdots & 0 & 0 & 0 & \cdots & 0 & \cdots & 0 \\
\vdots & \vdots & \cdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\
1 & 1 & \cdots & 1 & 1 & 0 & 0 & \cdots & k+1 & 0 & 0 & \cdots & 0 & \cdots & 0 \\
0 & 1 & \cdots & 1 & 1 & 0 & 0 & \cdots & 0 & k & 0 & \cdots & 0 & \cdots & 0 \\
0 & 1 & \cdots & 1 & 1 & 0 & 0 & \cdots & 0 & 0 & k & \cdots & 0 & \cdots & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \cdots & \vdots & \vdots & \vdots & \ddots & \vdots & \cdots & \vdots \\
0 & 1 & \cdots & 1 & 1 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & k & \cdots & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \cdots & \vdots & \vdots & \vdots & \cdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & 1 & 1 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & \cdots & 2
\end{bmatrix}.$$

Suppose $\tau_{(i)}$ is given in general as

$$\tau_{(i)} = \begin{pmatrix} a_{i,1} \\ a_{i,2} \\ \vdots \\ a_{i,k} \\ a_{i,0} \\ b_{i,1} \\ b_{i,2} \\ \vdots \\ b_{i,n_{(1)}} \\ b_{i,n_{(1)}+1} \\ b_{i,n_{(1)}+2} \\ \vdots \\ b_{i,n_{(2)}} \\ \vdots \\ b_{i,n_{(k)}} \end{pmatrix},$$

then

$$\mathcal{X}'\mathcal{X}\tau_{(i)} = \begin{bmatrix} n_{(1)}a_{i,1} + \sum_{j=1}^{n_{(1)}} b_{i,j} \\ n_{(2)}a_{i,2} + \sum_{j=1}^{n_{(2)}} b_{i,j} \\ \vdots \\ n_{(k)}a_{i,k} + \sum_{j=1}^{n_{(k)}} b_{i,j} \\ n_{(k)}a_{i,0} + \sum_{j=1}^{n_{(k)}} b_{i,j} \\ \sum_{j=1}^{k} a_{i,j} + a_{i,0} + (k+1)b_{i,1} \\ \sum_{j=1}^{k} a_{i,j} + a_{i,0} + (k+1)b_{i,2} \\ \vdots \\ \sum_{j=1}^{k} a_{i,j} + a_{i,0} + (k+1)b_{i,n_{(1)}} \\ \sum_{j=2}^{k} a_{i,j} + a_{i,0} + (k)b_{i,n_{(1)}+1} \\ \sum_{j=2}^{k} a_{i,j} + a_{i,0} + (k)b_{i,n_{(1)}+2} \\ \vdots \\ \sum_{j=2}^{k} a_{i,j} + a_{i,0} + (k)b_{i,n_{(2)}} \\ \vdots \\ a_{i,k} + a_{i,0} + (2)b_{i,n_{(k)}} \end{bmatrix} = l_{(i)} = \begin{bmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ -1 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ \vdots \\ 0 \end{bmatrix} \begin{matrix} \\ \\ \leftarrow i\text{th element} \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \end{matrix}. \qquad (3.A.1)$$

The remaining work is to solve for $\tau_{(i)}$ in the equation set of (3.A.1).

1. if $i = k, \beta_{(i)} - \beta_0 = \beta_{(k)} - \beta_0$, it's trivial that

$$a_{k,1} = a_{k,2} = \cdots = a_{k,k-1} = 0,\ a_{k,k} = \frac{1}{n_{(k)}},\ a_{k,0} = -\frac{1}{n_{(k)}},\ b_{k,1} = b_{k,2} = \cdots = b_{k,n_{(k)}} = 0.$$

2. if $1 \leq i \leq k-1$, we have the following equations to solve:

$$\left\{\begin{array}{rlcc}
n_{(1)}a_{i,1}+ & \sum_{j=1}^{n_{(1)}} b_{i,j} & = & 0 \\
 & \vdots & & \\
n_{(i-1)}a_{i,i-1}+ & \sum_{j=1}^{n_{(i-1)}} b_{i,j} & = & 0 \\
n_{(i)}a_{i,i}+ & \sum_{j=1}^{n_{(i)}} b_{i,j} & = & 1 \\
n_{(i+1)}a_{i,i+1}+ & \sum_{j=1}^{n_{(i+1)}} b_{i,j} & = & 0 \\
 & \vdots & & \\
n_{(k)}a_{i,k}+ & \sum_{j=1}^{n_{(k)}} b_{i,j} & = & 0 \\
n_{(k)}a_{i,0}+ & \sum_{j=1}^{n_{(k)}} b_{i,j} & = & -1 \\
\sum_{j=1}^{k} a_{i,j}+ & a_{i,0} + (k+1)b_{i,1} & = & 0 \\
\sum_{j=1}^{k} a_{i,j}+ & a_{i,0} + (k+1)b_{i,2} & = & 0 \\
 & \vdots & & \\
\sum_{j=1}^{k} a_{i,j}+ & a_{i,0} + (k+1)b_{i,n_{(1)}} & = & 0 \\
\sum_{j=2}^{k} a_{i,j}+ & a_{i,0} + (k)b_{i,n_{(1)}+1} & = & 0 \\
\sum_{j=2}^{k} a_{i,j}+ & a_{i,0} + (k)b_{i,n_{(1)}+2} & = & 0 \\
 & \vdots & & \\
\sum_{j=2}^{k} a_{i,j}+ & a_{i,0} + (k)b_{i,n_{(2)}} & = & 0 \\
 & \vdots & & \\
a_{i,k}+ & a_{i,0} + (2)b_{i,n_{(k)}} & = & 0
\end{array}\right.$$

The solution to the above set of equations is:

$$a_{i,1} = a_{i,2} = \cdots = a_{i,i-1} = 0,$$

$$a_{i,i} = \frac{1}{n_{(i)}},$$

$$a_{i,i+1} = \frac{1}{(k-i+1)n_{(i+1)}} - \frac{1}{(k-i+1)n_{(i)}},$$

for $i + 2 \leq g \leq k$,

$$a_{i,g} = \frac{1}{(k-g+2)n_{(g)}} - \frac{1}{(k-i+1)n_{(i)}} - \sum_{h=i+1}^{g-1} \frac{1}{(k-h+1)(k-h+2)n_{(h)}},$$

$$a_{i,0} = -\frac{1}{(k-i+1)n_{(i)}} - \sum_{h=i+1}^{k} \frac{1}{(k-h+1)(k-h+2)n_{(h)}},$$

$$b_{i,1} = \cdots = b_{i,n_{(1)}} = b_{i,n_{(1)}+1} = \cdots = b_{i,n_{(2)}} = \cdots = b_{i,n_{(i)}} = 0,$$

$$b_{i,n_{(i)}+1} = \cdots = b_{i,n_{(i+1)}} = \frac{1}{(k-i+1)n_{(i)}},$$

for $i+1 \leq g \leq k-1$,

$$b_{i,n_{(g)}+1} = \cdots = b_{i,n_{(g+1)}} = \frac{1}{(k-i+1)n_{(i)}} + \sum_{h=i+1}^{g} \frac{1}{(k-h+1)(k-h+2)n_{(h)}}.$$

No matter what i is, we can always find a vector $\tau_{(i)}$ such that $\mathcal{X}'\mathcal{X}\tau_{(i)} = l_{(i)}$ holds. Therefore $\beta_{(i)} - \beta_0$ is estimable. □

3.A.2 Multivariate comparison of the means in neurobiological characteristics

Suppose there are $M(M \geq 1)$ neurobiological characteristics measured in each subject and we are interested in estimating the differences between the population with dysfunction in the construct of interest and the healthy population in all the characteristics simultaneously.

Let $y_{(i)jm}$ and $\epsilon_{(i)jm}$ denote the mth characteristic measurement and random error for the subject in $D_{(i)}$ and the jth match, respectively. We assume the diagnosis effect as well as the match effect are different for each characteristic, so $\beta_{(i)m}$ and γ_{jm} denote respectively the diagnosis effect for $D_{(i)}$ and match effect for the jth match on characteristic m. Measurements taken on different subjects are assumed to be independently distributed. The random error vector $(\epsilon_{(i)j1}, \epsilon_{(i)j2}, \cdots, \epsilon_{(i)jM})$ on a single subject is assumed to follow the multivariate normal distribution with the following mean vector and covariance matrix.

$$E \begin{bmatrix} \epsilon_{(i)j1} \\ \epsilon_{(i)j2} \\ \vdots \\ \epsilon_{(i)jM} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad Cov \begin{bmatrix} \epsilon_{(i)j1} \\ \epsilon_{(i)j2} \\ \vdots \\ \epsilon_{(i)jM} \end{bmatrix} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1M} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{M1} & \sigma_{M2} & \cdots & \sigma_{MM} \end{bmatrix}.$$

A multivariate ANOVA model can be employed.

$$\begin{aligned} y_{(i)jm} &= \beta_{(i)m} + \gamma_{jm} + \epsilon_{(i)jm}, \quad i = 1, 2, \cdots, k, \quad j = 1, 2, \cdots, n_{(i)}, \quad m = 1, 2, \cdots, M; \\ y_{0jm} &= \beta_{0m} + \gamma_{jm} + \epsilon_{0jm}, \quad j = 1, 2, \cdots, n_{(k)}, \quad m = 1, 2, \cdots, M. \end{aligned} \tag{3.A.2}$$

If we rewrite the model in matrix form, it is easy to see that the design matrix $\mathcal{X}$ stays the same as in the univariate case. The model in matrix form can be written as:

$$\underbrace{\begin{bmatrix} y_{(1)11} & \cdots & y_{(1)1M} \\ y_{(1)21} & \cdots & y_{(1)2M} \\ \vdots & \vdots & \vdots \\ y_{(1)n_{(1)}1} & \cdots & y_{(1)n_{(1)}M} \\ y_{(2)11} & \cdots & y_{(2)1M} \\ y_{(2)21} & \cdots & y_{(2)2M} \\ \vdots & \vdots & \vdots \\ y_{(2)n_{(2)}1} & \cdots & y_{(2)n_{(2)}M} \\ \vdots & \vdots & \vdots \\ y_{(k)11} & \cdots & y_{(k)1M} \\ y_{(k)21} & \cdots & y_{(k)2M} \\ \vdots & \vdots & \vdots \\ y_{(k)n_{(k)}1} & \cdots & y_{(k)n_{(k)}M} \\ y_{011} & \cdots & y_{01M} \\ y_{021} & \cdots & y_{02M} \\ \vdots & \vdots & \vdots \\ y_{0n_{(k)}1} & \cdots & y_{0n_{(k)}M} \end{bmatrix}}_{\boldsymbol{y_1} \ \cdots \ \boldsymbol{y_M}} = \mathcal{X} \underbrace{\begin{bmatrix} \beta_{(1)1} & \cdots & \beta_{(1)M} \\ \beta_{(2)1} & \cdots & \beta_{(2)M} \\ \vdots & \vdots & \vdots \\ \beta_{(k)1} & \cdots & \beta_{(k)M} \\ \beta_{01} & \cdots & \beta_{0M} \\ \gamma_{11} & \cdots & \gamma_{1M} \\ \gamma_{21} & \cdots & \gamma_{2M} \\ \vdots & \vdots & \vdots \\ \gamma_{n_{(1)}1} & \cdots & \gamma_{n_{(1)}M} \\ \vdots & \vdots & \vdots \\ \gamma_{n_{(2)}1} & \cdots & \gamma_{n_{(2)}M} \\ \vdots & \vdots & \vdots \\ \gamma_{n_{(k)}1} & \cdots & \gamma_{n_{(k)}M} \end{bmatrix}}_{\boldsymbol{\beta_1} \ \cdots \ \boldsymbol{\beta_M}} + \underbrace{\begin{bmatrix} \epsilon_{(1)11} & \cdots & \epsilon_{(1)1M} \\ \epsilon_{(1)21} & \cdots & \epsilon_{(1)2M} \\ \vdots & \vdots & \vdots \\ \epsilon_{(1)n_{(1)}1} & \cdots & \epsilon_{(1)n_{(1)}M} \\ \epsilon_{(2)11} & \cdots & \epsilon_{(2)1M} \\ \epsilon_{(2)21} & \cdots & \epsilon_{(2)2M} \\ \vdots & \vdots & \vdots \\ \epsilon_{(2)n_{(2)}1} & \cdots & \epsilon_{(2)n_{(2)}M} \\ \vdots & \vdots & \vdots \\ \epsilon_{(k)11} & \cdots & \epsilon_{(k)1M} \\ \epsilon_{(k)21} & \cdots & \epsilon_{(k)2M} \\ \vdots & \vdots & \vdots \\ \epsilon_{(k)n_{(k)}1} & \cdots & \epsilon_{(k)n_{(k)}M} \\ \epsilon_{011} & \cdots & \epsilon_{01M} \\ \epsilon_{021} & \cdots & \epsilon_{02M} \\ \vdots & \vdots & \vdots \\ \epsilon_{0n_{(k)}1} & \cdots & \epsilon_{0n_{(k)}M} \end{bmatrix}}_{\boldsymbol{\epsilon_1} \ \cdots \ \boldsymbol{\epsilon_M}}.$$

Here for $1 \leq m \leq M, \boldsymbol{y_m}$ represents all the observations on characteristic m, $\boldsymbol{\beta_m}$ is the vector of parameters on characteristic m and $\boldsymbol{\epsilon_m}$ is the random error for $\boldsymbol{y_m}$. It is easy to see that the marginal distribution of each $\boldsymbol{\epsilon_m}$ is multivariate normal with mean $\mathbf{0}$ and covariance matrix $\sigma_{mm} I_{n_{(k)}+\sum_{i=1}^{k} n_{(i)}}$. However, when the M random vectors are stacked together, the covariance matrix for the joint distribution of $(\boldsymbol{\epsilon_1}, \boldsymbol{\epsilon_2}, \cdots, \boldsymbol{\epsilon_M})$ is not diagonal as every $\boldsymbol{\epsilon_m}$ contains the measurements from the same subject. The covariance matrix for $(\boldsymbol{\epsilon_1}, \boldsymbol{\epsilon_2}, \cdots, \boldsymbol{\epsilon_M})$ is

$$Cov \begin{bmatrix} \boldsymbol{\epsilon_1} \\ \boldsymbol{\epsilon_2} \\ \vdots \\ \boldsymbol{\epsilon_M} \end{bmatrix} = \begin{bmatrix} \sigma_{11} I_{n_{(k)}+\sum_{i=1}^{k} n_{(i)}} & \sigma_{12} I_{n_{(k)}+\sum_{i=1}^{k} n_{(i)}} & \cdots & \sigma_{1M} I_{n_{(k)}+\sum_{i=1}^{k} n_{(i)}} \\ \sigma_{21} I_{n_{(k)}+\sum_{i=1}^{k} n_{(i)}} & \sigma_{22} I_{n_{(k)}+\sum_{i=1}^{k} n_{(i)}} & \cdots & \sigma_{2M} I_{n_{(k)}+\sum_{i=1}^{k} n_{(i)}} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{M1} I_{n_{(k)}+\sum_{i=1}^{k} n_{(i)}} & \sigma_{M2} I_{n_{(k)}+\sum_{i=1}^{k} n_{(i)}} & \cdots & \sigma_{MM} I_{n_{(k)}+\sum_{i=1}^{k} n_{(i)}} \end{bmatrix}. \quad (3.A.3)$$

Note that this covariance matrix is a square matrix with dimension $M * (n_{(k)} + \sum_{i=1}^{k} n_{(i)})$.

If we use $\boldsymbol{\Delta} = (\Delta_1, \Delta_2, \cdots, \Delta_M)$ to denote the true differences in the M characteristics between the population with dysfunction and the healthy population and $\hat{\boldsymbol{\Delta}} = (\hat{\Delta}_1, \hat{\Delta}_2, \cdots, \hat{\Delta}_M)$ to be the estimate of $\boldsymbol{\Delta}$, then based on the ANOVA model in (3.A.2), $\boldsymbol{\Delta}$ is nothing but

$$\boldsymbol{\Delta} = \begin{bmatrix} \Delta_1 \\ \Delta_2 \\ \vdots \\ \Delta_M \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^k \pi_{(i)}(\beta_{(i)1} - \beta_{01}) \\ \sum_{i=1}^k \pi_{(i)}(\beta_{(i)2} - \beta_{02}) \\ \vdots \\ \sum_{i=1}^k \pi_{(i)}(\beta_{(i)M} - \beta_{0M}) \end{bmatrix}.$$

Due to the same reason as in the univariate case, each $\beta_{(i)m} - \beta_{0m}$ is estimable and so each $\hat{\Delta}_m$ can be expressed in $\boldsymbol{y_m}$ as in (3.2.7), i.e.,

$$\hat{\Delta}_m = \boldsymbol{\pi}' T' \mathcal{X}' \boldsymbol{y_m}, \quad 1 \leq m \leq M.$$

Therefore,

$$\hat{\boldsymbol{\Delta}} = \begin{bmatrix} \hat{\Delta}_1 \\ \hat{\Delta}_2 \\ \vdots \\ \hat{\Delta}_M \end{bmatrix} = \begin{bmatrix} \boldsymbol{\pi}' T' \mathcal{X}' \boldsymbol{y_1} \\ \boldsymbol{\pi}' T' \mathcal{X}' \boldsymbol{y_2} \\ \vdots \\ \boldsymbol{\pi}' T' \mathcal{X}' \boldsymbol{y_M} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\pi}' T' \mathcal{X}' & \boldsymbol{0'} & \cdots & \boldsymbol{0'} \\ \boldsymbol{0'} & \boldsymbol{\pi}' T' \mathcal{X}' & \cdots & \boldsymbol{0'} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0'} & \boldsymbol{0'} & \cdots & \boldsymbol{\pi}' T' \mathcal{X}' \end{bmatrix} \begin{bmatrix} \boldsymbol{y_1} \\ \boldsymbol{y_2} \\ \vdots \\ \boldsymbol{y_M} \end{bmatrix}.$$

Here $\boldsymbol{0'}$ is a row vector with dimension $n_{(k)} + \sum_{i=1}^k n_{(i)}$ and all the elements equal to 0.

3.A.3 Proof of Equivalence between Quantile Regression Model Statements

Result 3.3. *If the pth quantiles of F_0, the distribution of the error term in the healthy population and of F_1, the distribution of the error term in the mixture population are both 0, the model statement in (3.3.4) and (3.3.5) are the same.*

Proof. For F_0, the pth quantile is 0 means that

$$\begin{aligned}
p &= F_0(0) \\
&= P(\epsilon_{00j} \le 0) \\
&= P(Y_{00j} - \beta_0^p - M_j^p \le 0) \\
&= P(Y_{00j} \le \beta_0^p - M_j^p) \\
&= P(Y_{00j} \le Q_{0j}),
\end{aligned}$$

which means that the pth quantile of the neurobiological characteristic in the healthy population and jth match $Q_{0j} = \beta_0^p + M_j^p$.

For F_1, it is the distribution of the error term of any observation from the mixture population with dysfunction and the jth match. According to previous discussions, F_1 is a mixture distribution with k components and each component specifies the distribution of ϵ_{1ij}.

The pth quantile of F_1 is 0 means that

$$\begin{aligned}
p &= F_1(0) \\
&= \sum_{i=1}^{k} \pi_i P(\epsilon_{1ij} \le 0) \\
&= \sum_{i=1}^{k} \pi_i P(Y_{1ij} - \beta_1^p - M_j^p \le 0) \\
&= \sum_{i=1}^{k} \pi_i P(Y_{1ij} \le \beta_1^p + M_j^p). \qquad (3.A.4)
\end{aligned}$$

Here (3.A.4) means that the pth quantile of the neurobiological characteristic in the mixture population is $\beta_1^p + M_j^p$, i.e., $Q_{1j} = \beta_1^p + M_j^p$.

