Learning Logic Rules for Disease Classification: With an Application to Developing Criteria Sets for the Diagnostic and Statistical Manual of Mental Disorders

Christine Mauro

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2015

ABSTRACT

Learning Logic Rules for Disease Classification: With an Application to Developing Criteria Sets for the Diagnostic and Statistical Manual of Mental Disorders

Christine Mauro

This dissertation develops several new statistical methods for disease classification that directly account for the unique logic structure of criteria sets found in the Diagnostic and Statistical Manual of Mental Disorders. For psychiatric disorders, a clinically significant anatomical or physiological deviation cannot be used to determine disease status. Instead, clinicians rely on criteria sets from the Diagnostic and Statistical Manual of Mental Disorders to make diagnoses. Each criteria set is comprised of several symptom domains, with the domains determined by expert opinion or psychometric analyses. In order to be diagnosed, an individual must meet the minimum number of symptoms, or threshold, required for each domain. If both the overall number of domains and the number of symptoms within each domain are small, an exhaustive search to determine these thresholds is feasible, with the thresholds chosen to minimize the overall misclassification rate. However, for more complicated scenarios, such as incorporating a continuous biomarker into the diagnostic criteria, a novel technique is necessary. In this dissertation, we propose several novel approaches to empirically determine these thresholds.

Within each domain, we start by fitting a linear discriminant function based upon a sample of individuals in which disease status and the number of symptoms present in that domain are both known. Since one must meet the criteria for all domains, an overall positive diagnosis is only issued if the prediction in each domain is positive. Therefore, the overall decision rule is the intersection of all the domain specific rules. We fit this model using several approaches. In the first approach, we directly apply the framework of the support vector machine (SVM). This results in a non-convex minimization problem, which we can approximate by an iterative algorithm based on the Difference of Convex functions algorithm. In the second approach, we recognize that

the expected population loss function can be re-expressed in an alternative form. Based on this alternative form, we propose two more iterative algorithms, SVM Iterative and Logistic Iterative. Although the number of symptoms per domain for the current clinical application is small, the proposed iterative methods are general and flexible enough to be adapted to complicated settings such as using continuous biomarker data, high-dimensional data (for example, imaging markers or genetic markers), other logic structures, or non-linear discriminant functions to assist in disease diagnosis.

Under varying simulation scenarios, the Exhaustive Search and both proposed methods, SVM Iterative and Logistic Iterative, have good performance characteristics when compared with the oracle decision rule. We also examine one simulation in which the Exhaustive Search is not feasible and find that SVM Iterative and Logistic Iterative perform quite well. Each of these methods is then applied to a real data set in order to construct a criteria set for Complicated Grief, a new psychiatric disorder of interest. As the domain structure is currently unknown, both a two domain and three domain structure is considered. For both domain structures, all three methods choose the same thresholds. The resulting criteria sets are then evaluated on an independent data set of cases and shown to have high sensitivities. Using this same data, we also evaluate the sensitivity of three previously published criteria sets for Complicated Grief. Two of the three published criteria sets show poor sensitivity, while the sensitivity of the third is quite good. To fully evaluate our proposed criteria sets, as well as the previously published sets, a sample of controls is necessary so that specificity can also be assessed. The collection of this data is currently ongoing. We conclude the dissertation by considering the influence of study design on criteria set development and its evaluation. We also discuss future extensions of this work such as handling complex logic structures and simultaneously discovering both the domain structure and domain thresholds.

Table of Contents

List of Figures

List of Tables

Acknowledgments

First and foremost, I need to sincerely thank my dissertation adviser, Dr. Yuanjia Wang, for her guidance and assistance over the last several years. Through countless meetings and emails, she has taught me what it means to be a great biostatistician, a respected statistical collaborator, and an exceptional mentor.

I would also like to deeply thank the chair of my dissertation committee, Dr. Bruce Levin, and the other members of my committee: Dr. Roger Vaughan, Dr. Melanie Wall, and Dr. Katherine Shear. My dissertation has been greatly improved as a result of their advice and suggestions. I would especially like to thank Dr. Shear for bringing this problem to my attention and for allowing me to use her data.