It is noteworthy that the pth quantile of the neurobiological characteristic in D_i and

Match j is not necessarily $\beta_1^p + M_j^p$ because based on (3.A.4), $P(Y_{1ij} \leq \beta_1^p + M_j^p)$ is not necessarily p. Actually the pth quantile of the neurobiological characteristic in D_i and Match j can be anything. Therefore, although no DSM diagnosis information is involved in the our model, the model does not indicate the differences in the pth quantile of the neurobiological characteristic between each D_i and the healthy population are the same. □

3.A.4 Proof of Different DSM Diagnosis Effect in Quantile Regression

Result 3.4. *Estimation of Δ^p does not require the difference in the pth quantile of the distribution of the neurobiological characteristic between each DSM diagnosis and the healthy population to be the same.*

Proof. Suppose the neurobiological characteristic in the healthy population and in $D_i (i = 1, 2, \cdots, k)$ are distributed as F_0 and F_{1i}, respectively. According to the previous discussions, the distribution of the neurobiological characteristic in the population with dysfunction, F_1, is a mixture of the F_{1i}'s with relative proportion π_i. So

$$F_1 = \sum_{i=1}^{k} \pi_i F_{1i}.$$

Suppose the pth quantiles of the neurobiological characteristic in the population with dysfunction and the healthy population are x_1^p and x_0^p, respectively. Then

$$\begin{aligned} F_0(x_0^p) =& p, \\ F_1(x_1^p) =& \sum_{i=1}^{k} \pi_i F_{1i}(x_1^p) = p. \end{aligned} \tag{3.A.5}$$

And $\Delta^p = x_1^p - x_0^p$ is the difference we would like to estimate between the two populations. That is to say, as long as the model can estimate x_1^p and x_0^p, we can obtain $\hat{\Delta}^p$. However, the estimation of x_1^p does not require in (3.A.5) that $F_{1i}(x_1^p) = p$. Actually the pth quantile of the neurobiological characteristic in each D_i can be anything as long as they satisfy the equation in (3.A.5). Therefore, nothing needs to be assumed for the difference in the pth quantile of the distribution of the neurobiological characteristic between each DSM diagnosis and the healthy population in estimating Δ^p. □

4.0 OPTIMAL DESIGN WITH CONSTRAINTS

In this chapter, we deal with the optimal design problem which was briefly mentioned in Chapter 1.

4.1 MOTIVATION FOR THE OPTIMAL DESIGN

In Section 3.2 of Chapter 3, we have already worked out a formula to estimate the difference in the means of the neurobiological characteristics between the healthy population and the population with dysfunction in the construct of interest. The formula gives us an estimate of the difference with known sample sizes $n_1, n_2, \cdots, n_k$ under the triangular design shown in Table 3.4. However, in designing actual trials, theses sample sizes are unknown. They have to be chosen before a study is designed. A very interesting and practical statistical question in designing a study to compare the means in the neurobiological characteristic between the two populations is how to sample from the two populations using the post-mortem tissue databases. For the healthy population, it is straightforward because subjects in this population are considered to be homogeneous. For the population with dysfunction in the construct of interest, as discussed in previous chapters, it is a mixture of several DSM diagnoses. Thus how to sample from a mixture, i.e., how many subjects are needed for each DSM diagnosis, remains a question.

Moreover, in actual designs involving post-mortem tissue databases, the number of the subjects in those databases for each DSM diagnosis is limited. Also we may not be able to examine every subject that is available to us because the resources for a study are limited. Therefore there are two constraints in designing a study with post-mortem tissue samples where these constraints are described in Section 4.2. The question becomes how to design an experiment optimally to study the differences in the means of neurobiological characteristics between the population with dysfunction and healthy populations under the two types of constraints. Here the criteria we use for the optimality is minimum variance in the estimated differences of the neurobiological characteristics between the two populations. To calculate the sample sizes optimally, the same triangular design is used as in Chapter 3.

4.2 LAYOUT OF THE OPTIMAL DESIGN

4.2.1 An illustration of the Optimal Design

Before detailing the notation of the optimal design problem, we first illustrate the problem with the Stanley Brain Collection (abbreviated as SBC) database. The Stanley Brain Collection is a widely used resource for researchers who study schizophrenia, bipolar disorder and major depressive disorder. Currently there are five cohorts available: the Neuropathology Consortium consisting of 60 brains (15 in each of schizophrenia, bipolar disorder, depression and control, matched by age, sex, race, postmortem interval, pH, side of brain and mRNA quality), the Array Collection consisting of 105 cases (35 in each of schizophrenia, bipolar disorder, and control), the Depression Collection consisting of 36 cases (12 in each of depression with psychotic features, depression without psychotic features and control), the Inferior Parietal Collection of 48 cases (fixed inferior parietal sections from 24 schizophrenia and 24 controls) and a New Collection of 57 cases (19 in each of schizophrenia, bipolar disorder and control). For each of the five cohorts, frozen sections of tissue are available from different areas of the brain. Neurobiological characteristics such as RNAs or proteins can be measured in these tissue sections. Researchers can request these tissue samples for further scientific investigations.

Now if we are interested in investigating the fundamental biological mechanism that leads to psychosis, some RNA levels from the post-mortem tissue samples could be measured to compare the psychotic population with the healthy population. If we were to use the Stanley Brain Collection to conduct a study using the occipital section of a brain, the Inferior Parietal Collection is not applicable because it only has the parietal region of the brain. Therefore, there are at most 69 schizophrenic subjects, 69 bipolar disordered subjects, 27 depressed subjects and 81 healthy control subjects we can choose from. Suppose the budget allows 100 subjects to be processed, then we want to design a matched sample study with at most 100 subjects in total. We would like to determine the number of subjects needed from each DSM diagnosis and the healthy population so that the estimated differences in the RNA levels between the psychotic and healthy population have the smallest variance among all designs using 100 or fewer subjects.

4.2.2 Notation of Optimal Design

We follow the same notation that has been used in Table 3.1. Let N_i be the known number of available post-mortem subjects for D_i and n_i be the unknown sample size chosen for D_i. Let N_0 and n_0 be, respectively, the known number of available and unknown number of chosen subjects from the healthy population. Also we use n to denote the total number of subjects that the budget allows. Obviously $n_i \leq N_i$ for $0 \leq i \leq k$ and $\sum_{i=0}^{k} n_i \leq n$. The goal of this chapter is to find $(n_0, n_1, \cdots, n_k) \in \mathcal{F}$ s.t. $\text{Var}(\hat{\Delta})$ is minimized, where

$$\mathcal{F} = \{(n_0, n_1, \cdots, n_k) : n_i \leq N_i, 0 \leq i \leq k \text{ and } \sum_{i=0}^{k} n_i \leq n\}. \tag{4.2.1}$$

4.2.3 Hypothetical Numerical Example of Optimal Design

In order to clearly explain the optimal design problem and how we determine the sample size for each DSM diagnosis optimally, a hypothetical numerical example is used to illustrate the ideas. Suppose we are using the SBC database and the population with dysfunction in the construct of interest involves $k = 3$ DSM diagnoses which are schizophrenia, bipolar disorder and depression. We further assume that schizophrenia patients account for 70%, bipolar disordered patients account for 20% and depression patients account for 10% among this mixture population. Under the notation above, we know immediately that $\pi_1 = 0.1, \pi_2 = 0.2$ and $\pi_3 = 0.7$ because $\pi_1 < \pi_2 < \pi_3$ has to hold. Therefore schizophrenia is D_3, bipolar disorder is D_2 and depression is D_1 under our notation.

Suppose hypothetically that the number of available post-mortem subjects with appropriate tissue for each DSM diagnosis is $N_1 = 4, N_2 = 3, N_3 = 2$ and the number of available healthy control subjects is $N_0 = 5$. Also assume that the number of subjects our budget allows us to process is $n = 10$. Note that n is the total number of subjects that includes both the subjects with dysfunction and the healthy controls. Now our goal is to determine the sample size n_i for each DSM diagnosis and the healthy population such that $(n_0, n_1, n_2, n_3) \in \mathcal{F}$ in (4.2.1) and $\text{Var}(\hat{\Delta})$ is minimized.

4.2.4 Assumptions for Optimal Design

In addition to the assumption we have made in Section 3.2 for the triangular design, some more assumptions need to be made to determine the optimal sample sizes.

Assumption 4.1. $N_0 \geq \max(N_1, N_2, \cdots, N_k)$

This assumption parallels Result 3.1 so that we don't have to worry about the constraint for the healthy population in the sample size determination. In other words, we assume we are always able to find a healthy control subject for every match. This assumption is reasonable due to the fact that most post-mortem tissue databases have more healthy control subjects than any particular DSM diagnosis in this database. In the hypothetical example in Section 4.2.3, we assume $N_1 = 4, N_2 = 3, N_3 = 2, N_0 = 5$. Obviously $N_0 \geq \max(N_1, N_2, N_3)$ so that this assumption holds.

Result 3.1 and Assumption 4.1 together lead to the following dimensional change in the definition of the set $\mathcal{F}$ in (4.2.1).

Result 4.1. *Under Assumption 3.2, the set $\mathcal{F}$ in (4.2.1) can be written as $\mathcal{F}_a$:*

$$\mathcal{F}_a = \{(n_1, \cdots, n_k) : n_i \leq N_i, i = 1, \cdots, k \textit{ and } \sum_{i=1}^{k} n_i + n_{(k)} \leq n\}. \tag{4.2.2}$$

Proof. Let $N_{(i)}$ be the constraint of sample size for $D_{(i)}$. The condition $n_i \leq N_i$ in set $\mathcal{F}$ can be written equivalently as $n_{(i)} \leq N_{(i)}$ for $1 \leq i \leq k$. So we have $n_{(k)} \leq N_{(k)}$. Note that $N_{(k)}$ is just the constraint of the sample size in $D_{(k)}$ and thus it has to be smaller than or equal to the maximum of all the constraints of the DSM diagnoses, i.e., $N_{(k)} \leq \max(N_1, N_2, \cdots, N_k)$. Then $n_{(k)} \leq N_{(k)} \leq \max(N_1, N_2, \cdots, N_k)$. Replace n_0 by $n_{(k)}$ and with $N_0 \geq \max(N_1, N_2, \cdots, N_k)$, we could have $n_0 \leq \max(N_1, N_2, \cdots, N_k) \leq N_0$. This means Result 3.1 and Assumption 4.1, together with $n_i \leq N_i (1 \leq i \leq k)$ make the condition $n_0 \leq N_0$ automatically holds in set $\mathcal{F}$. Dropping the index 0 in $\mathcal{F}$ and setting n_0 to $n_{(k)}$ in the condition $\sum_{i=0}^{k} n_i \leq n$ give us the set $\mathcal{F}_a$ in (4.2.2). □

Now our goal becomes to find $(n_1, n_2, \cdots, n_k) \in \mathcal{F}_a$ s.t. $\text{Var}(\hat{\Delta})$ is minimized.

4.3 STUDY DESIGN WITH UNKNOWN SAMPLE SIZES

As described previously, we use the same triangular design as in Section 3.2 for the optimal design problem. The sample size for each DSM diagnosis is determined optimally as having the smallest variance of the estimated difference among all the possible triangular designs in the set $\mathcal{F}_a$ in (4.2.2). It is noteworthy that the design is "triangular" only in the sense of the ordered chosen sample size, i.e., the $n_{(i)}$'s. This is not an issue in Section 3.2 because there the sample sizes are known and thus the order of the sample sizes is known also. However, in this chapter, the sample sizes for the DSM diagnoses are to be determined and thus the order of them is still unknown when designing a study. Table 3.4 does not indicate how the n_i's are ordered. Recall $D_{(i)}$ is defined by the ascending chosen sample size $n_{(i)}$ and thus it is unknown which DSM diagnosis $D_{(i)}$ represents before the sample sizes are determined. $D_{(i)}$ may not necessarily be D_i, which is defined by the ascending relative proportion π_i. Without prior knowledge about how the n_i's are ordered, the triangular design in Table 3.4 is equivalent to one of $k!$ scenarios with different possible orderings of the n_i's. And the sample size n_i is determined by considering all the scenarios. Depending on the ordering of the sample sizes, set $\mathcal{F}_a$ also changes.

In the example of Section 4.2.3 with $k = 3$, we have $3! = 6$ possible orderings of n_1, n_2 and n_3, and hence $3! = 6$ possible $(D_{(1)}, D_{(2)}, D_{(3)})$. Ignoring the healthy controls, the triangular design could be one of the 6 scenarios listed in Table 4.1. Our goal is to find the optimal sample size under all of the 6 possible designs in set $\mathcal{F}_a$. Note that $\mathcal{F}_a$ changes in the 6 scenarios as $n_{(3)}$ could be one of n_1, n_2 or n_3.

Table 4.1: Possible Triangular Designs for the example in Section 4.2.3

order of n_1, n_2, n_3	design		
$n_1 \leq n_2 \leq n_3$	$D_{(1)} = D_1$	$D_{(2)} = D_2$	$D_{(3)} = D_3$
	$n_{(1)} = n_1$	$n_{(2)} = n_2$	$n_{(3)} = n_3$
	$N_{(1)} = 4$	$N_{(2)} = 3$	$N_{(3)} = 2$
	$\mathcal{F}_a = \{(n_1, n_2, n3) : n_1 \leq 4, n_2 \leq 3, n_3 \leq 2, \sum_{i=1}^{3} n_i + n_3 \leq 10\}$		
$n_1 \leq n_3 \leq n_2$	$D_{(1)} = D_1$	$D_{(2)} = D_3$	$D_{(3)} = D_2$
	$n_{(1)} = n_1$	$n_{(2)} = n_3$	$n_{(3)} = n_2$
	$N_{(1)} = 4$	$N_{(2)} = 2$	$N_{(3)} = 3$
	$\mathcal{F}_a = \{(n_1, n_2, n3) : n_1 \leq 4, n_2 \leq 3, n_3 \leq 2, \sum_{i=1}^{3} n_i + n_2 \leq 10\}$		
$n_2 \leq n_1 \leq n_3$	$D_{(1)} = D_2$	$D_{(2)} = D_1$	$D_{(3)} = D_3$
	$n_{(1)} = n_2$	$n_{(2)} = n_1$	$n_{(3)} = n_3$
	$N_{(1)} = 3$	$N_{(2)} = 4$	$N_{(3)} = 2$
	$\mathcal{F}_a = \{(n_1, n_2, n3) : n_1 \leq 4, n_2 \leq 3, n_3 \leq 2, \sum_{i=1}^{3} n_i + n_3 \leq 10\}$		
$n_2 \leq n_3 \leq n_1$	$D_{(1)} = D_2$	$D_{(2)} = D_3$	$D_{(3)} = D_1$
	$n_{(1)} = n_2$	$n_{(2)} = n_3$	$n_{(3)} = n_1$
	$N_{(1)} = 3$	$N_{(2)} = 2$	$N_{(3)} = 4$
	$\mathcal{F}_a = \{(n_1, n_2, n3) : n_1 \leq 4, n_2 \leq 3, n_3 \leq 2, \sum_{i=1}^{3} n_i + n_1 \leq 10\}$		
$n_3 \leq n_1 \leq n_2$	$D_{(1)} = D_3$	$D_{(2)} = D_1$	$D_{(3)} = D_2$
	$n_{(1)} = n_3$	$n_{(2)} = n_1$	$n_{(3)} = n_2$
	$N_{(1)} = 2$	$N_{(2)} = 4$	$N_{(3)} = 3$
	$\mathcal{F}_a = \{(n_1, n_2, n3) : n_1 \leq 4, n_2 \leq 3, n_3 \leq 2, \sum_{i=1}^{3} n_i + n_2 \leq 10\}$		
$n_3 \leq n_2 \leq n_1$	$D_{(1)} = D_3$	$D_{(2)} = D_2$	$D_{(3)} = D_1$
	$n_{(1)} = n_3$	$n_{(2)} = n_2$	$n_{(3)} = n_1$
	$N_{(1)} = 2$	$N_{(2)} = 3$	$N_{(3)} = 4$
	$\mathcal{F}_a = \{(n_1, n_2, n3) : n_1 \leq 4, n_2 \leq 3, n_3 \leq 2, \sum_{i=1}^{3} n_i + n_1 \leq 10\}$		

4.4 VARIANCE OF $\hat{\Delta}$

As derived in Section 3.2, if $\mathcal{X}$ is the design matrix specified in (3.2.4), $\tau_{(i)}$ is the vector such that $\widehat{\beta_{(i)} - \beta_0} = \tau'_{(i)}\mathcal{X}'\boldsymbol{y}$ and matrix $T = (\tau_{(1)}\, \tau_{(2)}\, \cdots\, \tau_{(k)})$, then $\hat{\Delta}$ can be expressed as a linear function of the data vector $\boldsymbol{y}$ as in (3.2.7). Because the covariance matrix of $\boldsymbol{y}$ is known to be $\sigma^2 I_{n_{(k)}+\sum_{i=1}^{k} n_{(i)}}$, the variance of $\hat{\Delta}$ can be written as:

$$\begin{aligned} Var(\hat{\Delta}) &= Var(\boldsymbol{\pi}'T'\mathcal{X}'\boldsymbol{y}) \\ &= \boldsymbol{\pi}'T'\mathcal{X}'\mathcal{X}T\boldsymbol{\pi}\sigma^2, \end{aligned} \tag{4.4.1}$$

where $\boldsymbol{\pi}' = (\pi_{(1)}\, \pi_{(2)}\, \cdots\, \pi_{(k)})$.

Now we want to derive $T'\mathcal{X}'\mathcal{X}T$ in terms of $n_{(i)}$ so that Var($\hat{\Delta}$) can be expressed with $\pi_{(i)}$ and $n_{(i)}$ and thus be minimized. The derivation is shown in Appendix 4.A.1 and Var($\hat{\Delta}$) is given in (4.A.3) as:

$$Var(\hat{\Delta}) = \sigma^2 \sum_{i=1}^{k} \frac{[\pi_{(1)} + \pi_{(2)} + \cdots + \pi_{(i-1)} + (k-i+2)\pi_{(i)}]^2}{n_{(i)}(k-i+1)(k-i+2)}.$$

Note again that the variance formula involves the ordered unknown sample size $n_{(i)}$ and relative proportion $\pi_{(i)}$, and thus can not be minimized directly. We need to express the above formula in terms of n_i and π_i so that Var($\hat{\Delta}$) can be minimized. Depending on how the unknown sample sizes are ordered, there are $k!$ possible expressions of the Var($\hat{\Delta}$) formula in terms of n_i and π_i. Recall that depending on how the unknown sample sizes are ordered and which DSM diagnosis has the largest sample size, there are k possible $\mathcal{F}_a$'s. Therefore each Var($\hat{\Delta}$) expression is minimized in the corresponding set $\mathcal{F}_a$.

In the hypothetical example in Section 4.2.3 where $k = 3$, the variance of $\hat{\Delta}$ becomes

$$Var(\hat{\Delta}) = \frac{[4\pi_{(1)}]^2}{3*4*n_{(1)}}\sigma^2 + \frac{[\pi_{(1)} + 3\pi_{(2)}]^2}{2*3*n_{(2)}}\sigma^2 + \frac{[\pi_{(1)} + \pi_{(2)} + 2\pi_{(3)}]^2}{1*2*n_{(3)}}\sigma^2.$$

However, the above formula cannot be minimized directly because we don't know the order of the n_1, n_2 and n_3 beforehand, so that we don't know which DSM diagnosis $D_{(i)}$ corresponds to and thus we don't know what $\pi_{(i)}$ is. As there are $3! = 6$ possible ways to order n_1, n_2 and n_3,

Table 4.2: Expression of Var($\hat{\Delta}$) for the example in Section 4.2.3

order of n_1, n_2, n_3	Var($\hat{\Delta}$)
$n_1 \leq n_2 \leq n_3$	$\frac{[4\pi_1]^2}{3*4*n_1}\sigma^2 + \frac{[\pi_1+3\pi_2]^2}{2*3*n_2}\sigma^2 + \frac{[\pi_1+\pi_2+2\pi_3]^2}{1*2*n_3}\sigma^2$
$n_1 \leq n_3 \leq n_2$	$\frac{[4\pi_1]^2}{3*4*n_1}\sigma^2 + \frac{[\pi_1+3\pi_3]^2}{2*3*n_3}\sigma^2 + \frac{[\pi_1+\pi_2+2\pi_2]^2}{1*2*n_2}\sigma^2$
$n_2 \leq n_1 \leq n_3$	$\frac{[4\pi_2]^2}{3*4*n_2}\sigma^2 + \frac{[\pi_2+3\pi_1]^2}{2*3*n_1}\sigma^2 + \frac{[\pi_2+\pi_1+2\pi_3]^2}{1*2*n_3}\sigma^2$
$n_2 \leq n_3 \leq n_1$	$\frac{[4\pi_2]^2}{3*4*n_2}\sigma^2 + \frac{[\pi_2+3\pi_3]^2}{2*3*n_3}\sigma^2 + \frac{[\pi_2+\pi_3+2\pi_1]^2}{1*2*n_1}\sigma^2$
$n_3 \leq n_1 \leq n_2$	$\frac{[4\pi_3]^2}{3*4*n_3}\sigma^2 + \frac{[\pi_3+3\pi_1]^2}{2*3*n_1}\sigma^2 + \frac{[\pi_3+\pi_1+2\pi_2]^2}{1*2*n_2}\sigma^2$
$n_3 \leq n_2 \leq n_1$	$\frac{[4\pi_3]^2}{3*4*n_3}\sigma^2 + \frac{[\pi_3+3\pi_2]^2}{2*3*n_2}\sigma^2 + \frac{[\pi_3+\pi_2+2\pi_1]^2}{1*2*n_2}\sigma^2$

the variance formula in this example has 6 possible expressions in terms of $n_1, n_2, n_3, \pi_1, \pi_2$ and π_3, which are listed in Table 4.2.

In this example, to minimize Var($\hat{\Delta}$), we need to specify the exact expressions given in Table 4.2 and minimize each expression in an corresponding set $\mathcal{F}_a$, which is listed in Table 4.1.

4.5 MINIMIZATION OF VAR($\hat{\Delta}$)

The sample size for each DSM diagnosis is determined by minimizing Var($\hat{\Delta}$) in (4.A.3) in set $\mathcal{F}_a$. In theory, we can enumerate all the possible points in $\mathcal{F}_a$, calculate the variance of $\hat{\Delta}$ according to (4.A.3) and select the one with the smallest variance. However, simple enumeration is very computationally expensive, especially when k, n and each of N_i is large. Another way to achieve the minimization in set $\mathcal{F}_a$ is through convex optimization, which is to minimize the variance formula (4.A.3). In this section, we first introduce the concepts of convex set and convex function, and show the reason why direct minimization of Var($\hat{\Delta}$) in $\mathcal{F}_a$ is difficult. Then we try to minimize Var($\hat{\Delta}$) in a upper set $\mathcal{F}_u$ of $\mathcal{F}_a$ and see if the minimum of Var($\hat{\Delta}$) in $\mathcal{F}_u$ is in $\mathcal{F}_a$. If it is, then we are done. If the minimum is outside $\mathcal{F}_a$, then it means the sample size for at least one DSM diagnosis is beyond the corresponding constraint. In this case, we set the sample size for such DSM diagnosis to be its upper bound and recompute the minimum. In other words, the minimum is recalculated by minimizing $\mathrm{Var}(\hat{\Delta})$ on the boundary of $\mathcal{F}_a$. The minimum of $\mathrm{Var}(\hat{\Delta})$ in the upper set $\mathcal{F}_u$ can be shown to exist in a particular subset $\mathcal{F}_0$.

4.5.1 Difficulty in Minimization of Var($\hat{\Delta}$) in $\mathcal{F}_a$

4.5.1.1 Convex Set and Convex Function Convex minimization is done on convex functions with domains of convex sets. Convex functions and convex sets are defined as below.

Definition 4.1 (convex set)**.** *A set C is convex if the line segment between any two points in C lies in C, i.e., for any $x_1, x_2 \in C$ and $0 \leq \theta \leq 1$, we have $\theta x_1 + (1-\theta)x_2 \in C$*

Definition 4.2 (convex function)**.** *If $\mathbf{D}_f$ denotes the domain of a function f, then f is convex if $\mathbf{D}_f$ is convex and for any $x, y \in \mathbf{D}_f$ and $0 \leq \theta \leq 1$, we have*

$$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta)f(y). \tag{4.5.1}$$

A function f is strictly convex if strict inequality holds in (4.5.1) *for $x \neq y$ and $0 < \theta < 1$.*

To show that the variance formula in (4.A.3) is a convex function on $\mathcal{F}_a$, we show $\mathcal{F}_a$ is a convex set first. Although the set $\mathcal{F}_a$ as defined in (4.2.2) contains the order statistic $n_{(k)}$ and can be expressed as k different sets depending on which sample size $n_{(k)}$ is, in the proof that it is a convex set, there is no need to specify what $n_{(k)}$ is. Here $n_{(k)}$ can just be used as a symbol and whether it is n_1, n_2 or any other n_i does not matter. In the proof, $\max(n_1, \cdots, n_k)$ is used to stand for $n_{(k)}$.

Result 4.2. *$\mathcal{F}_a$ is a convex set.*

Proof. Suppose both $(n_{1a}, n_{2a}, \cdots, n_{ka})$ and $(n_{1b}, n_{2b}, \cdots, n_{kb})$ are points in $\mathcal{F}_a$, then by definition of set $\mathcal{F}_a$,

$$\begin{aligned}
\sum_{i=1}^{k} n_{ia} + \max(n_{1a}, \cdots, n_{ka}) &\leq n, \\
\sum_{i=1}^{k} n_{ib} + \max(n_{1b}, \cdots, n_{kb}) &\leq n, \\
n_{ia} \leq N_i, \quad i &= 1, 2, \cdots, k, \\
n_{ib} \leq N_i, \quad i &= 1, 2, \cdots, k.
\end{aligned}$$

For any $0 \leq \theta \leq 1$, we have

$$\begin{aligned}
&\theta \sum_{i=1}^{k} n_{ia} + (1-\theta) \sum_{i=1}^{k} n_{ib} + \max(\theta n_{1a} + (1-\theta) n_{1b}, \cdots, \theta n_{ka} + (1-\theta) n_{kb}) \\
\leq &\theta(n - \max(n_{1a}, \cdots, n_{ka})) + (1-\theta)(n - \max(n_{1b}, \cdots, n_{kb})) \\
&+ \max(\theta n_{1a} + (1-\theta) n_{1b}, \cdots, \theta n_{ka} + (1-\theta) n_{kb}) \\
= &n - \max(\theta n_{1a}, \cdots, \theta n_{ka})) - \max((1-\theta) n_{1a}, \cdots, (1-\theta) n_{ka})) \\
&+ \max(\theta n_{1a} + (1-\theta) n_{1b}, \cdots, \theta n_{ka} + (1-\theta) n_{kb}) \\
\leq &n - \max(\theta n_{1a} + (1-\theta) n_{1b}, \cdots, \theta n_{ka} + (1-\theta) n_{kb}) \\
&+ \max(\theta n_{1a} + (1-\theta) n_{1b}, \cdots, \theta n_{ka} + (1-\theta) n_{kb}) \\
= &n;
\end{aligned}$$

also for $i = 1, 2, \cdots, k$,

$$\theta n_{ia} + (1-\theta) n_{ib} \leq \theta N_i + (1-\theta) N_i = N_i.$$

Therefore, $(\theta n_{1a} + (1-\theta) n_{1b}, \theta n_{2a} + (1-\theta) n_{2b}, \cdots, \theta n_{ka} + (1-\theta) n_{kb})$ is also in $\mathcal{F}_a$ and thus $\mathcal{F}_a$ is a convex set. □

Usually it is not easy to show a function is convex using Definition 4.2. Fortunately, there is an important property about convex function given in Boyd & Vandenberghe (2004) which makes it easy to show convexity of a function. The property is stated as a lemma below and we use this property to show that the variance formula (4.A.3) is convex.

Lemma 4.1 (Boyd & Vandenberghe (2004))**.** *If the Hessian matrix of a function, which is the second order derivative matrix of the function, exists at each point in its domain, then the function is convex if and only if the Hessian matrix is positive semidefinite. The function is strictly convex if and only if the Hessian matrix is positive definite.*

Now we show the variance function in (4.A.3) is strictly convex on the set $\mathcal{F}_a$. Although the function in (4.A.3) is written generally in terms of $n_{(i)}$ and $\pi_{(i)}$, the actual parameters are n_i and π_i as (4.A.3) is a function on the set of $\mathcal{F}_a$. The $n_{(i)}$ can be any n_i depending on the ordering of $n_1, n_2, \cdots, n_k$ and thus formula (4.A.3) has $k!$ possible expressions in terms of n_i and π_i as discussed previously. To avoid showing convexity for each of the $k!$ possible expressions, we use $n_{(i)}$ as a symbol as in the proof of $\mathcal{F}_a$ being a convex set.

Result 4.3. *Var*$(\hat{\Delta})$ *expressed in* (4.A.3) *is strictly convex on* $\mathcal{F}_a$.

Proof. The first derivative of $\text{Var}(\hat{\Delta})$ with respect to $n_{(i)}$ is

$$\frac{\partial Var(\hat{\Delta})}{\partial n_{(i)}} = -\frac{[\pi_{(1)} + \pi_{(2)} + \cdots + \pi_{(i-1)} + (k-i+2)\pi_{(i)}]^2}{n_{(i)}^2 (k-i+1)(k-i+2)}\sigma^2.$$

The Hessian matrix $\bigtriangledown^2 Var(\hat{\Delta})$ is:

$$\bigtriangledown^2 Var(\hat{\Delta}) = \begin{bmatrix} \frac{2[(k+1)\pi_{(1)}]^2}{n_{(1)}^3 k(k+1)}\sigma^2 & 0 & \cdots & 0 \\ 0 & \frac{2[\pi_{(1)}+k\pi_{(2)}]^2}{n_{(2)}^3 (k-1)k}\sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{2[\pi_{(1)}+\pi_{(2)}+\cdots+\pi_{(k-1)}+2\pi_{(k)}]^2}{n_{(k)}^3 1*2}\sigma^2 \end{bmatrix}.$$

Again although we use $n_{(i)}$ in the Hessian matrix, $(n_{(1)}, n_{(2)}, \cdots, n_{(k)})$ can be any ordering of $(n_1, n_2, \cdots, n_k)$. It is easy to see that $\bigtriangledown^2 Var(\hat{\Delta})$ exists at any point in $\mathcal{F}_a$ and is a diagonal matrix with all the diagonal elements positive, so $\bigtriangledown^2 Var(\hat{\Delta})$ is positive definite according to Lemma 4.1. Because $\mathcal{F}_a$ is a convex set, $\text{Var}(\hat{\Delta})$ is strictly convex on $\mathcal{F}_a$. □

4.5.1.2 KKT Conditions Optimization of a convex function can be solved using the Karush-Kuhn-Tucker (KKT) conditions, which involves a set of equations formed with the Lagrange multiplier. As Boyd & Vandenberghe (2004) pointed out, for a strictly convex function, the KKT conditions are sufficient and necessary conditions for a solution to be optimal, given that some regulatory conditions hold. The KKT conditions allow both equality and inequality constraints in the optimization, which fits the need in our problem.

In order to minimize $Var(\hat{\Delta})$ expressed in (4.A.3) in the set $\mathcal{F}_a$, consider the Lagrangian $L(n_{(1)}, \cdots, n_{(k)}, \lambda_0, \cdots, \lambda_{k-1}, \eta_1, \cdots, \eta_k)$, where,

$$
\begin{aligned}
L =& \sigma^2 \sum_{i=1}^{k} \frac{[\pi_{(1)} + \pi_{(2)} + \cdots + \pi_{(i-1)} + (k-i+2)\pi_{(i)}]^2}{n_{(i)}(k-i+1)(k-i+2)} \\
&+ \lambda_0(n_{(1)} + n_{(2)} + \cdots + n_{(k-1)} + 2n_{(k)} - n) \\
&+ \lambda_1(n_{(1)} - n_{(2)}) \\
&+ \lambda_2(n_{(2)} - n_{(3)}) \\
&+ \cdots \\
&+ \lambda_{k-1}(n_{(k-1)} - n_{(k)}) \\
&+ \eta_1(n_{(1)} - N_{(1)}) \\
&+ \eta_2(n_{(2)} - N_{(2)}) \\
&+ \cdots \\
&+ \eta_k(n_{(k)} - N_{(k)}).
\end{aligned}
$$

Following Boyd & Vandenberghe (2004), we find the KKT conditions form a set of equations:

$$\left\{\begin{array}{l}
\frac{\partial L}{\partial n_{(1)}} = -\frac{[(k+1)\pi_{(1)}]^2}{n_{(1)}^2 k(k+1)}\sigma^2 + \lambda_0 + \lambda_1 + \eta_1 = 0 \\
\frac{\partial L}{\partial n_{(2)}} = -\frac{[\pi_{(1)}+k\pi_{(2)}]^2}{n_{(2)}^2 (k-1)k}\sigma^2 + \lambda_0 - \lambda_1 + \lambda_2 + \eta_2 = 0 \\
\vdots \\
\frac{\partial L}{\partial n_{(k-1)}} = -\frac{[\pi_{(1)}+\cdots+\pi_{(k-2)}+3\pi_{(k-1)}]^2}{n_{(k-1)}^2 2*3}\sigma^2 + \lambda_0 - \lambda_{k-2} + \lambda_{k-1} + \eta_k = 0 \\
\frac{\partial L}{\partial n_{(k)}} = -\frac{[\pi_{(1)}+\cdots+\pi_{(k-1)}+2\pi_{(k)}]^2}{n_{(k)}^2 1*2}\sigma^2 + 2\lambda_0 - \lambda_{k-1} + \eta_k = 0 \\
\lambda_0 \geq 0, \lambda_1 \geq 0, \cdots, \lambda_{k-1} \geq 0, \eta_1 \geq 0, \cdots, \eta_k \geq 0 \\
n_{(1)} + n_{(2)} + \cdots + n_{(k-1)} + 2n_{(k)} - n \leq 0 \\
n_{(1)} - n_{(2)} \leq 0 \\
\vdots \\
n_{(k-1)} - n_{(k)} \leq 0 \\
n_{(1)} - N_{(1)} \leq 0 \\
\vdots \\
n_{(k)} - N_{(k)} \leq 0 \\
\lambda_0(n_{(1)} + n_{(2)} + \cdots + n_{(k-1)} + 2n_{(k)} - n) = 0 \\
\lambda_1(n_{(1)} - n_{(2)}) = 0 \\
\vdots \\
\lambda_{k-1}(n_{(k-1)} - n_{(k)}) = 0 \\
\eta_1(n_{(1)} - N_{(1)}) = 0 \\
\vdots \\
\eta_k(n_{(k)} - N_{(k)}) = 0
\end{array}\right. \tag{4.5.2}$$

Here again to avoid writing the Lagrangian and equation set for each of the $k!$ possible expressions of formula (4.A.3), the order statistic $(n_{(1)}, n_{(2)}, \cdots, n_{(k)})$ is used. In the actual process of solving the equation set, the order of $(n_1, n_2, \cdots, n_k)$ needs to be known. The solution to the above equation set is the minimum of $\text{Var}(\hat{\Delta})$ for a given ordering of $(n_1, n_2, \cdots, n_k)$ in the corresponding set $\mathcal{F}_a$. In other words, for each ordering of $(n_1, n_2, \cdots, n_k)$, an equation set in (4.5.2) can be formed and thus there are $k!$ possible equation sets. Intuitively, we should solve each of the equation sets and obtain $k!$ minima. The minimum of the $k!$ minima is the optimal point that gives the desired sample sizes. However, on one hand, as k increases, the number of equation sets increases very quickly and it would be tedious to consider all the orderings of $(n_1, n_2, \cdots, n_k)$. On the other hand, even if the order of

$(n_1, n_2, \cdots, n_k)$ is known and thus $\pi_{(i)}$ is known, the equation set (4.5.2) is difficult to solve as shown next.

4.5.1.3 Difficulty with Minimizing Var($\hat{\Delta}$) in $\mathcal{F}_a$ As can be seen from equation set (4.5.2), for a given ordering of $(n_1, n_2, \cdots, n_k)$, there are $3k$ unknown parameters, $3k$ equations and $4k$ inequality constraints. The $3k$ equations can be used to solve for the $3k$ unknowns provided the solution exists. However, to have a solution for the KKT conditions equation set, we also need to check if the inequalities hold at the solution point. A common way to solve the equation set is to consider all the scenarios each of which depends on if each one of the Lagrangian multipliers is positive or zero. Because there are $2k$ Lagrangian multipliers, there are in total 2^{2k} scenarios to consider. Therefore, even if the order of $(n_1, n_2, \cdots, n_k)$ is known, solving (4.5.2) is very complicated, especially when k is large. We illustrate that in Appendix 4.A.3 for the simplest case where $k = 2$, it is not easy to get a solution.

As can be seen from the illustration, when the order of $(n_1, n_2, \cdots, n_k)$ is known, in addition to a large number of scenarios to consider, there is still a challenge in obtaining a solution. The challenge is that what the solution is depends on the relationship among $N_1, \cdots, N_k$ and n, especially the order of the k upper bounds $N_1, \cdots, N_k$. Without knowing the relationship and order, we are unable to write down a closed form solution.

4.5.2 Set $\mathcal{F}_u$ and $\mathcal{F}_o$

As discussed in previous sections, there are two problems we need to solve to get the optimal sample sizes through convex optimization. The first problem is to solve the equation set (4.5.2) for a given ordering of $(n_1, n_2, \cdots, n_k)$ and obtain a minimum. The second one is to find the minimum of all the $k!$ minima. One way to get around the first problem is to remove some of the constraints first. In other words, we first minimize Var($\hat{\Delta}$) in an upper set of $\mathcal{F}_a$, which we call $\mathcal{F}_u$, and see if the minimum is in $\mathcal{F}_a$. If it is, then the minimum in $\mathcal{F}_a$ is found. Otherwise, it means the calculated sample sizes for some DSM diagnoses are larger than their corresponding upper bound. In this case we fix the sample sizes for these

DSM diagnoses at their corresponding upper bounds and recalculate the sample sizes in $\mathcal{F}_u$. That is, we recalculate the sample size in $\mathcal{F}_u$ but also on the boundary of $\mathcal{F}_a$. The definition of set $\mathcal{F}_u$ is:

$$\mathcal{F}_u = \{(n_1, \cdots, n_k) : \sum_{i=1}^{k} n_i + n_{(k)} \leq n\}.$$

It can be seen from the definition of set $\mathcal{F}_u$ that the constraint $n_i \leq N_i$ is removed. Therefore, it is easier to solve for a minimum in $\mathcal{F}_u$ as there are fewer parameters and we don't have to worry about the N_i's. However, similar as $\mathcal{F}_a$, set $\mathcal{F}_u$ lacks the ordering information of $(n_1, n_2, \cdots, n_k)$, so that we still don't know which n_i is $n_{(i)}$ and which π_i is $\pi_{(i)}$ when minimizing (4.A.3) in $\mathcal{F}_u$. Therefore, the second problem in obtaining the optimal sample size mentioned above is still unsolved by minimizing in the upper set $\mathcal{F}_u$. Depending on the orderings of $(n_1, n_2, \cdots, n_k)$, set $\mathcal{F}_u$ can be partitioned into $k!$ mutually exclusive subsets each with known $n_{(i)}$ and $p_{(i)}$. Minimizing in $\mathcal{F}_u$ with a known ordering of $(n_1, n_2, \cdots, n_k)$ is just minimizing in a particular subset of $\mathcal{F}_u$ and the global minimum in $\mathcal{F}_u$ is the smallest among the $k!$ minima from each subset. Because $k!$ could be a very large number, it is unwise to minimize in all of the $k!$ subsets. Our work would be reduced if we know which subset would give the smallest minimum. Then we can first minimize (4.A.3) in this subset and see if the minimum is in $\mathcal{F}_a$. It is shown in the next section that the minimum of $\text{Var}(\hat{\Delta})$ in $\mathcal{F}_u$ exists in the particular subset $\mathcal{F}_o$ which is defined as:

$$\mathcal{F}_o = \{(n_1, \cdots, n_k) : \sum_{i=1}^{k} n_i + n_{(k)} \leq n \textbf{ and } n_1 \leq n_2 \leq \cdots \leq n_k\}.$$

Here we can see that $\mathcal{F}_0$ is a very special subset of $\mathcal{F}_u$ in the sense that the sample sizes have the same order as the relative proportions. When the sample sizes exist in $\mathcal{F}_0$, it is obvious that $D_{(i)} = D_i, n_{(i)} = n_i, \pi_{(i)} = \pi_i$ and $N_{(i)} = N_i$. For instance, among the population with dysfunction in the construct of interest, D_1, the DSM diagnosis with the smallest proportion would need the fewest subject. Since we want to minimize $\text{Var}(\hat{\Delta})$ in $\mathcal{F}_u$ and $\mathcal{F}_0$ through convex optimization, we need to show both $\mathcal{F}_u$ and $\mathcal{F}_0$ are convex sets. The proof is very similar to that of Result 4.2.

Result 4.4. *$\mathcal{F}_u$ and $\mathcal{F}_0$ are both convex sets.*

Proof. Suppose both $(n_{1a}, n_{2a}, \cdots, n_{ka})$ and $(n_{1b}, n_{2b}, \cdots, n_{kb})$ are points in $\mathcal{F}_u$, then by definition,

$$\sum_{i=1}^{k} n_{ia} + \max(n_{1a}, \cdots, n_{ka}) \leq n,$$
$$\sum_{i=1}^{k} n_{ib} + \max(n_{1b}, \cdots, n_{kb}) \leq n.$$

For any $0 \leq \theta \leq 1$, we have

$$\begin{aligned}
&\theta \sum_{i=1}^{k} n_{ia} + (1-\theta) \sum_{i=1}^{k} n_{ib} + \max(\theta n_{1a} + (1-\theta) n_{1b}, \cdots, \theta n_{ka} + (1-\theta) n_{kb}) \\
\leq &\theta (n - \max(n_{1a}, \cdots, n_{ka})) + (1-\theta)(n - \max(n_{1b}, \cdots, n_{kb})) \\
&+ \max(\theta n_{1a} + (1-\theta) n_{1b}, \cdots, \theta n_{ka} + (1-\theta) n_{kb}) \\
= &n - \max(\theta n_{1a}, \cdots, \theta n_{ka})) - \max((1-\theta) n_{1a}, \cdots, (1-\theta) n_{ka})) \\
&+ \max(\theta n_{1a} + (1-\theta) n_{1b}, \cdots, \theta n_{ka} + (1-\theta) n_{kb}) \\
\leq &n - \max(\theta n_{1a} + (1-\theta) n_{1b}, \cdots, \theta n_{ka} + (1-\theta) n_{kb}) \\
&+ \max(\theta n_{1a} + (1-\theta) n_{1b}, \cdots, \theta n_{ka} + (1-\theta) n_{kb}) \\
= &n.
\end{aligned}$$

Therefore, $(\theta n_{1a} + (1-\theta) n_{1b}, \theta n_{2a} + (1-\theta) n_{2b}, \cdots, \theta n_{ka} + (1-\theta) n_{kb})$ is also in $\mathcal{F}_u$ and thus $\mathcal{F}_u$ is a convex set.