My grateful thanks are also extended to Dr. Donglin Zeng for his theoretical guidance and assistance on this project.

To my grandmother, aunts, uncles, and cousins, thank you for always reminding me what is most important in life. I especially need to thank my siblings, Jessica, Joey, and Mikey, for always providing me with laughter when I need it the most. Lastly, I need to wholeheartedly thank my parents. Without their sacrifices, never ending support, and constant motivation, this would not have been possible.

For my family

Chapter 1

Introduction

1.1 Overview

This dissertation develops several new statistical methods for disease classification that directly account for the unique logic structure of criteria sets found in the Diagnostic and Statistical Manual (DSM) of Mental Disorders. In the remainder of this chapter, we provide an overview of the clinical research problem (Section 1.2) and then briefly review the statistical methodology on which our proposed methods will be based (Section 1.3). Chapter Two presents the statistical methodology in four parts. In part 2.1, the statistical framework of the clinical problem is set up. In section 2.2, three novel statistical methods are presented to address the clinical problem. Following that, in section 2.3, we consider the theoretical properties of these methods. Finally in section 2.4, we discuss how to choose between each of the proposed methods. In Chapter Three, we examine the finite sample performance of these methods using simulation studies under various scenarios and error structures. Following that, in section 4.1 we apply our proposed methods to a real data set to develop criteria sets for Complicated Grief, a new psychiatry disorder of interest. We then evaluate the resulting criteria sets on an independent validation data set in section 4.2. Using the same validation sample, we are also able to compare, for the first time, the performance characteristics

of three previously proposed criteria sets for Complicated Grief. In Chapter Five, we conclude with a discussion on criteria set development and present some thoughts on possible extensions of these methods.

1.2 Introduction to Clinical Research Problem

1.2.1 Current State of Disease Classification and the DSM Manual

In 2011, the National Research Council of the National Academies released a report entitled "*Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and and a New Taxonomy of Disease*," [National Research Council of the National Academies, 2011] This report was the result of a committee charged with exploring the feasibility and need for a New Taxonomy of human disease based on molecular biology. The motivation for their study was the explosion of molecular data on humans, particularly those associated with individual patients, and the sense that there are large, as-yet-untapped opportunities to use these data to improve health outcomes. They found that a new taxonomy that integrates multi-parameter molecular data with clinical data, environmental data, and health outcomes in a dynamic, iterative fashion is absolutely necessary. It was their argument that the new taxonomy system should describe and define diseases based on their intrinsic biology in addition to the more traditional physical "signs and symptoms." We believe statisticians should play an integral role in helping with this merging of the "old" and "new" and in tapping data to improve health outcomes.

In the field of Psychiatry, The Diagnostic and Statistical Manual (DSM) of Mental Disorders has often been referred to as "the bible" of psychiatric diagnoses [Kupfer *et al.*, 2008]. Interestingly, DSM includes the word "statistical" in its name. This is because the first two editions of the DSM, DSM-I and DSM-II, were proposed for purposes related primarily to counting cases to determine disease prevalences. Starting with DSM-III, it was recognized that DSM diagnoses serve many

other types of clinical and clinical research purposes as well. The "word 'statistical' in DSM now takes on greater meaning, for one goal of DSM is to facilitate drawing correct statistical inferences from what is observed" [Kraemer, 2007].

Similar to medical taxonomy in general, the DSM is also in a state of transition. In May 2013, DSM-5 replaced the previous manual, DSM-IV, which was released in 1994. Although it took almost 20 years to make the last revision, the DSM is now being viewed as a living document. Following the publication of the DSM-5, ongoing review groups will be established to coordinate and oversee periodic assessments of advancements. The review groups will determine if a more intensive assessment or changes to the diagnostic criteria are warranted. This change is reflected by the switch from Roman numerals to Arabic, of which the Arabic can be updated as incremental updates are made (5.1, 5.2, etc.) [American Psychiatric Association, 2010].