If we further require that $n_{1a} \leq n_{2a} \leq \cdots \leq n_{ka}$ and $n_{1b} \leq n_{2b} \leq \cdots \leq n_{kb}$, then both $(n_{1a}, n_{2a}, \cdots, n_{ka})$ and $(n_{1b}, n_{2b}, \cdots, n_{kb})$ are points in $\mathcal{F}_0$. It is easy to see that

$$\theta n_{(1a)} + (1-\theta) n_{(1b)} \leq \theta n_{(2a)} + (1-\theta) n_{(2b)} \leq \cdots \leq \theta n_{(ka)} + (1-\theta) n_{(kb)}.$$

So $(\theta n_{1a} + (1-\theta) n_{1b}, \theta n_{2a} + (1-\theta) n_{2b}, \cdots, \theta n_{ka} + (1-\theta) n_{kb})$ is also in $\mathcal{F}_0$. Then by definition, $\mathcal{F}_0$ is a convex set too. □

The relationship of $\mathcal{F}_u$, $\mathcal{F}_a$ and $\mathcal{F}_o$ can be represented in Figure 4.1. Here, the big circle stands for $\mathcal{F}_u$ and the small circle stands for $\mathcal{F}_a$. The shaded area stands for $\mathcal{F}_0$, which is a particular subset of $\mathcal{F}_u$ and also intersects with $\mathcal{F}_a$.

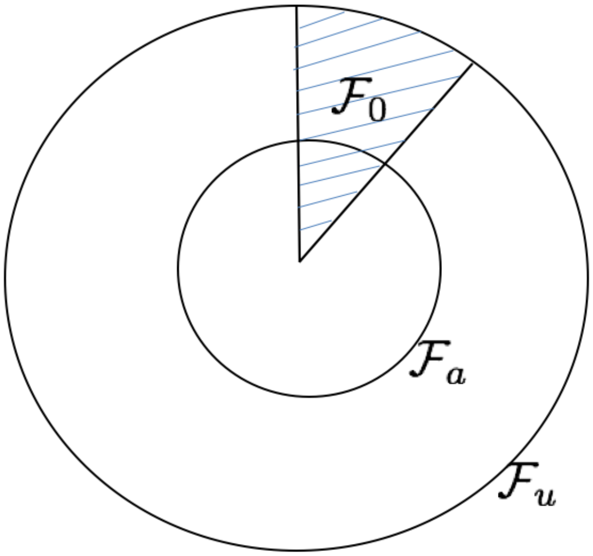


Figure 4.1: Relationship of $\mathcal{F}_u$, $\mathcal{F}_a$ and $\mathcal{F}_o$

4.5.3 Global Minimum of Var$(\hat{\Delta})$ in $\mathcal{F}_u$

In this section, we prove the global minimum of Var$(\hat{\Delta})$ in $\mathcal{F}_u$ is in the particularly ordered subset $\mathcal{F}_0$ and show how to calculate the minimum in $\mathcal{F}_0$. If this minimum is not inside $\mathcal{F}_a$, we show in the next section how to minimize Var$(\hat{\Delta})$ in $\mathcal{F}_u$ but on the boundary of $\mathcal{F}_a$, i.e.,with some $n_i = N_i$.

4.5.3.1 Permutation of π_i Based on the definition of $D_{(i)}$ and $\pi_{(i)}$, for any triangular design with given $(n_{(i)})$ and π_i, permuting the sample sizes is the same as permuting the relative proportions. For example, consider the simplest case where there are two DSM diagnoses with $(\pi_1, \pi_2) = (0.1, 0.9)$ and $(n_{(1)}, n_{(2)}) = (5, 10)$. There could be two possible designs, which are $n_1 = 5, n_2 = 10$ and $n_1 = 10, n_2 = 5$. For each possible design, we can represent it either using $D_{(i)}$ where the sample sizes are ascending or using D_i where the relative proportions are ascending. As can be shown in Table 4.3, table (a) and table (b) are two different representations of the design $n_1 = 5, n_2 = 10$ and table (c) and table (d) are two different ways to represent the design $n_1 = 10, n_2 = 5$. Either permuting the relative proportions in Table 4.3(a) or permuting the sample sizes in Table 4.3(b) converts the design with $n_1 = 5, n_2 = 10$ to the one with $n_1 = 10, n_2 = 5$. In order to show that the global minimum of $\text{Var}(\hat{\Delta})$ in $\mathcal{F}_u$ is in $\mathcal{F}_0$, a subset with special ordering of $(n_1, n_2, \cdots, n_k)$, we could permute either the relative proportions or the sample sizes. Here we choose to permute the relative proportions because the study is represented using $D_{(i)}$ as shown in Table 3.4.

Table 4.3: Illustration of permuting sample sizes and relative proportions

(a) design $n_1 = 5, n_2 = 10$

$D_{(1)}$	$D_{(2)}$	Control
5	10	10
0.1	0.9	

(b) design $n_1 = 5, n_2 = 10$

D_1	D_2	Control
5	10	10
0.1	0.9	

(c) switching π_1, π_2 in (a)

$D_{(1)}$	$D_{(2)}$	Control
5	10	10
0.9	0.1	

(d) switching n_1, n_2 in (b)

D_1	D_2	Control
10	5	10
0.1	0.9	

As there are a total of $k!$ possible permutations of $\pi_1, \pi_2, \cdots, \pi_k$, there are $k!$ ways the vector $(\pi_{(1)}, \pi_{(2)}, \cdots, \pi_{(k)})$ can be realized. Through pairwise switching, different realizations can be converted to each other. In order to show that the minimum of $\text{Var}(\hat{\Delta})$ in $\mathcal{F}_u$ exists in $\mathcal{F}_o$, i.e., when $\pi_{(i)} = \pi_i$, we first show that every pairwise switch of the values of $\pi_{(p)}$ and $\pi_{(q)}$ enlarges $\text{Var}(\hat{\Delta})$ if $\pi_{(p)} < \pi_{(q)}$ and $p < q$. We then show $\pi_{(i)} = \pi_i$ can be achieved through a few steps of such pairwise switches and each step decreases the variance.

Result 4.5. *For given $(n_{(1)}, n_{(2)}, \cdots, n_{(k)})$ and $(\pi_1, \pi_2, \cdots, \pi_k)$, if $p < q$, Var$(\hat{\Delta})$ with $\pi_{(p)} < \pi_{(q)}$ is smaller than the Var$(\hat{\Delta})$ after switching the values of $\pi_{(p)}$ and $\pi_{(q)}$.*

Proof. See Appendix 4.A.4. □

This result says $\text{Var}(\hat{\Delta})$ would be different in the $k!$ subsets each with a different ordering of the known $(n_{(1)}, n_{(2)}, \cdots, n_{(k)})$. It also indicates that if we hold the values of $n_{(1)}, n_{(2)}, \cdots, n_{(k)}$ fixed, we can always obtain a smaller variance of $\hat{\Delta}$ by switching a pair of $\pi_{(i)}$'s and finally achieve the minimum. And the switch should go in the direction such that the DSM diagnosis with the smaller relative proportion would have the smaller sample size. In other words, the variance reduces if the pairwise switch of $\pi_{(i)}$'s makes the order of the sample sizes and the order of the relative proportions conform. When the sample sizes from some DSM diagnoses are equal, the pairwise switching of the $\pi_{(i)}$'s within these groups does not change the variance. This implies that if all the sample sizes are equal, it does not matter how $\pi_{(i)}$ is defined because the variance stays the same.

To illustrate the above result, consider the hypothetical example in Section 4.2.3. Suppose in the following four designs shown in Table 4.4, $n_{(1)} = 1, n_{(2)} = 2, n_{(3)} = 3$. The control group is not included in the tables, but by our assumption, it has 3 subjects. Assume $\sigma = 1$, the variance of $\hat{\Delta}$ in $\mathcal{F}_u$ is computed for each case.

Table 4.4: Variance comparison for four designs

(a) $\mathrm{Var}(\hat{\Delta}) = 0.996$

0.7	0.2	0.1
$D_{(1)}$	$D_{(2)}$	$D_{(3)}$
1	2	3

(b) $\mathrm{Var}(\hat{\Delta}) = 0.977$

0.7	0.1	0.2
$D_{(1)}$	$D_{(2)}$	$D_{(3)}$
1	2	3

(c) $\mathrm{Var}(\hat{\Delta}) = 0.657$

0.1	0.7	0.2
$D_{(1)}$	$D_{(2)}$	$D_{(3)}$
1	2	3

(d) $\mathrm{Var}(\hat{\Delta}) = 0.536$

0.1	0.2	0.7
$D_{(1)}$	$D_{(2)}$	$D_{(3)}$
1	2	3

In Table 4.4, adjacent tables are different only in the order of one pair of $\pi_{(i)}$'s. The $\mathrm{Var}(\hat{\Delta})$ in design (a) is larger than that in design (b) because in design (a), $\pi_{(2)}$ is bigger than $\pi_{(3)}$, which means the DSM diagnosis with a larger relative proportion has a smaller sample size. As shown in Result 4.5, as long as the order of n_i's and that of π_i's conform after the switch, the $\mathrm{Var}(\hat{\Delta})$ is reduced.

Result 4.5 only deals with pairwise switch. It still remains unknown which permutation of $(\pi_{(1)}, \pi_{(2)}, \cdots, \pi_{(k)})$ has the smallest $\mathrm{Var}(\hat{\Delta})$ for a given set of $(n_{(1)}, n_{(2)}, \cdots, n_{(k)})$. The following result shows that if $\pi_{(i)} = \pi_i$ for all i, $\mathrm{Var}(\hat{\Delta})$ is the smallest. That is to say, the minimum of $\mathrm{Var}(\hat{\Delta})$ in $\mathcal{F}_u$ exists in $\mathcal{F}_o$.

Result 4.6. $\underset{\mathcal{F}_u}{\operatorname{argmin}}\ \mathit{Var}(\hat{\Delta}) \in \mathcal{F}_o$.

Proof. Represent the design in the triangular form as in Table 3.4 so that the DSM diagnoses are listed in the ordered sample sizes. For a given set of $(n_{(1)}, n_{(2)}, \cdots, n_{(k)})$, start from π_1, the smallest relative proportion. Suppose $\pi_{(i)} = \pi_1$, then switch $\pi_{(i)}$ pairwisely with each of the $\pi_{(j)}$'s where $\pi_{(j)} < \pi_1$ and $1 \leq j < i$ until $\pi_{(1)} = \pi_1$. In each step we get a smaller $\text{Var}(\hat{\Delta})$ based on Result 4.5 because the switch makes the order of relative proportions and that of the sample sizes conform. Repeat the process for π_2 until $\pi_{(2)} = \pi_2$. Again we get a smaller $\text{Var}(\hat{\Delta})$ in each step. Repeat the process for $\pi_3, \cdots, \pi_k$ until $\pi_{(k)} = \pi_k$ and we will have the smallest $\text{Var}(\hat{\Delta})$. The process is illustrated in Table 4.4, where $\pi_1 = 0.1$ is switched to $\pi_{(1)}$ from (a) to (c) and $\pi_2 = 0.2$ is switched to $\pi_{(2)}$ from (c) to (d). Finally $\pi_i = \pi_{(i)}$ and there is nothing that can be switched to achieve a smaller $\text{Var}(\hat{\Delta})$.

Through the above process, we obtain the smallest $\text{Var}(\hat{\Delta})$ when $\pi_{(i)} = \pi_i$ for a given set of $(n_{(1)}, n_{(2)}, \cdots, n_{(k)})$. By definition π_i is the relative proportion for the DSM diagnosis with sample size n_i and $\pi_{(i)}$ is the relative proportion for the DSM diagnosis with sample size $n_{(i)}$, thus $\pi_{(i)} = \pi_i$ means the DSM diagnosis with sample size $n_{(i)}$ is the same DSM diagnosis with sample size n_i. Therefore $\pi_{(i)} = \pi_i$ is equivalent as $n_{(i)} = n_i$ and $D_{(i)} = D_i$. Because $n_{(1)} \leq n_{(2)} \leq \cdots \leq n_{(k)}$, the n_i's are also ordered such that $n_1 \leq n_2 \leq \cdots \leq n_k$. In other words, for a given set of $(n_{(1)}, n_{(2)}, \cdots, n_{(k)})$, $\text{Var}(\hat{\Delta})$ is the smallest when $n_1 \leq n_2 \leq \cdots \leq n_k$, i.e., the sample size for each DSM diagnosis increases as the relative proportion increases.

Based on the above argument, the point that minimizes $\text{Var}(\hat{\Delta})$ in $\mathcal{F}_u$ must be in $\mathcal{F}_o$. Otherwise, we can always find a permutation in $\mathcal{F}_0$ to have the same $(n_{(1)}, n_{(2)}, \cdots, n_{(k)})$ but give a smaller variance. Therefore, $\underset{\mathcal{F}_u}{\text{argmin}}\ \text{Var}(\hat{\Delta}) \in \mathcal{F}_o$. □

4.5.3.2 Minimization of Var($\hat{\Delta}$) in $\mathcal{F}_o$ Since $\mathcal{F}_o$ is a convex set, $\text{Var}(\hat{\Delta})$ can be minimized by convex optimization through solving the KKT equation set again. Because the sample sizes $n_1, n_2, \cdots, n_k$ are ordered in $\mathcal{F}_o$, the variance formula (4.A.3) can be written directly without the order statistic notation as:

$$Var(\hat{\Delta}) = \sigma^2 \sum_{i=1}^{k} \frac{[\pi_1 + \pi_2 + \cdots + \pi_{i-1} + (k-i+2)\pi_i]^2}{n_i(k-i+1)(k-i+2)}.$$

If we use $L_0(n_1, \cdots, n_k, \lambda_0, \cdots, \lambda_{k-1})$ to denote the Lagrangian in this case, where

$$\begin{aligned}L_0(n_1, \cdots, n_k, \lambda_0, \cdots, \lambda_{k-1}) =& \sigma^2 \sum_{i=1}^{k} \frac{[\pi_1 + \pi_2 + \cdots + \pi_{i-1} + (k-i+2)\pi_i]^2}{n_i(k-i+1)(k-i+2)} \\ &+ \lambda_0(n_1 + n_2 + \cdots + n_{k-1} + 2n_k - n) \\ &+ \lambda_1(n_1 - n_2) \\ &+ \cdots \\ &+ \lambda_{k-1}(n_{k-1} - n_k),\end{aligned}$$

then following Boyd & Vandenberghe (2004), the KKT condition says:

$$\begin{cases} \frac{\partial L_0}{\partial n_1} = -\frac{[(k+1)\pi_1]^2}{n_1^2 k(k+1)}\sigma^2 + \lambda_0 + \lambda_1 = 0 \\ \frac{\partial L_0}{\partial n_2} = -\frac{[\pi_1 + k\pi_2]^2}{n_2^2 (k-1)k}\sigma^2 + \lambda_0 - \lambda_1 + \lambda_2 = 0 \\ \vdots \\ \frac{\partial L_0}{\partial n_k} = -\frac{[\pi_1 + \cdots + \pi_{k-1} + 2\pi_k]^2}{n_k^2 1 * 2}\sigma^2 + 2\lambda_0 - \lambda_{k-1} = 0 \\ \lambda_0 \geq 0, \lambda_1 \geq 0, \cdots, \lambda_{k-1} \geq 0 \\ n_1 + n_2 + \cdots + n_{k-1} + 2n_k - n \leq 0 \\ n_1 - n_2 \leq 0 \\ \vdots \\ n_{k-1} - n_k \leq 0 \\ \lambda_0(n_1 + n_2 + \cdots + n_{k-1} + 2n_k - n) = 0 \\ \lambda_1(n_1 - n_2) = 0 \\ \vdots \\ \lambda_{k-1}(n_{k-1} - n_k) = 0 \end{cases} .$$

In theory, we need to find all the $(n_1^*, \cdots, n_k^*)$ and $(\lambda_0^*, \cdots, \lambda_{k-1}^*)$ that satisfies the above equation set. However, as Boyd & Vandenberghe (2004) showed any local minimum of a strictly convex function is a global minimum and there exists at most one global minimum of a strictly convex function, if we can find one point $(n_1^*, \cdots, n_k^*, \lambda_0^*, \cdots, \lambda_{k-1}^*)$ that solves the above equation set, that point is the global minimum. Obviously, $\lambda_0^*, \lambda_1^*, \cdots, \lambda_{k-1}^*$ cannot all be 0. Now we set $\lambda_1^* = \cdots = \lambda_{k-1}^* = 0$ and $\lambda_0^* > 0$, then we have

$$\begin{cases} -\frac{[(k+1)\pi_1]^2}{n_1^{*2}k(k+1)}\sigma^2 + \lambda_0^* = 0 \\ -\frac{[\pi_1+k\pi_2]^2}{n_2^{*2}(k-1)k}\sigma^2 + \lambda_0^* = 0 \\ \vdots \\ -\frac{[\pi_1+\cdots+\pi_{k-2}+3\pi_{k-1}]^2}{n_{k-1}^{*2}2*3}\sigma^2 + \lambda_0^* = 0 \\ -\frac{[\pi_1+\cdots+\pi_{k-1}+2\pi_k]^2}{n_k^{*2}1*2}\sigma^2 + 2\lambda_0^* = 0 \\ n_1^* + n_2^* + \cdots + n_{k-1}^* + 2n_k^* - n = 0 \end{cases} .$$

This leads to,

$$\begin{cases} \lambda_0^* = [\frac{1}{n}(\sum_{i=1}^{k-1} \frac{\pi_1+\cdots+\pi_{i-1}+(k-i+2)\pi_i}{\sqrt{(k-i+1)(k-i+2)}} + 1 + \pi_k)]^2 > 0 \\ n_j^* = \frac{\frac{\pi_1+\cdots+\pi_{j-1}+(k-j+2)\pi_j}{\sqrt{(k-j+1)(k-j+2)}}}{\sum_{i=1}^{k-1} \frac{\pi_1+\cdots+\pi_{i-1}+(k-i+2)\pi_i}{\sqrt{(k-i+1)(k-i+2)}}+1+\pi_k} n \quad (1 \le j \le k-1) \\ n_k^* = \frac{(1+\pi_k)/2}{\sum_{i=1}^{k-1} \frac{\pi_1+\cdots+\pi_{i-1}+(k-i+2)\pi_i}{\sqrt{(k-i+1)(k-i+2)}}+1+\pi_k} n \end{cases} . \tag{4.5.3}$$

Note that because $\pi_1 < \pi_2 < \cdots < \pi_k$, we have for $1 \le j \le k-2$,

$$\begin{aligned} \frac{n_j^*}{n_{j+1}^*} &= \frac{\pi_1 + \cdots + \pi_{j-1} + (k-j+2)\pi_j}{\pi_1 + \cdots + \pi_{j-1} + \pi_j + (k-j+1)\pi_{j+1}} \sqrt{\frac{(k-j)(k-j+1)}{(k-j+1)(k-j+2)}} \\ &= \frac{\pi_1 + \cdots + \pi_{j-1} + \pi_j + (k-j+1)\pi_j}{\pi_1 + \cdots + \pi_{j-1} + \pi_j + (k-j+1)\pi_{j+1}} \sqrt{\frac{k-j}{k-j+2}} \\ &< 1 \end{aligned}$$

and

$$\begin{aligned} \frac{n_{k-1}^*}{n_k^*} &= \frac{\pi_1 + \cdots + \pi_{k-2} + 3\pi_{k-1}}{\pi_1 + \cdots + \pi_{k-2} + \pi_{k-1} + 2\pi_k} \frac{2}{\sqrt{2*3}} \\ &= \frac{\pi_1 + \cdots + \pi_{k-2} + \pi_{k-1} + 2\pi_{k-1}}{\pi_1 + \cdots + \pi_{k-2} + \pi_{k-1} + 2\pi_k} \frac{2}{\sqrt{2*3}} \\ &< 1. \end{aligned}$$

This means

$$n_1^* < n_2^* < \cdots < n_k^*.$$

The solution $(n_1^*, \cdots, n_k^*, \lambda_0^*, \cdots, \lambda_{k-1}^*)$ in (4.5.3) satisfies the KKT conditions, so according to Boyd & Vandenberghe (2004), this is the global minimum in $\mathcal{F}_o$. By Result 4.6, this is the global minimum of $\text{Var}(\hat{\Delta})$ in $\mathcal{F}_u$.