Prior to the release of the new version, several members of the DSM-V Task force published an article that outlined the set of revision principals used to guide the efforts of the DSM-5 work groups [Kupfer *et al.*, 2008]. Their very first principal, based on the overall goals extending from the APA/NIH/WHO, was that all recommendations will be grounded in empirical evidence. In addition, in discussing limitations of past versions of the manual, they wrote "The reliability of DSM as a clinical tool has been upheld but less emphasis has been given to its validity. Face validity has generally gone hand-in-hand with clinical reliability, but other forms of more stringent validity, including specificity and sensitivity, are lacking." Based on both the need to ground any recommendations in empirical evidence and the need to focus on diagnostic measures such as sensitivity and specificity, rather than reliability, we believe it essential for statisticians to be involved in developing these disease classification systems or criteria sets.

Psychiatric diagnoses in the DSM manual are defined as polythetic - categorical concepts. Polythetic refers to the fact that specific mental disorders are defined by multiple symptoms, and not

all listed symptoms are necessary to consider a mental disorder present in a specific individual. Rather, a specific combination and number of symptoms, less than the total number of symptoms of the disorder, must be observed to consider a diagnosis present. This is in contrast to a monothetic classification system, in which all criteria must be met in order to have a positive diagnosis. Categorical refers to the fact that all mental disorders in the DSM are binary, "either/or" concepts. Disorders are considered present in individuals when the right combination and number of symptoms are present, and absent when those symptoms are not present in the correct combination or number. Some limitations that arise when using this type of model include comorbidity, within-category heterogeneity, and the validity of subthreshold symptomatology [Krueger and Bezdjian, 2009]. Comorbidity refers to patients meeting the criteria for two or more diagnoses. Within-category heterogeneity refers to the fact that patients with the same label are often heterogeneous with respect to key clinical features, such as severity and prognosis. Lastly, patients who do not meet the thresholds can still be significantly impaired.

In light of these limitations, [Kraemer, 2007] proposes an enhancement to DSM that she believes would enhance the reliability and validity of DSM diagnoses: "the addition of a dimensional adjunct to each of the traditional categorical diagnoses of the DSM." She argues that by including a dimensional scale, along with a categorical diagnosis, the quality of a diagnosis will be improved. The only time a dimensional scale will not add quality is if there is no meaningful clinical variation among those who are diagnosed and no clinical variation among those who are not diagnosed, which is virtually never the case. Although dimensional diagnoses did not end up being added to DSM-5, the issues raised by this paper still exist and should be considered in future revisions to the manual.

Another common criticism of the DSM is that all symptoms within a domain have equal impact in terms of their diagnostic ability because a simple sum of symptoms present is calculated. By setting a cut-point, as polythetic disorders do, we are essentially more concerned with quantity

rather than the quality of the items and are assuming that all items have similar frequency and discriminating power. Numerous studies have found this assumption to be flawed [Clark and McKenzie, 1994; Aggen *et al.*, 2005; Cooper *et al.*, 2010]. They all argue that diagnostic algorithms, such as those found in the DSM, should instead incorporate weights that represent the relative importance of a symptom in terms of its diagnostic ability.

Further, criteria sets in the current DSM all rely on the self-reporting of symptoms by patients to their clinicians. In an American Journal of Psychiatry editorial, [First and Zimmerman, 2006] argue for the possible inclusion of laboratory tests for some diagnostic criteria in addition to the current list of symptoms. The advantages, they list, are that "laboratory tests are more objective, would facilitate detection of mental disorders in primary care settings, and would highlight the neurobiological basis of psychiatric disorders." They further argue that rather than considering whether a laboratory test by itself is sufficiently sensitive and specific to make a particular psychiatric diagnosis, we should instead consider finding the combination of clinical signs and symptoms and laboratory tests that optimally defines the disorder of interest. In other words, rather than having a diagnosis based solely on a list of symptoms, or solely on a lab test, some combination of the two actually might be ideal and have improved performance over what already exists.