4.5.3.3 Numerical Example of Minimization in $\mathcal{F}_o$ The above section says that if we want to calculate the sample size optimally for each DSM diagnosis in estimating the difference between the population with dysfunction and the healthy population using the triangular design, and if we only have the constraint for the total number of subjects that we can process, then the number of subjects in each DSM diagnosis can be calculated as in (4.5.3) in theory. In the numerical example in Section 4.2.3 with $\pi_1 = 0.1, \pi_2 = 0.2, \pi_3 = 0.7$ and $n = 10$, the sample size that gives the smallest variance of $\hat{\Delta}$ in $\mathcal{F}_u$ is calculated as:

$$
\begin{aligned}
n_1^\star &= \frac{\frac{(3-1+2)*0.1}{\sqrt{(3-1+1)(3-1+2)}}}{\frac{(3-1+2)*0.1}{\sqrt{(3-1+1)(3-1+2)}} + \frac{0.1+(3-2+2)*0.2}{\sqrt{(3-2+1)(3-2+2)}} + 1 + 0.7} * 10 = 0.550,\\
n_2^\star &= \frac{\frac{0.1+(3-2+2)*0.2}{\sqrt{(3-2+1)(3-2+2)}}}{\frac{(3-1+2)*0.1}{\sqrt{(3-1+1)(3-1+2)}} + \frac{0.1+(3-2+2)*0.2}{\sqrt{(3-2+1)(3-2+2)}} + 1 + 0.7} * 100 = 1.360,\\
n_3^\star &= \frac{\frac{1+0.7}{2}}{\frac{(3-1+2)*0.1}{\sqrt{(3-1+1)(3-1+2)}} + \frac{0.1+(3-2+2)*0.2}{\sqrt{(3-2+1)(3-2+2)}} + 1 + 0.7} * 100 = 4.045.
\end{aligned}
$$

According to our assumption, the healthy population also has sample size $n_3^\star$. The numbers computed with formula in (4.5.3) are not integers, because they give the theoretically smallest variance in $\mathcal{F}_u$, which is not required to be an integer set. When working with actual study designs, the integer point "closest" to $(n_1^\star, n_2^\star, \cdots, n_k^\star)$ should be used.

4.5.4 Minimization of Var($\hat{\Delta}$) in $\mathcal{F}_u$ with some $n_i = N_i$

As discussed previously, after the global minimum of Var($\hat{\Delta}$) in $\mathcal{F}_u$ is calculated as in (4.5.3), whether the minimum is inside $\mathcal{F}_a$ or not needs to be checked. If it is inside $\mathcal{F}_a$, which corresponds to the intersection of $\mathcal{F}_a$ and the shaded area in Figure 4.1, then we have also obtained the global minimum in $\mathcal{F}_a$. Otherwise, if the minimum is inside $\mathcal{F}_0$ but outside $\mathcal{F}_a$, it means the sample sizes for some DSM diagnoses at this global minimum are larger than the corresponding upper bounds. Thus we need to minimize Var($\hat{\Delta}$) in $\mathcal{F}_u$ again but this time also on the boundary of $\mathcal{F}_a$. To do so we only need to force the sample sizes for the DSM diagnoses with $n_{(i)}^\star \geq N_{(i)}$ to be the corresponding $N_{(i)}$. After a new minimum is obtained, we need to check if the new minimum is inside $\mathcal{F}_a$ or not. If the new minimum is inside $\mathcal{F}_a$, then we have found the minimum desired. Otherwise, we repeat the process of minimizing $\text{Var}(\hat{\Delta})$ in $\mathcal{F}_u$ with some $n_i = N_i$ and check if the newly obtained minimum is inside $\mathcal{F}_a$ or not.

When forcing the sample sizes for some DSM diagnoses to be equal to the corresponding constraints, we are still minimizing in $\mathcal{F}_u$, so the convex optimization through solving the KKT equation set is still applicable. However, as discussed in preceding sections, the ordering information of the sample sizes is still lacked and which π_i is $\pi_{(i)}$ remains unknown. Theoretically we need to minimize in each of $k!$ subsets of $\mathcal{F}_u$ with some $n_i = N_i$. But practically not all the $k!$ possible permutations needs to be considered because for those DSM diagnoses with sample sizes forced at the corresponding upper bounds, their sample size ordering is known, and for those DSM diagnoses whose sample sizes are to be determined again, we can still order them according to their relative proportion based on Result 4.5. Suppose there are s DSM diagnoses with sample sizes forced at the upper bounds, then there are $\binom{k}{s}$ ways that the k DSM diagnoses can be order such that s of the them have increasing upper bounds and the remaining $k - s$ have increasing relative proportions. In other words, there are $\binom{k}{s}$ subsets of $\mathcal{F}_u$ to be considered in minimization of $\text{Var}(\hat{\Delta})$ with some $n_i = N_i$, which is significantly smaller than $k!$. In each subset, we know which π_i is $\pi_{(i)}$ and we have an index $\mathcal{S}$ to denote the index whose $n_{(i)} = N_{(i)}$, where the definition of $\mathcal{S}$ is as follows:

Definition 4.3. *Suppose there are s DSM diagnoses whose sample sizes are forced at the corresponding upper bounds, we use $\mathcal{S} = \{i_1, i_2, \cdots, i_s : i_1 < i_2 < \cdots < i_s\}$ to denote the index set such that for $i \in \mathcal{S}, n_{(i)} = N_{(i)}$.*

According to the above argument, there is no one step closed form solution for minimization of $\text{Var}(\hat{\Delta})$ in $\mathcal{F}_u$ with some $n_i = N_i$. What we have to do is to consider all the $\binom{k}{s}$ permutations of the sample sizes and see which one has the smallest variance of $\hat{\Delta}$. For each one of these $\binom{k}{s}$ cases, $\pi_{(i)}$ is known. However, even with $\pi_{(i)}$ known, if the KKT conditions equation set is written down and to be solved, it is easy to see that sometimes there is no solution for it because the inequalities in the equation set may not all be satisfied. One way to get around is to ignore the order restriction in the KKT conditions equation set first. If

$I(i \in \mathcal{S})$ denotes the indicator function such that

$$\begin{cases} I(i \in \mathcal{S}) = 1 & i \in \mathcal{S} \\ I(i \in \mathcal{S}) = 0 & i \notin \mathcal{S} \end{cases},$$

then ignoring the order restriction means Lagrangian $L_{\mathcal{S}}(n_{(1)}, \cdots, n_{(k)}, \lambda_0, \kappa_1, \cdots, \kappa_k)$ is considered to solve each one of the $\binom{k}{s}$ KKT conditions equation sets, which is

$$\begin{aligned} L_{\mathcal{S}} =& \sigma^2 \sum_{i=1}^{k} \frac{[\pi_{(1)} + \pi_{(2)} + \cdots + \pi_{(i-1)} + (k-i+2)\pi_{(i)}]^2}{n_{(i)}(k-i+1)(k-i+2)} \\ &+ \lambda_0(n_{(1)} + n_{(2)} + \cdots + n_{(k-1)} + 2n_{(k)} - n) \\ &+ \sum_{i=1}^{k} \kappa_i I(i \in \mathcal{S})(n_{(i)} - N_{(i)}). \end{aligned}$$

What we do then is to solve the KKT equation set generated by $L_{\mathcal{S}}$ and see if the solution has the increasing order as we have assumed. If it does, then this solution is considered eligible, $\text{Var}(\hat{\Delta})$ at this solution is calculated and compared with that from other eligible solutions. Otherwise this case is not eligible for comparison of the variances.

4.5.4.1 Algorithm to Minimize Var($\hat{\Delta}$) in $\mathcal{F}_u$ with some $n_i = N_i$

To minimize $\text{Var}(\hat{\Delta})$ in $\mathcal{F}_u$ with some $n_i = N_i$, we need to solve a KKT conditions equation set for each one of the $\binom{k}{s}$ permutations of the DSM diagnoses first. The KKT equations generated by $L_{\mathcal{S}}$ are

$$\begin{cases} \frac{\partial L_{\mathcal{S}}}{\partial n_{(i)}} = -\frac{[\pi_{(1)}+\pi_{(2)}+\cdots+\pi_{(i-1)}+(k-i+2)\pi_{(i)}]^2}{n_{(i)}^2(k-i+1)(k-i+2)}\sigma^2 + \lambda_0 + \kappa_i I(i \in \mathcal{S}) = 0, \quad 1 \le i \le k-1 \\ \frac{\partial L}{\partial n_{(k)}} = -\frac{[\pi_{(1)}+\cdots+\pi_{(k-1)}+2\pi_{(k)}]^2}{n_{(k)}^2 1*2}\sigma^2 + 2\lambda_0 + \kappa_k I(k \in \mathcal{S}) = 0 \\ \lambda_0 \ge 0 \\ n_{(1)} + n_{(2)} + \cdots + n_{(k-1)} + 2n_{(k)} - n \le 0 \\ n_{(i)} - N_{(i)} = 0, \quad i \in \mathcal{S} \\ \lambda_0(n_{(1)} + n_{(2)} + \cdots + n_{(k-1)} + 2n_{(k)} - n) = 0 \end{cases}. \tag{4.5.4}$$

Obviously the solution $\lambda_0^\star \neq 0$, otherwise the first equation in (4.5.4) does not hold when $i \notin \mathcal{S}$. There are two scenarios that needs to be considered when solving (4.5.4).

1. $I(k \in \mathcal{S}) = 1$

   Then we have

$$\begin{aligned} n^{\star}_{(k)} &= N_{(k)} \\ n^{\star 2}_{(i)} &= \frac{[\pi_{(1)} + \pi_{(2)} + \cdots + \pi_{(i-1)} + (k-i+2)\pi_{(i)}]^2}{\lambda^{\star}_0 (k-i+1)(k-i+2)} \sigma^2, \quad i \notin \mathcal{S} \\ \sum_{i \notin \mathcal{S}} n^{\star}_{(i)} &= n - \sum_{j \in \mathcal{S}} N_{(j)} - N_{(k)}. \end{aligned}$$

   This leads to

$$\begin{aligned} n^{\star}_{(i)} &= N_{(i)}, \quad i \in \mathcal{S} \\ n^{\star}_{(i)} &= \frac{\frac{\pi_{(1)}+\pi_{(2)}+\cdots+\pi_{(i-1)}+(k-i+2)\pi_{(i)}}{\sqrt{(k-i+1)(k-i+2)}}}{\sum_{g \notin \mathcal{S}} \frac{\pi_{(1)}+\pi_{(2)}+\cdots+\pi_{(g-1)}+(k-g+2)\pi_{(g)}}{\sqrt{(k-g+1)(k-g+2)}}} (n - \textstyle\sum_{j \in \mathcal{S}} N_{(j)} - N_{(k)}), \quad i \notin \mathcal{S} \\ \lambda^{\star}_0 &= \frac{(\sum_{g \notin \mathcal{S}} \frac{\pi_{(1)}+\pi_{(2)}+\cdots+\pi_{(g-1)}+(k-g+2)\pi_{(g)}}{\sqrt{(k-g+1)(k-g+2)}})^2}{(n - \sum_{j \in \mathcal{S}} N_{(j)} - N_{(k)})^2} \sigma^2. \end{aligned} \tag{4.5.5}$$

2. $I(k \in \mathcal{S}) = 0$

   Under this situation, we have

$$\begin{aligned} n^{\star 2}_{(i)} &= \frac{[\pi_{(1)} + \pi_{(2)} + \cdots + \pi_{(i-1)} + (k-i+2)\pi_{(i)}]^2}{\lambda^{\star}_0 (k-i+1)(k-i+2)} \sigma^2, \quad i \notin \mathcal{S}, i \neq k; \\ n^{\star 2}_{(k)} &= \frac{[\pi_{(1)} + \pi_{(2)} + \cdots + \pi_{(k-1)} + 2\pi_{(k)}]^2}{2\lambda^{\star}_0 1 * 2} \sigma^2, \quad k \notin \mathcal{S}; \\ \sum_{i \notin \mathcal{S}} n^{\star}_{(i)} + n^{\star}_{(k)} &= n - \sum_{j \in \mathcal{S}} N_{(j)}. \end{aligned}$$

   and this leads to

$$\begin{aligned} n^{\star}_{(i)} &= N_{(i)}, \quad i \in \mathcal{S} \\ n^{\star}_{(i)} &= \frac{\frac{\pi_{(1)}+\pi_{(2)}+\cdots+\pi_{(i-1)}+(k-i+2)\pi_{(i)}}{\sqrt{(k-i+1)(k-i+2)}}}{\sum_{g \notin \mathcal{S}, g \neq k} \frac{\pi_{(1)}+\pi_{(2)}+\cdots+\pi_{(g-1)}+(k-g+2)\pi_{(g)}}{\sqrt{(k-g+1)(k-g+2)}} + 1 + \pi_{(k)}} (n - \textstyle\sum_{j \in \mathcal{S}} N_{(i)}), \quad i \notin \mathcal{S}, i \neq k \\ n^{\star}_{(k)} &= \frac{\frac{1+\pi_{(k)}}{2}}{\sum_{g \notin \mathcal{S}, g \neq k} \frac{\pi_{(1)}+\pi_{(2)}+\cdots+\pi_{(g-1)}+(k-g+2)\pi_{(g)}}{\sqrt{(k-g+1)(k-g+2)}} + 1 + \pi_{(k)}} (n - \textstyle\sum_{j \in \mathcal{S}} N_{(i)}), \quad k \notin \mathcal{S} \\ \lambda^{\star}_0 &= \frac{(\sum_{g \notin \mathcal{S}, g \neq k} \frac{\pi_{(1)}+\pi_{(2)}+\cdots+\pi_{(g-1)}+(k-g+2)\pi_{(g)}}{\sqrt{(k-g+1)(k-g+2)}} + 1 + \pi_{(k)})^2}{(n - \sum_{j \in \mathcal{S}} N_{(j)})^2} \sigma^2. \end{aligned} \tag{4.5.6}$$

Because we ignore $n^\star_{(1)} \leq n^\star_{(2)} \leq \cdots \leq n^\star_{(k)}$ in the KKT equations, we have to check whether the obtained $n^\star_{(i)}$'s indeed have this increasing order. If it does, then the $n^\star_{(i)}$ above is the solution we want. Otherwise, there is no solution for this particular equation set. After the solution for each of the $\binom{k}{s}$ KKT equation sets is obtained, the minimum in $\mathcal{F}_u$ with the sample sizes for some DSM diagnoses reaching the upper bounds is the smallest among these solutions. The algorithm to obtain this minimum can be stated in the following result.

Result 4.7. *Suppose there are s DSM diagnoses with sample sizes known to be equal to the corresponding constraints, then order all the k DSM diagnoses such that the s DSM diagnoses would have increasing constraints and the other $k - s$ DSM diagnoses would have increasing relative proportions. There are $\binom{k}{s}$ ways to order the k DSM diagnoses.*

For the rth ordering ($r = 1, 2, \cdots, \binom{k}{s}$), let $\mathcal{S}^{(r)}$ be the index set $\mathcal{S}$ defined in Definition 4.3 and $n^{(r)}_{(i)}$ be the solution obtained from (4.5.5) or (4.5.6), depending on whether $k \in \mathcal{S}^{(r)}$. Let $\mathcal{R} = \{r : n^{(r)}_{(1)} \leq n^{(r)}_{(2)} \leq \cdots \leq n^{(r)}_{(k)}, 1 \leq r \leq \binom{k}{s}\}$, then only for $r \in \mathcal{R}$, we calculate $Var^{(r)}(\hat{\Delta})$ with formula (4.A.3). The minimum in $\mathcal{F}_u$ with some sample sizes equal to the corresponding constraints is $\underset{r \in \mathcal{R}}{\operatorname{argmin}}\ Var^{(r)}(\hat{\Delta})$.

The hypothetical example in Section 4.2.3 is used to illustrate the process stated in Result 4.7.

4.5.4.2 Numerical Example of Minimization in $\mathcal{F}_u$ with some $n_i = N_i$

Recall in Section 4.5.3.3, we have already obtain the global minimum in $\mathcal{F}_0$ with $n = 10$ as $n_1 = 0.550, n_2 = 1.360, n_3 = 4.045$. Because we have $N_1 = 4, N_2 = 3, N_3 = 2$, then it immediately follows that the global minimum is outside $\mathcal{F}_a$ because $n_3 > N_3 = 2$. Now we need to force $n_3 = 2$ and minimize $\text{Var}(\hat{\Delta})$ in $\mathcal{F}_u$ again.

The process of minimization in $\mathcal{F}_u$ with $n_3 = N_3 = 2$ can be shown in Table 4.5. The number in parenthesis is the relative proportion which could identify the specific DSM diagnosis. Now there is only $s = 1$ DSM diagnosis with the sample size known to be equal to the constraint, so there are $\binom{3}{1} = 3$ possible orderings of the DSM diagnoses. For each ordering the solution is obtained and listed in a row. It can be seen in this case for $r = 2$ and $r = 3$ the solution has the increasing order assumed and $\mathcal{R} = \{2, 3\}$. If $\sigma = 1$, then the variance of $\hat{\Delta}$ is 0.620 for $r = 2$ and 0.650 for $r = 3$. Therefore $(n_1 = 0.702, n_2 = 3.649, n_3 = 2)$ is the minimum in $\mathcal{F}_u$ with $n_3 = N_3 = 2$. Note that $n_2 = 3.649 > N_2$ and thus this minimum with $n_3 = N_3$ is again outside $\mathcal{F}_a$. If we want to obtain the optimal sample size in $\mathcal{F}_a$, we need to minimize again in $\mathcal{F}_u$ but with $n_2 = N_2 = 3, n_3 = N_3 = 2$.

Table 4.5: Numerical example for minimization in $\mathcal{F}_u$ with $n_3 = N_3 = 2$

r	$n_{(1)}(\pi_{(1)})$	$n_{(2)}(\pi_{(1)})$	$n_{(3)}(\pi_{(1)})$	$\mathcal{S}$	Var($\hat{\Delta}$)
1	1.727(0.1)	4.273(0.2)	2(0.7)	{3}	×
2	0.702(0.1)	2(0.7)	3.649(0.2)	{2}	**0.620**
3	2(0.7)	2.031(0.1)	2.985(0.2)	{1}	0.650

Based on the argument before and as can be seen from the illustration, there is no one step closed form solution to the minimization of Var($\hat{\Delta}$) in $\mathcal{F}_u$ with some of the sample sizes equal to the corresponding upper bounds. What specific ordering of the DSM diagnoses the final minimum has depends on the specific values of the constraints as well as the distances in the relative proportions between any two of the DSM diagnoses.

4.6 PROPOSED ALGORITHM FOR OPTIMAL DESIGN AND ILLUSTRATION

With our assumptions described previously and methods to minimize Var($\hat{\Delta}$) in $\mathcal{F}_u$ under the different situations, we are now ready to state the proposed algorithm to find the optimal sample size for each DSM diagnosis that minimizes Var($\hat{\Delta}$) in $\mathcal{F}_a$ under the constraints. The hypothetical example in Section 4.2.3 is again used to illustrate the algorithm. An R program is also provided to implement the algorithm in Appendix 4.A.5.

4.6.1 Proposed Algorithm for Optimal Design

The proposed algorithm to obtain the optimal sample size for each DSM diagnosis is stated in Result 4.8.

Result 4.8. *With the parameters defined as in Table 3.1 and this chapter, in a triangular design with k DSM diagnoses and one healthy control group, the optimal sample sizes to estimate the difference in a neurobiological characteristic between the population with dysfunction in the construct of interest and the healthy population can be calculated with the following algorithm:*

0. *Assume $n_1^{(0)} \leq n_2^{(0)} \leq \cdots \leq n_k^{(0)}$, calculate the global minimum $n_i^{(0)}$ in $\mathcal{F}_u$ using (4.5.3); Now start with $j = 0$ and iterate the following two steps:*
1. *Compare $n_i^{(j)}$ with N_i for $1 \leq i \leq k$;*
   a. *if $n_i^{(j)} \leq N_i$ for all i, then $(n_1^{(j)}, n_2^{(j)}, \cdots, n_k^{(j)}) \in \mathcal{F}_a$ and stop. This is the final sample size desired;*
   b. *if $n_i^{(j)} > N_i$ for some index i, then go to step 2.*
2. *Use the method stated in Result 4.7 to minimize Var$(\hat{\Delta})$ in $\mathcal{F}_u$ with $n_i^{(j+1)} = N_i$ for the index i such that $n_i^{(j)} > N_i$. Obtain the minimum $(n_1^{(j+1)}, n_2^{(j+1)}, \cdots, n_k^{(j+1)})$ and go to step 1.*

The algorithm says to obtain the optimal sample sizes, we always start from subset $\mathcal{F}_0$ to obtain the global minimum in $\mathcal{F}_u$ as if there is no constraint on the sample size for each DSM diagnosis. Then the sample size for each DSM diagnosis keeps being updated after comparing it with the corresponding constraint until all the sample sizes are no greater than the corresponding constraints. Note that step 2 might involve multiple minimizations as illustrated in Table 4.5. Another thing that is worth noting is that the values obtained by the proposed algorithm are not necessarily integers. Rounding is needed where appropriate.

4.6.2 Illustration of the Proposed Algorithm

We again use the hypothetical example in Section 4.2.3 to illustrate the algorithm. Because in this example, N_1, N_2, N_3 and n are small enough, all the possible points in $\mathcal{F}_a$ are enumerated and listed in Table 4.7. We compare what is obtained from the proposed algorithm with the sample sizes obtained by enumeration.

Suppose $\sigma = 1$. Table 4.6 below lists each step in determining the sample sizes using the proposed algorithm. Each row is a separate step with step index in the first column.

Because the study needs to be a triangular design, the DSM diagnoses in each row are represented using $D_{(1)}, D_{(2)}, D_{(3)}$. The three terms in each cell are respectively the relative proportion, the constraint of sample size and the sample size computed in that step for each DSM diagnosis. The relative proportion is used to identify the specific DSM diagnosis $D_{(i)}$ refers to.

Table 4.6: Illustration of the proposed algorithm for the example in Section 4.2.3

Step	$(\pi_{(1)}, N_{(1)}, n_{(1)})$	$(\pi_{(2)}, N_{(2)}, n_{(2)})$	$(\pi_{(3)}, N_{(3)}, n_{(3)})$	$\text{Var}(\hat{\Delta})$
0	(0.1,4,0.550)	(0.2,3,1.360)	(0.7,2,4.045)	0.442
1	(0.1,4,0.702)	(0.7,2,2)	(0.2,3,3.649)	0.620
2	(0.1,4,2)	(0.7,2,2)	(0.2,3,3)	0.650

As can be seen in Table 4.6, the final sample sizes determined by the proposed algorithm are 2, 3 and 2 for depression ($\pi_1 = 0.1$), bipolar disorder ($\pi_2 = 0.2$) and schizophrenia ($\pi_3 = 0.7$), respectively. Based on our assumptions, the healthy population has 3 subjects also. The enumeration listed in Table 4.7 shows that there are in total 18 possible points in set $\mathcal{F}_a$ and we calculate $\text{Var}(\hat{\Delta})$ for each point. The point with the smallest variance is the same as what is obtained with the proposed algorithm.

Comparing Table 4.6 and Table 4.7, we can see that our proposed algorithm needs fewer steps. However, the proposed algorithm does not necessarily return a solution with integers because the minimization is not done on an integer set. So at last, we need to decide whether we should round up or down for each DSM diagnosis. Because the total number of subjects are constrained, rounding up in one DSM diagnosis means rounding down in another one. As the number of DSM diagnoses k increases, it becomes more complicated to decide which DSM diagnoses need rounding up and which ones need rounding down.

Table 4.7: Enumeration of sample sizes for the example in Section 4.2.3

n_1	n_2	n_3	$n_{(3)}$	$\text{Var}(\hat{\Delta})$
1	1	1	1	1.540
1	1	2	2	0.818
1	2	1	2	1.180
1	2	2	2	0.777
1	3	1	3	1.060
1	3	2	3	0.657
2	1	1	2	1.238
2	1	2	2	0.797
2	2	1	2	1.097
2	2	2	2	0.770
2	3	1	3	0.977
2	**3**	**2**	**3**	**0.650**
3	1	1	3	1.137
3	1	2	3	0.696
3	2	1	3	0.996
3	2	2	3	0.669
3	3	1	3	0.949
4	1	1	4	1.086

4.A APPENDIX

4.A.1 Derivation of Var($\hat{\Delta}$)

It's obvious that $T'\mathcal{X}'\mathcal{X}T$ is a symmetric matrix and based on the definition of T,

$$T'\mathcal{X}'\mathcal{X}T = \begin{pmatrix} \tau'_{(1)} \\ \tau'_{(2)} \\ \vdots \\ \tau'_{(k)} \end{pmatrix} \mathcal{X}'\mathcal{X} \begin{pmatrix} \tau'_{(1)} & \tau'_{(2)} & \cdots & \tau'_{(k)} \end{pmatrix}$$
$$= \begin{pmatrix} \tau'_{(1)} \\ \tau'_{(2)} \\ \vdots \\ \tau'_{(k)} \end{pmatrix} \begin{pmatrix} l_{(1)} & l_{(2)} & \cdots & l_{(k)}. \end{pmatrix}$$

Using the definition of $\tau_{(i)}$ and $l_{(g)}$, we have $\tau'_{(i)}l_{(g)} = a_{i,g} - a_{i,0}$, so

$$T'\mathcal{X}'\mathcal{X}T = \begin{bmatrix} a_{1,1}-a_{1,0} & a_{1,2}-a_{1,0} & \cdots & a_{1,k}-a_{1,0} \\ a_{2,1}-a_{2,0} & a_{2,2}-a_{2,0} & \cdots & a_{2,k}-a_{2,0} \\ \vdots & \vdots & \vdots & \vdots \\ a_{k,1}-a_{k,0} & a_{k,2}-a_{k,0} & \cdots & a_{k,k}-a_{k,0} \end{bmatrix}.$$

By plugging in the solution of $\tau_{(i)}$ from the proof in Appendix 3.A.1, we get

$$(T'\mathcal{X}'\mathcal{X}T)_{ii} = (1 + \frac{1}{k-i+1})\frac{1}{n_{(i)}} + \sum_{h=i+1}^{k} \frac{1}{(k-h+1)(k-h+2)n_{(h)}} \tag{4.A.1}$$

and for $i < g$,

$$(T'\mathcal{X}'\mathcal{X}T)_{ig} = \frac{1}{(k-g+1)n_{(g)}} + \sum_{h=g+1}^{k} \frac{1}{(k-h+1)(k-h+2)n_{(h)}}. \tag{4.A.2}$$

Because $T'\mathcal{X}'\mathcal{X}T$ is symmetric, $(T'\mathcal{X}'\mathcal{X}T)_{ig} = (T'\mathcal{X}'\mathcal{X}T)_{gi}$ for $i > g$.

$Var(\hat{\Delta})$ is a quadratic form in $\boldsymbol{\pi}$ and by plugging (4.A.1) and (4.A.2) into (4.4.1) we have:

$$
\begin{aligned}
Var(\hat{\Delta}) =& \sigma^2 \boldsymbol{\pi}' T' \mathcal{X}' \mathcal{X} T \boldsymbol{\pi} \\
=& \sigma^2 \sum_{i=1}^{k} \sum_{g=1}^{k} (T' \mathcal{X}' \mathcal{X} T)_{ig} \pi_{(i)} \pi_{(g)} \\
=& \sigma^2 [\sum_{i=1}^{k} (T' \mathcal{X}' \mathcal{X} T)_{ii} \pi_{(i)}^2 + 2 \sum_{i=1}^{k-1} \sum_{g=i+1}^{k} (T' \mathcal{X}' \mathcal{X} T)_{ig} \pi_{(i)} \pi_{(g)}] \\
=& \sigma^2 \sum_{i=1}^{k} [(1 + \frac{1}{k-i+1}) \frac{1}{n_{(i)}} + \sum_{h=i+1}^{k} \frac{1}{(k-h+1)(k-h+2) n_{(h)}}] \pi_{(i)}^2 \\
& + 2\sigma^2 \sum_{i=1}^{k-1} \sum_{g=i+1}^{k} [\frac{1}{(k-g+1) n_{(g)}} + \sum_{h=g+1}^{k} \frac{1}{(k-h+1)(k-h+2) n_{(h)}}] \pi_{(i)} \pi_{(g)} \\
=& \sigma^2 \sum_{i=1}^{k} \frac{[\pi_{(1)} + \pi_{(2)} + \cdots + \pi_{(i-1)} + (k-i+2)\pi_{(i)}]^2}{n_{(i)} (k-i+1)(k-i+2)}.
\end{aligned}
\tag{4.A.3}
$$

4.A.2 Multivariate variance of the mean differences

Suppose there are $M (M \geq 1)$ neurobiological characteristics measured in each subject and we are interested in estimating the differences between the population with dysfunction in the construct of interest and the healthy population in all the characteristics simultaneously. Now the question of interest becomes whether the sample size determination in the multivariate case remains the same as in the univariate case.

Following Appendix 3.A.2, the joint distribution of $(\boldsymbol{y_1}, \boldsymbol{y_2}, \cdots, \boldsymbol{y_M})$ is different from that of $(\boldsymbol{\epsilon_1}, \boldsymbol{\epsilon_2}, \cdots, \boldsymbol{\epsilon_M})$ only in the mean vector. So $\text{Cov}(\boldsymbol{y_1}, \boldsymbol{y_2}, \cdots, \boldsymbol{y_M})$ is the same as $\text{Cov}(\boldsymbol{\epsilon_1}, \boldsymbol{\epsilon_2}, \cdots, \boldsymbol{\epsilon_M})$ in (3.A.3). Therefore the covariance matrix of $\hat{\boldsymbol{\Delta}}$ is

$$
\begin{aligned}
Cov(\hat{\boldsymbol{\Delta}}) = & \begin{bmatrix} \boldsymbol{\pi}'T'\mathcal{X}' & \mathbf{0}' & \cdots & \mathbf{0}' \\ \mathbf{0}' & \boldsymbol{\pi}'T'\mathcal{X}' & \cdots & \mathbf{0}' \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}' & \mathbf{0}' & \cdots & \boldsymbol{\pi}'T'\mathcal{X}' \end{bmatrix} \\
& * \begin{bmatrix} \sigma_{11} I_{n_{(k)}+\sum_{i=1}^{k} n_{(i)}} & \sigma_{12} I_{n_{(k)}+\sum_{i=1}^{k} n_{(i)}} & \cdots & \sigma_{1M} I_{n_{(k)}+\sum_{i=1}^{k} n_{(i)}} \\ \sigma_{21} I_{n_{(k)}+\sum_{i=1}^{k} n_{(i)}} & \sigma_{22} I_{n_{(k)}+\sum_{i=1}^{k} n_{(i)}} & \cdots & \sigma_{2M} I_{n_{(k)}+\sum_{i=1}^{k} n_{(i)}} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{M1} I_{n_{(k)}+\sum_{i=1}^{k} n_{(i)}} & \sigma_{M2} I_{n_{(k)}+\sum_{i=1}^{k} n_{(i)}} & \cdots & \sigma_{MM} I_{n_{(k)}+\sum_{i=1}^{k} n_{(i)}} \end{bmatrix} \\
& * \begin{bmatrix} \boldsymbol{\pi}'T'\mathcal{X}' & \mathbf{0}' & \cdots & \mathbf{0}' \\ \mathbf{0}' & \boldsymbol{\pi}'T'\mathcal{X}' & \cdots & \mathbf{0}' \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}' & \mathbf{0}' & \cdots & \boldsymbol{\pi}'T'\mathcal{X}' \end{bmatrix}' \\
& = \begin{bmatrix} \sigma_{11}\boldsymbol{\pi}'T'\mathcal{X}'\mathcal{X}T\boldsymbol{\pi} & \sigma_{12}\boldsymbol{\pi}'T'\mathcal{X}'\mathcal{X}T\boldsymbol{\pi} & \cdots & \sigma_{1M}\boldsymbol{\pi}'T'\mathcal{X}'\mathcal{X}T\boldsymbol{\pi} \\ \sigma_{21}\boldsymbol{\pi}'T'\mathcal{X}'\mathcal{X}T\boldsymbol{\pi} & \sigma_{22}\boldsymbol{\pi}'T'\mathcal{X}'\mathcal{X}T\boldsymbol{\pi} & \cdots & \sigma_{2M}\boldsymbol{\pi}'T'\mathcal{X}'\mathcal{X}T\boldsymbol{\pi} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{M1}\boldsymbol{\pi}'T'\mathcal{X}'\mathcal{X}T\boldsymbol{\pi} & \sigma_{M2}\boldsymbol{\pi}'T'\mathcal{X}'\mathcal{X}T\boldsymbol{\pi} & \cdots & \sigma_{MM}\boldsymbol{\pi}'T'\mathcal{X}'\mathcal{X}T\boldsymbol{\pi} \end{bmatrix} \\
& = \boldsymbol{\pi}'T'\mathcal{X}'\mathcal{X}T\boldsymbol{\pi} \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1M} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{M1} & \sigma_{M2} & \cdots & \sigma_{MM} \end{bmatrix}.
\end{aligned}
$$

Comparing the above expression to that for the univariate case in (4.4.1), it immediately follows that for both cases, we minimize the same quantity $\boldsymbol{\pi}'T'\mathcal{X}'\mathcal{X}T\boldsymbol{\pi}$. In other words, the optimal sample size determination does not depend on the number of neurobiological characteristics in the study. Therefore, only the univariate case is focused on in the main text.

4.A.3 Illustration of solving equation set (4.5.2) with $k = 2$

When $k = 2$, i.e., there are only two DSM diagnoses and a healthy population, so that equation set (4.5.2) becomes the following:

$$\begin{cases} \frac{\partial L}{\partial n_{(1)}} = -\frac{[3\pi_{(1)}]^2}{n_{(1)}^2 2*3}\sigma^2 + \lambda_0 + \lambda_1 + \eta_1 = 0 \\ \frac{\partial L}{\partial n_{(2)}} = -\frac{[\pi_{(1)}+2\pi_{(2)}]^2}{n_{(2)}^2 1*2}\sigma^2 + 2\lambda_0 - \lambda_1 + \eta_2 = 0 \\ \lambda_0 \geq 0, \lambda_1 \geq 0, \eta_1 \geq 0, \eta_2 \geq 0 \\ n_{(1)} + 2n_{(2)} - n \leq 0 \\ n_{(1)} - n_{(2)} \leq 0 \\ n_{(1)} - N_{(1)} \leq 0 \\ n_{(2)} - N_{(2)} \leq 0 \\ \lambda_0(n_{(1)} + 2n_{(2)} - n) = 0 \\ \lambda_1(n_{(1)} - n_{(2)}) = 0 \\ \eta_1(n_{(1)} - N_{(1)}) = 0 \\ \eta_2(n_{(2)} - N_{(2)}) = 0 \end{cases} . \tag{4.A.4}$$

If we use $n^\star_{(1)}, n^\star_{(2)}, \lambda^\star_0, \lambda^\star_1, \eta^\star_1, \eta^\star_2$ to denote the solution to (4.A.4), then we could have the following 16 scenarios depending on if each one of $\lambda^\star_0, \lambda^\star_1, \eta^\star_1, \eta^\star_2$ is positive or zero.

1. If $\lambda^\star_0 = 0, \lambda^\star_1 = 0, \eta^\star_1 = 0, \eta^\star_2 = 0$

   This scenario does not exist because the term involving $n_{(i)}$ in the first two equations in (4.A.4) cannot be 0.

2. If $\lambda^\star_0 = 0, \lambda^\star_1 = 0, \eta^\star_1 = 0, \eta^\star_2 > 0$

   This scenario does not exist because the term involving $n_{(1)}$ in the first equation in (4.A.4) cannot be 0.

3. If $\lambda^\star_0 = 0, \lambda^\star_1 = 0, \eta^\star_1 > 0, \eta^\star_2 = 0$

   This scenario does not exist because the term involving $n_{(2)}$ in the second equation in (4.A.4) cannot be 0.

4. If $\lambda_0^\star = 0, \lambda_1^\star = 0, \eta_1^\star > 0, \eta_2^\star > 0$

   Based on the first and last two equations in (4.A.4), we have

$$\begin{aligned} n_{(1)}^\star &= N_{(1)}, \\ n_{(2)}^\star &= N_{(2)}, \\ \eta_1^\star &= \frac{[3\pi_{(1)}]^2}{n_{(1)}^{\star 2} 2 * 3}\sigma^2, \\ \eta_2^\star &= \frac{[\pi_{(1)} + 2\pi_{(2)}]^2}{n_{(2)}^{\star 2} 1 * 2}\sigma^2. \end{aligned}$$

   Considering the inequality constraints, we still need to require $N_{(1)} \leq N_{(2)}$ and $N_{(1)} + 2N_{(2)} \leq n$ for the above $n_{(1)}^\star, n_{(2)}^\star, \eta_1^\star, \eta_2^\star$ to be a solution. Therefore, we actually have $n_{(1)}^\star = \min(N_1, N_2)$ and $n_{(2)}^\star = \max(N_1, N_2)$. This means for each DSM diagnosis the actual sample size is equal to the number of subjects available and the budget allows us to process more than what is available. In this case, the definition of $D_{(1)}$ is also clear, which is just the DSM diagnosis with fewer available post-mortem subjects.

5. If $\lambda_0^\star = 0, \lambda_1^\star > 0, \eta_1^\star = 0, \eta_2^\star = 0$

   This scenario does not exist because based on the second equation in (4.A.4), if $\lambda_0^\star = 0, \eta_2^\star = 0$, then $\lambda_1^\star < 0$.

6. If $\lambda_0^\star = 0, \lambda_1^\star > 0, \eta_1^\star = 0, \eta_2^\star > 0$

   When $\lambda_1^\star > 0$, we have $n_{(1)}^\star = n_{(2)}^\star$, which means that the DSM diagnoses and the healthy population have equal sample sizes. If $\eta_2^\star > 0$, then $n_{(1)}^\star = n_{(2)}^\star = N_{(2)} \leq N_{(1)}$. This means

$$\begin{aligned} n_{(1)}^\star = n_{(2)}^\star &= \min(N_1, N_2), \\ \lambda_1^\star &= \frac{[3\pi_{(1)}]^2}{n_{(1)}^{\star 2} 2 * 3}\sigma^2, \\ \eta_2^\star &= \frac{[3\pi_{(1)}]^2}{n_{(1)}^{\star 2} 2 * 3}\sigma^2 + \frac{[\pi_{(1)} + 2\pi_{(2)}]^2}{n_{(2)}^{\star 2} 1 * 2}\sigma^2. \end{aligned}$$

   Considering other inequality constraints, we still need $3 * \min(N_1, N_2) \leq n$ for the above $n_{(1)}^\star, n_{(2)}^\star, \eta_1^\star, \eta_2^\star$ to be a solution. This means the lesser of the number of available subjects for the two DSM diagnoses should not exceed one third of the total number of subjects we can afford. Note in this case that the three groups have the same sample size and it does not matter which π_i is $\pi_{(i)}$.

7. If $\lambda_0^\star = 0, \lambda_1^\star > 0, \eta_1^\star > 0, \eta_2^\star = 0$

   This scenario does not exist because based on the second equation in (4.A.4), if $\lambda_0^\star = 0, \eta_2^\star = 0$, then $\lambda_1^\star < 0$.

8. If $\lambda_0^\star = 0, \lambda_1^\star > 0, \eta_1^\star > 0, \eta_2^\star > 0$

   Based on the last three equations in (4.A.4), we have

$$\begin{aligned} n_{(1)}^\star = n_{(2)}^\star &= N_1 = N_2, \\ \lambda_1^\star + \eta_1^\star &= \frac{[3\pi_{(1)}]^2}{n_{(1)}^{\star 2} 2 * 3}\sigma^2, \\ -\lambda_1^\star + \eta_2^\star &= \frac{[\pi_{(1)} + 2\pi_{(2)}]^2}{n_{(2)}^{\star 2} 1 * 2}\sigma^2. \end{aligned}$$

   There is no unique solution to $\lambda_1^\star, \eta_1^\star$ and $\eta_2^\star$. Also we still need to have $3 * N_1 = 3 * N_2 \leq n$. This means the upper bound for the sample sizes in the DSM diagnoses should be equal and not exceed one third of the total number of subjects we can afford.

9. If $\lambda_0^\star > 0, \lambda_1^\star = 0, \eta_1^\star = 0, \eta_2^\star = 0$

   When $\lambda_0^\star > 0$, we have $n_{(1)}^\star + 2n_{(2)}^\star = n$. With the first two equations in (4.A.4), we have

$$\begin{aligned} n_{(1)}^\star &= \frac{\frac{\sqrt{6}}{2}\pi_{(1)}}{\frac{\sqrt{6}}{2}\pi_{(1)} + \pi_{(1)} + 2\pi_{(2)}} n, \\ n_{(2)}^\star &= \frac{\frac{\pi_{(1)} + 2\pi_{(2)}}{2}}{\frac{\sqrt{6}}{2}\pi_{(1)} + \pi_{(1)} + 2\pi_{(2)}} n, \\ \lambda_0^\star &= \frac{\sigma^2}{n}(\frac{\sqrt{6}}{2}\pi_{(1)} + \pi_{(1)} + 2\pi_{(2)}). \end{aligned}$$

   Again, we have to require $n_{(1)}^\star \leq n_{(2)}^\star, n_{(1)}^\star \leq N_{(1)}$ and $n_{(2)}^\star \leq N_{(2)}$. Plugging in the solution of $n_{(1)}^\star$ and $n_{(2)}^\star$ from above into the equation $\pi_{(1)} + \pi_{(2)} = 1$, this means $\pi_{(1)} \leq \frac{2}{\sqrt{6}+1}$. In other words, the relative proportions of the DSM diagnoses also need to satisfy some condition for the solution to exist in this scenario. Note in this case that we have obtained a closed form solution of $n_{(i)}^\star$, however, the solution has $\pi_{(i)}$ in it. We need to know which one of the π_i's the $\pi_{(i)}$ corresponds to so that the variance achieves the minimum. In this example, there are $k! = 2$ possible ways to permute π_1 and π_2, so we have to check which permutation yields a smaller variance.

10. If $\lambda_0^\star > 0, \lambda_1^\star = 0, \eta_1^\star = 0, \eta_2^\star > 0$

    When $\lambda_0^\star > 0$, we have $n_{(1)}^\star + 2n_{(2)}^\star = n$. And $\eta_2^\star > 0$ means $n_{(2)}^\star = N_{(2)}$. With other equations in (4.A.4), we have

$$\begin{aligned}
n_{(1)}^\star &= n - 2N_{(2)}, \\
n_{(2)}^\star &= N_{(2)}, \\
\lambda_0^\star &= \frac{[3\pi_{(1)}]^2}{n_{(1)}^{\star 2} 2 * 3}\sigma^2, \\
\eta_2^\star &= \frac{[\pi_{(1)} + 2\pi_{(2)}]^2}{n_{(2)}^{\star 2} 1 * 2}\sigma^2 - 2\frac{[3\pi_{(1)}]^2}{n_{(1)}^{\star 2} 2 * 3}\sigma^2.
\end{aligned}$$

    Considering the inequalities, we need

$$\begin{gathered}
n - 2N_{(2)} \leq N_{(2)}, \\
n - 2N_{(2)} \leq N_{(1)}, \\
\frac{[\pi_{(1)} + 2\pi_{(2)}]^2}{n_{(2)}^{\star 2} 1 * 2}\sigma^2 - 2\frac{[3\pi_{(1)}]^2}{n_{(1)}^{\star 2} 2 * 3}\sigma^2 > 0.
\end{gathered}$$

    which leads to

$$\begin{aligned}
N_{(1)} + 2N_{(2)} &\geq n, \\
\frac{n}{3} \leq N_{(2)} &< \frac{1 + \pi_{(2)}}{2 + \sqrt{6} + (2 - \sqrt{6})\pi_{(2)}} n.
\end{aligned}$$

    Under this scenario, to get a solution, $N_{(1)}, N_{(2)}$ and n have to satisfy some conditions in terms of their relationship. Also the two permutations of π_1, π_2 need to be checked and see which one produces a smaller variance.

11. If $\lambda_0^\star > 0, \lambda_1^\star = 0, \eta_1^\star > 0, \eta_2^\star = 0$

    Similar as the previous scenario, we have

$$\begin{aligned}
n_{(1)}^\star &= N_{(1)}, \\
n_{(2)}^\star &= \frac{n - N_{(1)}}{2}, \\
\lambda_0^\star &= \frac{1}{2}\frac{[\pi_{(1)} + 2\pi_{(2)}]^2}{n_{(2)}^{\star 2} 1 * 2}\sigma^2, \\
\eta_1^\star &= \frac{[3\pi_{(1)}]^2}{n_{(1)}^{\star 2} 2 * 3}\sigma^2 - \frac{1}{2}\frac{[\pi_{(1)} + 2\pi_{(2)}]^2}{n_{(2)}^{\star 2} 1 * 2}\sigma^2.
\end{aligned}$$

Considering the inequalities, we need to have

$$\begin{aligned} N_{(1)} + 2N_{(2)} &\geq n, \\ N_{(1)} &\leq \frac{n}{3}, \\ N_{(1)} &< \frac{\sqrt{6}\pi_{(1)}}{4 + (\sqrt{6} - 2)\pi_{(1)}} n. \end{aligned}$$

Again the relationship between $N_{(1)}, N_{(2)}$ and n need to satisfy some conditions and the two permutations of π_1, π_2 need to be checked.

12. If $\lambda_0^\star > 0, \lambda_1^\star = 0, \eta_1^\star > 0, \eta_2^\star > 0$

    It is easy to see in this case that

$$\begin{aligned} n_{(1)}^\star &= N_{(1)}, \\ n_{(2)}^\star &= N_{(2)}, \\ N_{(1)} + 2N_{(2)} &= n, \\ \lambda_0^\star + \eta_1^\star &= \frac{[3\pi_{(1)}]^2}{n_{(1)}^{\star 2} 2 * 3} \sigma^2, \\ 2\lambda_0^\star + \eta_2^\star &= \frac{[\pi_{(1)} + 2\pi_{(2)}]^2}{n_{(2)}^{\star 2} 1 * 2} \sigma^2. \end{aligned}$$

Even though $n_{(1)}^\star$ and $n_{(2)}^\star$ can be obtained, there are two equations about the three unknown $\lambda_0^\star, \eta_1^\star, \eta_2^\star$. So there is no unique solution to the above equations. And in this case because $n_{(1)}^\star \leq n_{(2)}^\star$, it must be that $N_{(1)} \leq N_{(2)}$. So the solution $n_{(1)}^\star$ and $n_{(2)}^\star$ are actually

$$\begin{aligned} n_{(1)}^\star &= \min(N_1, N_2), \\ n_{(2)}^\star &= \max(N_1, N_2). \end{aligned}$$

and we need to have $N_1 + N_2 + \max(N_1, N_2) = n$ for a solution to exist in this case.

13. If $\lambda_0^\star > 0, \lambda_1^\star > 0, \eta_1^\star = 0, \eta_2^\star = 0$

    With $\lambda_0^\star > 0$ and $\lambda_1^\star > 0$, we have $n_{(1)}^\star = n_{(2)}^\star = \frac{n}{3}$. Plugging them into the first two equations in (4.A.4), we have

$$\lambda_0^\star = 3\sigma^2(\frac{[3\pi_{(1)}]^2}{n^2 2 * 3} + \frac{[\pi_{(1)} + 2\pi_{(2)}]^2}{n^2 1 * 2}),$$
$$\lambda_1^\star = 3\sigma^2(\frac{2 * [3\pi_{(1)}]^2}{n^2 2 * 3} - \frac{[\pi_{(1)} + 2\pi_{(2)}]^2}{n^2 1 * 2}).$$

    Considering the inequalities, we still need to have

$$n \leq 3N_1,$$
$$n \leq 3N_2,$$
$$\pi_{(1)} > \frac{2}{\sqrt{6}+1}.$$

    In other words, in this case, the upper bound for the sample size in the DSM diagnoses should at least be one third of the total number of subjects we can afford. And the relative proportion also need to satisfy some condition.

14. If $\lambda_0^\star > 0, \lambda_1^\star > 0, \eta_1^\star = 0, \eta_2^\star > 0$

    As above, we have $n_{(1)}^\star = n_{(2)}^\star = \frac{n}{3}$. Additionally with $\eta_2^\star > 0$, we have $n_{(2)}^\star = N_{(2)}$. So

$$N_{(1)} \geq N_{(2)} = \frac{n}{3},$$
$$\lambda_0^\star + \lambda_1^\star = \frac{9 * [3\pi_{(1)}]^2}{n^2 2 * 3}\sigma^2,$$
$$2\lambda_0^\star - \lambda_1^\star + \eta_2^\star = \frac{9 * [\pi_{(1)} + 2\pi_{(2)}]^2}{n^2 1 * 2}\sigma^2.$$

    Equivalently, $\max(N_1, N_2) \geq \min(N_1, N_2) = \frac{n}{3}$. Again, there are more unknowns than the number of equations, and thus there is no unique solution for $\lambda_0^\star, \lambda_1^\star$ and $\eta_2^\star$.

15. If $\lambda_0^\star > 0, \lambda_1^\star > 0, \eta_1^\star > 0, \eta_2^\star = 0$

    As above, we have $n_{(1)}^\star = n_{(2)}^\star = \frac{n}{3}$. Additionally with $\eta_1^\star > 0$, we have $n_{(1)}^\star = N_{(1)}$. So

$$\begin{aligned} N_{(2)} \geq N_{(1)} &= \frac{n}{3}, \\ \lambda_0^\star + \lambda_1^\star + \eta_1^\star &= \frac{9 * [3\pi_{(1)}]^2}{n^2 2 * 3}\sigma^2, \\ 2\lambda_0^\star - \lambda_1^\star &= \frac{9 * [\pi_{(1)} + 2\pi_{(2)}]^2}{n^2 1 * 2}\sigma^2. \end{aligned}$$

    Equivalently, $\max(N_1, N_2) \geq \min(N_1, N_2) = \frac{n}{3}$. Again, there are more unknowns than the number of equations, and thus there is no unique solution for $\lambda_0^\star, \lambda_1^\star$ and $\eta_1^\star$.

16. If $\lambda_0^\star > 0, \lambda_1^\star > 0, \eta_1^\star > 0, \eta_2^\star > 0$

    With all of $\lambda_0^\star, \lambda_1^\star, \eta_1^\star, \eta_2^\star$ greater than 0, we have

$$\begin{aligned} n_{(1)}^\star = n_{(2)}^\star = N_{(1)} = N_{(2)} &= \frac{n}{3}, \\ \lambda_0^\star + \lambda_1^\star + \eta_1^\star &= \frac{9 * [3\pi_{(1)}]^2}{n^2 2 * 3}\sigma^2, \\ 2\lambda_0^\star - \lambda_1^\star + \eta_2^\star &= \frac{9 * [\pi_{(1)} + 2\pi_{(2)}]^2}{n^2 1 * 2}\sigma^2. \end{aligned}$$

    Again, there are two equations with four unknowns, so no unique solution of $\lambda_0^\star, \lambda_1^\star, \eta_1^\star, \eta_2^\star$ exists. Also the upper bounds of the sample size in both DSM diagnoses have to be exactly one third of the total number of subjects we can afford.

The above example with $k = 2$ illustrates the difficulty in minimization of $\text{Var}\hat{\Delta}$ in $\mathcal{F}_a$ using the KKT conditions even if the order of $(n_1, n_2, \cdots, n_k)$ is known. As can be seen, each of the 16 scenarios require the N_1, N_2 and n to satisfy some conditions to have a solution. Because we don't know the relationship between N_1, N_2 and n as well as the order of N_1 and N_2, it is hard to know which of the 16 scenarios above would give us a solution.

4.A.4 Proof of Result 4.5

According to the prerequisites stated in Result 4.5 Since $n_{(1)}, \cdots, n_{(k)}$ are given, we have the same set of order statistics before and after the pairwise switch. Let V and V^* denote the variance of $\hat{\Delta}$, $\pi_{(i)}$ and $\pi^*_{(i)}$ be the relative proportions of $D_{(i)}$ before and after the switch, respectively. Using the variance formula in (4.A.3), V and V^* are:

$$\begin{aligned} V =& \sigma^2 \sum_{i=1}^{k} \frac{[\pi_{(1)} + \pi_{(2)} + \cdots + \pi_{(i-1)} + (k-i+2)\pi_{(i)}]^2}{n_{(i)}(k-i+1)(k-i+2)}, \\ V^* =& \sigma^2 \sum_{i=1}^{k} \frac{[\pi^*_{(1)} + \pi^*_{(2)} + \cdots + \pi^*_{(i-1)} + (k-i+2)\pi^*_{(i)}]^2}{n_{(i)}(k-i+1)(k-i+2)}. \end{aligned}$$

The switch can be illustrated as:

$$\begin{array}{ccccccc} \pi_{(1)} & \cdots & \pi_{(p)} & \cdots & \pi_{(q)} & \cdots & \pi_{(k)} \\ D_{(1)} & \cdots & D_{(p)} & \cdots & D_{(q)} & \cdots & D_{(k)} \\ n_{(1)} & \cdots & n_{(p)} & \cdots & n_{(q)} & \cdots & n_{(k)} \end{array} \xrightarrow{\text{switch}} \begin{array}{ccccccc} \pi_{(1)} & \cdots & \pi_{(q)} & \cdots & \pi_{(p)} & \cdots & \pi_{(k)} \\ D_{(1)} & \cdots & D_{(p)} & \cdots & D_{(q)} & \cdots & D_{(k)} \\ n_{(1)} & \cdots & n_{(p)} & \cdots & n_{(q)} & \cdots & n_{(k)} \end{array}$$

It's easy to see that only the values of $\pi_{(p)}$ and $\pi_{(q)}$ are switched, so the relationship between $\pi^*_{(i)}$ and $\pi_{(i)}$ is

$$\pi^*_{(p)} = \pi_{(q)}, \pi^*_{(q)} = \pi_{(p)} \text{ and } \pi^*_{(i)} = \pi_{(i)} \text{ for } i \neq p, q. \tag{4.A.5}$$

Therefore

$$\begin{aligned} V - V^* =& \sigma^2 \sum_{i=1}^{k} \frac{[\pi_{(1)} + \pi_{(2)} + \cdots + \pi_{(i-1)} + (k-i+2)\pi_{(i)}]^2}{n_{(i)}(k-i+1)(k-i+2)} \\ & - \sigma^2 \sum_{i=1}^{k} \frac{[\pi^*_{(1)} + \pi^*_{(2)} + \cdots + \pi^*_{(i-1)} + (k-i+2)\pi^*_{(i)}]^2}{n_{(i)}(k-i+1)(k-i+2)} \\ =& \sigma^2 \sum_{i=1}^{k} \frac{\mathrm{S}_\pi(i) * \mathrm{D}_\pi(i)}{n_{(i)}(k-i+1)(k-i+2)}, \end{aligned}$$

where

$$\begin{aligned} \mathrm{S}_\pi(i) =& \pi_{(1)} + \cdots + \pi_{(i-1)} + (k-i+2)\pi_{(i)} + \pi^*_{(1)} + \pi^*_{(2)} + \cdots + \pi^*_{(i-1)} + (k-i+2)\pi^*_{(i)} \\ \mathrm{D}_\pi(i) =& \pi_{(1)} + \cdots + \pi_{(i-1)} + (k-i+2)\pi_{(i)} - \pi^*_{(1)} - \pi^*_{(2)} - \cdots - \pi^*_{(i-1)} - (k-i+2)\pi^*_{(i)}. \end{aligned}$$

Plugging in the relationship between $\pi^*_{(i)}$ and $\pi_{(i)}$ in (4.A.5), we have

$$\mathrm{S}_\pi(i) = \begin{cases} 2(\pi_{(1)} + \cdots + \pi_{(p-1)}) + (k-p+2)(\pi_{(p)} + \pi_{(q)}) & i = p \\ 2(\pi_{(1)} + \cdots + \pi_{(p-1)}) + (\pi_{(p)} + \pi_{(q)}) & \\ +2(\pi_{(p+1)} + \cdots + \pi_{(i-1)}) + 2(k-i+2)\pi_{(i)} & p+1 \leq i \leq q-1 \\ 2(\pi_{(1)} + \cdots + \pi_{(p-1)}) + (\pi_{(p)} + \pi_{(q)}) & \\ +2(\pi_{(p+1)} + \cdots + \pi_{(q-1)}) + (k-q+2)(\pi_{(p)} + \pi_{(q)}) & i = q \end{cases}$$

and

$$\mathrm{D}_\pi(i) = \begin{cases} 0 & i < p \\ (k-p+2)(\pi_{(p)} - \pi_{(q)}) & i = p \\ \pi_{(p)} - \pi_{(q)} & p+q \leq i \leq q-1 \\ -(k-q+1)(\pi_{(p)} - \pi_{(q)}) & i = q \\ 0 & i > q \end{cases}.$$

Note that $\mathrm{S}_\pi(i) > 0$ for all i and we don't need to compute $\mathrm{S}_\pi(i)$ for $i < p$ or $i > q$ because $\mathrm{D}_\pi(i)$ is 0 so their product $\mathrm{S}_\pi(i) * \mathrm{D}_\pi(i)$ is 0 when $i < p$ or $i > q$. Now $V - V^*$ becomes

$$\begin{aligned} V - V^* =& \sigma^2 \sum_{i=1}^{k} \frac{\mathrm{S}_\pi(i) * \mathrm{D}_\pi(i)}{n_{(i)}(k-i+1)(k-i+2)} \\ =& \sigma^2 \sum_{i=p}^{q} \frac{\mathrm{S}_\pi(i) * \mathrm{D}_\pi(i)}{n_{(i)}(k-i+1)(k-i+2)}. \end{aligned} \tag{4.A.6}$$

We next show that $\sum_{i=p}^{q} \frac{\mathrm{S}_\pi(i) * \mathrm{D}_\pi(i)}{(k-i+1)(k-i+2)} = 0$ and $V - V^* \leq 0$. Plugging in the expression of $\mathrm{S}_\pi(i)$ and $\mathrm{D}_\pi(i)$, we have

$$\sum_{i=p}^{q} \frac{\mathrm{S}_\pi(i) * \mathrm{D}_\pi(i)}{(k-i+1)(k-i+2)} = (\pi_{(p)} - \pi_{(q)})*$$

$$\Bigg(\frac{2(\pi_{(1)} + \cdots + \pi_{(p-1)}) + (k-p+2)(\pi_{(p)} + \pi_{(q)})}{k-p+1}$$

$$+ \sum_{i=p+1}^{q-1} \frac{2(\pi_{(1)} + \cdots + \pi_{(p-1)}) + (\pi_{(p)} + \pi_{(q)}) + 2(\pi_{(p+1)} + \cdots + \pi_{(i-1)}) + 2(k-i+2)\pi_{(i)}}{(k-i+1)(k-i+2)}$$

$$- \frac{2(\pi_{(1)} + \cdots + \pi_{(p-1)}) + (\pi_{(p)} + \pi_{(q)}) + 2(\pi_{(p+1)} + \cdots + \pi_{(q-1)}) + (k-q+2)(\pi_{(p)} + \pi_{(q)})}{k-q+2} \Bigg)$$

$$= (\pi_{(p)} - \pi_{(q)})*$$

$$\Bigg([\frac{2}{k-p+1} + \sum_{i=p+1}^{q-1} \frac{2}{(k-i+1)(k-i+2)} - \frac{2}{k-q+2}](\pi_{(1)} + \cdots + \pi_{(p-1)})$$

$$+ [\frac{k-p+2}{k-p+1} + \sum_{i=p+1}^{q-1} \frac{1}{(k-i+1)(k-i+2)} - \frac{k-q+3}{k-q+2}](\pi_{(p)} + \pi_{(q)})$$

$$+ \sum_{h=p+1}^{q-2} [\frac{2}{k-h+1} + \sum_{i=h+1}^{q-1} \frac{2}{(k-i+1)(k-i+2)} - \frac{2}{k-q+2}]\pi_{(h)}$$

$$+ [\frac{2}{k-(q-1)+1} - \frac{2}{k-q+2}]\pi_{(q-1)} \Bigg)$$

$$= (\pi_{(p)} - \pi_{(q)})*$$

$$\Bigg([\frac{2}{k-p+1} + \sum_{i=p+1}^{q-1} (\frac{2}{k-i+1} - \frac{2}{k-i+2}) - \frac{2}{k-q+2}](\pi_{(1)} + \cdots + \pi_{(p-1)})$$

$$+ [\frac{k-p+2}{k-p+1} + \sum_{i=p+1}^{q-1} (\frac{1}{k-i+1} - \frac{1}{k-i+2}) - \frac{k-q+3}{k-q+2}](\pi_{(p)} + \pi_{(q)})$$

$$+ \sum_{h=p+1}^{q-2} [\frac{2}{k-h+1} + \sum_{i=h+1}^{q-1} (\frac{2}{k-i+1} - \frac{2}{k-i+2}) - \frac{2}{k-q+2}]\pi_{(h)}$$

$$+ [\frac{2}{k-(q-1)+1} - \frac{2}{k-q+2}]\pi_{(q-1)} \Bigg)$$

$$= (\pi_{(p)} - \pi_{(q)}) * \Bigg(0 * (\pi_{(1)} + \cdots + \pi_{(p-1)}) + 0 * (\pi_{(p)} + \pi_{(q)}) + \sum_{h=p+1}^{q-2} 0 * \pi_{(h)} + 0 * \pi_{(q-1)} \Bigg)$$

$$= 0.$$

Dividing the sum into two parts, we have

$$\sum_{i=p}^{q} \frac{\mathrm{S}_\pi(i) * \mathrm{D}_\pi(i)}{(k-i+1)(k-i+2)} = \sum_{i=p}^{q-1} \frac{\mathrm{S}_\pi(i) * \mathrm{D}_\pi(i)}{(k-i+1)(k-i+2)} + \frac{\mathrm{S}_\pi(q) * \mathrm{D}_\pi(q)}{(k-q+1)(k-q+2)} = 0,$$

so

$$\sum_{i=p}^{q-1} \frac{\mathrm{S}_\pi(i) * \mathrm{D}_\pi(i)}{(k-i+1)(k-i+2)} = -\frac{\mathrm{S}_\pi(q) * \mathrm{D}_\pi(q)}{(k-q+1)(k-q+2)}. \tag{4.A.7}$$

Now because $p < q$, by the definition of order statistics, $n_{(p)} \leq n_{(p+1)} \leq \cdots \leq n_{(q)}$, we have

$$\frac{1}{n_{(p)}} \geq \frac{1}{n_{(p+1)}} \geq \cdots \geq \frac{1}{n_{(q)}},$$

so for $p \leq i \leq q-1$,

$$\frac{1}{n_{(i)}} - \frac{1}{n_{(q)}} \geq 0$$

$$\mathrm{D}_\pi(i) < 0 \text{ because } \pi_{(p)} < \pi_{(q)}.$$

Since $\mathrm{S}_\pi(i) > 0$, we have

$$(\frac{1}{n_{(i)}} - \frac{1}{n_{(q)}})\frac{\mathrm{S}_\pi(i) * \mathrm{D}_\pi(i)}{(k-i+1)(k-i+2)} \leq 0, \quad p \leq i \leq q-1.$$

The equality holds when $n_{(i)} = n_{(q)}, p \leq i \leq q-1$. Summing over i and using (4.A.6) and (4.A.7), we have

$$\begin{aligned}
0 &\geq \sum_{i=p}^{q-1}(\frac{1}{n_{(i)}} - \frac{1}{n_{(q)}})\frac{\mathrm{S}_\pi(i) * \mathrm{D}_\pi(i)}{(k-i+1)(k-i+2)} \\
&= \sum_{i=p}^{q-1} \frac{\mathrm{S}_\pi(i) * \mathrm{D}_\pi(i)}{n_{(i)}(k-i+1)(k-i+2)} - \frac{1}{n_{(q)}} \sum_{i=p}^{q-1} \frac{\mathrm{S}_\pi(i) * \mathrm{D}_\pi(i)}{(k-i+1)(k-i+2)} \\
&= \sum_{i=p}^{q-1} \frac{\mathrm{S}_\pi(i) * \mathrm{D}_\pi(i)}{n_{(i)}(k-i+1)(k-i+2)} + \frac{1}{n_{(q)}} \frac{\mathrm{S}_\pi(q) * \mathrm{D}_\pi(q)}{(k-q+1)(k-q+2)} \\
&= \sum_{i=p}^{q} \frac{\mathrm{S}_\pi(i) * \mathrm{D}_\pi(i)}{n_{(i)}(k-i+1)(k-i+2)} \\
&= V - V^*.
\end{aligned}$$

The equality holds when $n_{(p)} = n_{(p+1)} = \cdots = n_{(q)}$.

4.A.5 R program to implement the proposed algorithm

```
#This program is to implement the proposed algorithm and obtain the
 optimal sample sizes;
#function samplesizeinF0 calculates the samplesize in set F0 with given
 pi and n, and returns the sample sizes in F0 for D1,D2,...,Dk;
#function samplesizeonboundary calculates the sample sizes with some n=N;
#pivec is the vector for relative proportions and sorted increasingly;
#inputn is the obtained sample sizes (n1,n2,...,nk) from the last step;
#Nvec is (N1,N2,...,Nk);

library(gtools)
samplesizeinF0<-function(pivec,ntotal){
  k<-length(pivec)
  allovec<-rep(0,k)
  for (i in seq(1:k-1)){
    allovec[i]<-(sum(pivec[0:(i-1)])+(k-i+2)*pivec[i])/
      sqrt((k-i+1)*(k-i+2))
  }
  allovec[k]<-1+pivec[k]
  partitionvec<-allovec/sum(allovec)
  nstarvec<-c(rep(ntotal,k-1),ntotal/2)*partitionvec
  return(nstarvec)
}

samplesizeonboundary<-function(inputn,pivec,Nvec,ntotal){
    k<-length(pivec)
    boundindexset<-which(inputn>=Nvec)
    boundN_ordered<-sort(Nvec[boundindexset])
    boundpi_orderedbyN<-pivec[boundindexset]
```

```
                    [rank(Nvec[boundindexset])]
s<-length(boundindexset)
Comb<-combinations(n=k,r=s)
R<-nrow(Comb)
piMatrix=alloMatrix<-matrix(rep(0,R*k),nrow=R)
varpiMatrix=outputnMatrix<-matrix(rep(0,R*k),nrow=R)
nMatrix<-matrix(rep(0,R*(k+1)),nrow=R)
for (r in 1:R){
  piMatrix[r,Comb[r,]]<-boundpi_orderedbyN
  piMatrix[r,-Comb[r,]]<-pivec[-boundindexset]
  nMatrix[r,Comb[r,]]<-boundN_ordered
  nMatrix[r,k+1]<-ntotal-sum(boundN_ordered)-nMatrix[r,k]
}
for (j in 1:k){
  if (j==1){
  varpiMatrix[,j]=(rep(0,R)+(k-j+2)*piMatrix[,j])
                /sqrt((k-j+1)*(k-j+2))
  alloMatrix[,j]=(nMatrix[,j]==0)*1*varpiMatrix[,j]
  } else if ((k>2)&&(j==2)) {
  varpiMatrix[,j]=(piMatrix[,j-1]+(k-j+2)*piMatrix[,j])
                /sqrt((k-j+1)*(k-j+2))
  alloMatrix[,j]=(nMatrix[,j]==0)*1*varpiMatrix[,j]
  } else if (j==k){
  varpiMatrix[,j]=(1+piMatrix[,j])/sqrt(2)
  alloMatrix[,j]=(nMatrix[,j]==0)*1*(1+piMatrix[,j])/2
  } else {
  varpiMatrix[,j]=(apply(piMatrix[,seq(1,j-1)],1,sum)+
                (k-j+2)*piMatrix[,j])/sqrt((k-j+1)*(k-j+2))
  alloMatrix[,j]<-(nMatrix[,j]==0)*1*varpiMatrix[,j]
  }
```

```
    }
    outputnMatrix<-nMatrix[,1:k]+alloMatrix*nMatrix[,k+1]
                   /(apply(alloMatrix,1,sum)+alloMatrix[,k])
    Var<-apply(varpiMatrix^2/outputnMatrix,1,sum)
         *apply(outputnMatrix,1,function(x) prod(diff(x)>=0))
    rselect<-which(Var==min(Var[Var>0]))
    outputn<-outputnMatrix[rselect,][rank(piMatrix[rselect,])]
    return(outputn)
}

optimalsamplesize<-function(pi,N,n){
    n0<-samplesizeinF0(pivec=pi,ntotal=n)
    while(prod(n0<=N)==0){
      n0<-samplesizeonboundary(inputn=n0,pivec=pi,Nvec=N,ntotal=n)
    }
    return(n0)
}
```

5.0 CONCLUSIONS AND FUTURE WORK

5.1 CONCLUSIONS

In this dissertation, we develop some statistical methodologies to utilize the existing post-mortem tissue databases to facilitate mental health research under the RDoC framework. Because an RDoC study would focus on a particular psychiatric construct rather than any specific DSM diagnosis, we approach the population with dysfunction in a construct of interest by suitably considering all the DSM diagnoses related to this construct and their relative proportions within the construct.

We first propose a method to identify the neurobiological characteristics that are significantly associated with a construct of interest. Our method is to first apply the Laska's Min test on each neurobiological characteristic over all the involved DSM diagnoses and then to adjust for multiplicity with the BH procedure to protect the FDR. We show through simulations that when the neurobiological characteristics within each subject are positively correlated and each DSM diagnosis has its own healthy control group, the FDR is controlled at the desired level by our method. We successfully applied our method to a post-mortem tissue study about schizophrenia and schizoaffective disorder and identified two neurobiological characteristics among the 26 examined ones.

After identifying these significant neurobiological characteristics, we compare their means and quantiles between the population with dysfunction and the healthy population. These comparisons inform researchers how the two populations differ from each other. The findings in the comparisons can be used in various ways for later studies. For example, they can be used to define the normal range of the measures in the neurobiological characteristics. Any subjects outside the normal range for a neurobiological characteristic can be enrolled in later research to study why abnormality in this neurobiological characteristic might lead to the dysfunction of the construct of interest. We propose the triangular design to adjust for the matching covariates and batch effect in tissue processing for these comparisons as well as to be more efficient in tissue processing. In the comparison through the means of the neurobiological characteristics, we provide the formula to estimate the mean difference between the population with dysfunction and the healthy population. In the comparison of quantiles, we propose a heuristic approach to adjust for the disproportionate sample sizes in the estimation of the difference in the quantile of the neurobiological characteristic between the two populations. We show through simulations that our approach estimates the quantiles quite well.

At last, we develop an algorithm to determine the optimal sample size for each DSM diagnosis in the triangular design which gives minimum variance of the estimator of the mean difference between the two populations. The sample sizes are determined under two constraints: the number of available subjects for each DSM diagnosis from the post-mortem tissue databases and the total number of subjects the budget allows. We show by comparing to a simple enumeration method that our algorithm can indeed lead to the correct optimal sample sizes.

5.2 FUTURE WORK

Our current research follows the RDoC spirit by focusing on a particular construct and proposing statistical methods to identify and study the neurobiological characteristics that are significantly associated with this construct. More specifically our current research investigates what makes the DSM diagnoses similar to each other in symptoms and what makes the population with dysfunction in a construct of interest different from the healthy population in symptoms. In our research we divide the entire general population into two parts, one with dysfunction in the construct of interest and the other without. And the differences in the neurobiological characteristics between the population with dysfunction and the healthy population have to be significant enough to be identified by our proposed method and then compared later on. In other words, the assumption of "common significance across DSM diagnoses" by requiring the mean difference in the neurobiological characteristics between each DSM diagnosis and the healthy population to be significant might be too strong. Under this assumption, the current research is unable to study those neurobiological characteristics without significant mean difference in all the DSM diagnoses from the healthy population.

In the future, we would like to approach the RDoC spirit from a different perspective. It could be that some neurobiological characteristics are not significant enough to be identified by our proposed method because only a portion of the subjects in each DSM diagnosis are significantly different from the healthy population. The other portion of subjects in each DSM diagnosis are close enough to the healthy population. This portion of subjects close to the healthy population are dragging the entire DSM diagnosis toward the healthy population and thus make the neurobiological characteristics unable to be identified by the proposed method. In other words, there could be several underlying clusters in terms of the distribution of neurobiological characteristics over the general population, and each DSM diagnosis and the healthy population has a different combination of these underlying clusters. Through identifying these underlying clusters in the neurobiological measures, we can divide the general population into groups based on neurobiological information and see what results these differences among clusters could lead to. These clusters would provide an explanation of the heterogeneity within each DSM diagnosis and the overlaps among different DSM diagnoses.

In order to identify the clusters, we plan to apply a clustering analysis to the neurobiological measures of the subjects from the post-mortem tissue databases. More than one neurobiological characteristics can be used to define these clusters. Again suppose we have k DSM diagnoses and we use D_i to denote the ith DSM diagnosis with relative proportion π_i among the population with dysfunction, where $\pi_1 < \pi_2 < \cdots < \pi_k$ and $\sum_{i=1}^{k} \pi_i = 1$. Suppose there are n_i subjects from D_i and n_0 subjects from the healthy population. Let $\mathbf{Y}_{ij} = (Y_{ij1}, Y_{ij2}, \cdots, Y_{ijM})(i = 0, 1, \cdots, k; j = 1, 2, \cdots, n_i)$ be the vector of measurements of the M neurobiological characteristics in subject j from D_i or the healthy population. If $i = 0$ it is from the healthy population and if $i \neq 0$ it is from D_i.

Suppose there are G underlying clusters in the distribution of neurobiological characteristics and we use $C_g, g = 1, 2, \cdots, G$ to denote the distribution of the neurobiological

Table 5.1: Clustering Analysis of the Neurobiological Characteristics

	D_1	D_2	D_3	$\cdots$	D_k	Healthy Population
	$\pi_1(1-\Pi_0)$	$\pi_2(1-\Pi_0)$	$\pi_3(1-\Pi_0)$	$\cdots$	$\pi_k(1-\Pi_0)$	Π_0
C_1	ω_{11}	ω_{21}	ω_{31}	$\cdots$	ω_{k1}	ω_{01}
C_2	ω_{12}	ω_{22}	ω_{32}	$\cdots$	ω_{k2}	ω_{02}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
C_G	ω_{1G}	ω_{2G}	ω_{3G}	$\cdots$	ω_{kG}	ω_{0G}

characteristics in each cluster. Note that C_g should be an M dimensional distribution. Then the distribution of $\mathbf{Y}_{ij}$ can be thought of as a mixture of the G underlying clusters. That is,

$$\mathbf{Y}_{ij} \sim \sum_{g=1}^{G} \omega_{ig} C_g, \quad i = 0, 1, \cdots, k; j = 1, \cdots, n_i.$$

Here ω_{ig} is the probability of an observation $\mathbf{Y}_{ij}$ from cluster C_g. Obviously for $i = 0, 1, \cdots, k, \sum_{g=1}^{G} \omega_{ig} = 1$. The clustering analysis of the neurobiological characteristics can be laid out as in Table 5.1.

If among the general population, Π_0 is the proportion of healthy subjects, then each D_i accounts for $\pi_i(1-\Pi_0)$. As a result, $\omega_{0g}\Pi_0 + \sum_{i=1}^{k} \omega_{ig}\pi_i(1-\Pi_0)$ of the general population has the distribution C_g. The goal of the clustering analysis is to estimate ω_{ig} and C_g based on the sample we have from the post-mortem tissue databases. Again we need to address the issues of covariates adjustment and disproportionate sample sizes in the clustering analysis as we do in the comparison through quantiles in Section 3.3. For example, if there are only two DSM diagnoses and $\Pi_0 = 0.9, \pi_1 = 0.3, \pi_2 = 0.7$, then D_1 accounts for only 3% in the general population and D_2 accounts for 7%. However if we have $n_1 = n_2 = n_0 = 50$, then the sample sizes in the clustering analysis are disproportionate to the true sample sizes in the general population. Without any correction of the disproportionateness the estimated clusters would be invalid.

BIBLIOGRAPHY

Alzaid, A. A. & Proschan, F. (1994). Max-Infinite Divisibility and Multivariate Total Positivity. *Journal of Applied Probability* 31 pp. 721–730.

Association, A. P. (2013). *Diagnostic and Statistical Manual of Mental Disorders (DSM-5®).* American Psychiatric Publishing.

Benjamini, Y. & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)* 57 pp. 289–300.

Benjamini, Y. & Hochberg, Y. (2000). On the Adaptive Control of the False Discovery Rate in Multiple Testing with Independent Statistics. *Journal of Educational and Behavioral Statistics* 25 pp. 60–83.

Benjamini, Y. & Liu, W. (1999). A step-down multiple hypotheses testing procedure that controls the false discovery rate under independence. *Journal of Statistical Planning and Inference* 82 163–170.

Benjamini, Y. & Yekutieli, D. (2001). The Control of the False Discovery Rate in Multiple Testing under Dependency. *The Annals of Statistics* 29 pp. 1165–1188.

Boyd, S. & Vandenberghe, L. (2004). *Convex Optimization.* Cambridge University Press.

Bray, N. J., Leweke, F. M., Kapur, S. & Meyer-Lindenberg, A. (2010). The neurobiology of schizophrenia: new leads and avenues for treatment. *Current Opinion in Neurobiology* 20 810–815.

Brown, G. W. & Mood, A. M. (1951). On median tests for linear hypotheses. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability.* 159–166.

Davino, C., Furno, M. & Vistocco, D. (2013). *Quantile Regression.* Theory and Applications. John Wiley & Sons.

Dykstra, R. L., Hewett, J. E. & Thompson, W. A. J. (1973). Events which are Almost Independent. *The Annals of Statistics* 1 pp. 674–681.

Esary, J. D. & Proschan, F. (1972). Relationships Among Some Concepts of Bivariate Dependence. *The Annals of Mathematical Statistics* 43 pp. 651–655.

Ghosh, D. (2011). Generalized Benjamini-Hochberg procedures using spacings. *Technical Report, Penn State University* .

Green, E. K., Grozeva, D., Jones, I., Jones, L., Kirov, G., Caesar, S., Gordon-Smith, K., Fraser, C., Forty, L., Russell, E., Hamshere, M. L., Moskvina, V., Nikolov, I., Farmer, A., McGuffin, P., Consortium, W. T. C. C., Holmans, P. A., Owen, M. J., O'Donovan, M. C. & Craddock, N. (2010). The bipolar disorder risk allele at CACNA1C also confers risk of recurrent major depression and of schizophrenia. *Molecular Psychiatry* 15 1016–1022.

Hogg, R. V. (1975). Estimates of percentile regression lines using salary data. *Journal of the American Statistical Association* 70 56–59.

Joe, H. (1997). *Multivariate Models and Multivariate Dependence Concepts*. CRC Press.

Kessler, R. C., Berglund, P., Demler, O., Jin, R., Merikangas, K. R. & Walters, E. E. (2005). Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the National Comorbidity Survey Replication. *Archives of General Psychiatry* 62 593–602.

Khan, M. A. & Akella, S. (2009). Cannabis-induced bipolar disorder with psychotic features: a case report. *Psychiatry (Edgmont)* 6 44–48.

Koenker, R. (2005). *Quantile Regression*. Cambridge University Press.

Koenker, R. & Bassett, G., Jr (1978). Regression quantiles. *Econometrica: Journal of the Econometric Society* 46 33–50.

Laska, E. M. & Meisner, M. J. (1989). Testing Whether an Identified Treatment Is Best. *Biometrics* 45 pp. 1139–1151.

Lehmann, E. L. (1966). Some Concepts of Dependence. *The Annals of Mathematical Statistics* 37 pp. 1137–1153.

Lichtenstein, P., Yip, B. H., Björk, C., Pawitan, Y., Cannon, T. D., Sullivan, P. F. & Hultman, C. M. (2009). Common genetic determinants of schizophrenia and bipolar disorder in Swedish families: a population-based study. *Lancet* 373 234–239.

Ma, S., Huang, J. & Moran, M. S. (2009). Identification of genes associated with multiple cancers via integrative analysis. *BMC Genomics* 10 535.

National Institute of Mental Health (2011). NIMH Research Domain Criteria (RDoC). URL http://www.nimh.nih.gov/research-priorities/rdoc/index.shtml.

National Institute of Mental Health (2015). National Institute of Mental Health Strategic Plan for Research. URL http://www.nimh.nih.gov/about/strategic-planning-reports/introduction.shtml.

O'Donovan, M. C., Craddock, N., Norton, N., Williams, H., Peirce, T., Moskvina, V., Nikolov, I., Hamshere, M., Carroll, L., Georgieva, L., Dwyer, S., Holmans, P., Marchini, J. L., Spencer, C. C. A., Howie, B., Leung, H.-T., Hartmann, A. M., Möller, H.-J., Morris, D. W., Shi, Y., Feng, G., Hoffmann, P., Propping, P., Vasilescu, C., Maier, W., Rietschel, M., Zammit, S., Schumacher, J., Quinn, E. M., Schulze, T. G., Williams, N. M., Giegling, I., Iwata, N., Ikeda, M., Darvasi, A., Shifman, S., He, L., Duan, J., Sanders, A. R., Levinson, D. F., Gejman, P. V., Cichon, S., Nöthen, M. M., Gill, M., Corvin, A., Rujescu, D., Kirov, G., Owen, M. J., Buccola, N. G., Mowry, B. J., Freedman, R., Amin, F., Black, D. W., Silverman, J. M., Byerley, W. F., Cloninger, C. R. & Collaboration, M. G. o. S. (2008). Identification of loci associated with schizophrenia by genome-wide association and follow-up. *Nature Genetics* 40 1053–1055.

Owen, M. J., Craddock, N. & Jablensky, A. (2007). The genetic deconstruction of psychosis. *Schizophrenia Bulletin* 33 905–911.

Rhodes, D. R., Yu, J., Shanker, K., Deshpande, N., Varambally, R., Ghosh, D., Barrette, T., Pandey, A. & Chinnaiyan, A. M. (2004). Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. *Proceedings of the National Academy of Sciences of the USA* 101 9309–9314.

Sarkar, S. K. (2002). Some Results on False Discovery Rate in Stepwise Multiple Testing Procedures. *The Annals of Statistics* 30 pp. 239–257.

Shaked, M. (1982). A general theory of some positive dependence notions. *Journal of Multivariate Analysis* 12 199–218.

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 64 479–498.

Storey, J. D. (2003). The Positive False Discovery Rate: A Bayesian Interpretation and the q-Value. *The Annals of Statistics* 31 pp. 2013–2035.

van Os, J. & Kapur, S. (2009). Schizophrenia. *The Lancet* 374 635–645.

Wu, Q. & Sampson, A. R. (2012). Structured Modeling for Post-Mortem Brain Tissue Data. *Communications in Statistics - Theory and Methods* 41 1194–1213.

Yekutieli, D. (2008). False discovery rate control for non-positively regression dependent test statistics. *Journal of Statistical Planning and Inference* 138 405–415.