From a similar perspective as that of [First and Zimmerman, 2006], the current National Institute of Mental Health (NIMH) Strategic Plan calls for the development of new ways of classifying psychopathology based on dimensions of observable behavior and neurobiological measures. The Research Domain Criteria project (RDoC) has been launched to implement this strategy. The goal of this process is to define basic dimensions of functioning to be studied across multiple units of analysis, from genes to neural circuits to behaviors, cutting across disorders as traditionally defined in the DSM. The motivation for this is based on several mental health findings that do not map well onto current diagnostic categories. For example, some of the risk genes for psychotic disor-

ders appear to be associated with both schizophrenia and bipolar disorder and the same prefrontal region has been implicated in depression and PTSD. The goal of this project is to be able to translate new findings from neurobiological and behavioral research to an improved understanding of psychopathology and the development of new and/or optimally matched treatments for mental disorders. The development of RDoc will follow three guiding principles: First, RDoC is conceived as a dimensional system, spanning the range from normal to abnormal as opposed to the binary classifications found in the DSM. Second, RDoc will be agnostic about current disorder categories. The reason for this is because the intent is to generate classifications stemming from basic behavioral neuroscience, rather than starting with an illness and seeking its neurobiological underpinnings. Lastly, RDoc will use several different units of analyses in defining constructs for study, including imaging, physiological activity, behavior, and self-report of symptoms [National Institute of Mental Health, 2011]. Again, statisticians can and should play a role in this transition process.

1.2.2 Complicated Grief

The applied focus of this dissertation will be on developing a criteria set for Complicated Grief (CG), a new psychiatric disorder. However, our method is in no way limited to CG and can be extended to address the more general emerging issues in disease classification discussed above. Acute grief is a normal human response to the loss of a loved one that usually dissipates with time. In this context, grief should not be considered pathological or treated medically. However, for a small group of individuals, grief can be complicated in the sense that the symptoms are heightened and their duration prolonged [Shear *et al.*, 2011]. Those suffering from CG, also referred to as Prolonged Grief Disorder (PGD) [Prigerson *et al.*, 2009] or Persistent Complex Bereavement Disorder (PCBD) [American Psychiatric Association, 2013], often exhibit symptoms of strong yearning for the person who died, frequent thoughts or images of the deceased person, feelings of intense loneliness or

Category	Definition
A.	Event: Bereavement (loss of a significant other)
B.	Separation distress: The bereaved person experiences yearning (e.g., craving, pining, or longing for the deceased; physical or emotional suffering as a result of the desired, but unfulfilled, reunion with the deceased) daily or to a disabling degree.
C.	Cognitive, emotional, and behavioral symptoms: The bereaved person must have five (or more) of the following symptoms experienced daily or to a disabling degree:
	1. Confusion about one's role in life or diminished sense of self (i.e., feeling that a part of oneself has died)
	2. Difficulty accepting the loss
	3. Avoidance of reminders of the reality of the loss
	4. Inability to trust others since the loss
	5. Bitterness or anger related to the loss
	6. Difficulty moving on with life (e.g., making new friends, pursuing interests)
	7. Numbness (absence of emotion) since the loss
	8. Feeling that life is unfulfilling, empty, or meaningless since the loss
	9. Feeling stunned, dazed or shocked by the loss
D.	Timing: Diagnosis should not be made until at least six months have elapsed since the death.
E.	Impairment: The disturbance causes clinically significant impairment in social, occupational, or other important areas of functioning (e.g., domestic responsibilities).
F.	Relation to other mental disorders: The disturbance is not better accounted for by major depressive disorder, generalized anxiety disorder, or posttraumatic stress disorder.

doi:10.1371/journal.pmed.1000121.t003

Figure 1.1: Criteria for Prolonged Grief Disorder (PGD) proposed by [Prigerson *et al.*, 2009].

emptiness, and a feeling that life without the person has no purpose or meaning [Shear *et al.*, 2011]. As is reflected by the several names for CG, multiple criteria sets have been proposed for CG [Prigerson *et al.*, 2009; Shear *et al.*, 2011; American Psychiatric Association, 2013], and there is a lack of agreement among clinical experts. Each of these criteria sets are presented in Figures 1.1 - 1.3 and discussed in more detail below. As a result of this disagreement, there is currently no gold standard diagnosis.

Despite the fact that CG is a "new" disorder and a gold standard diagnosis does not exist, there are several ratings scales that help to distinguish CG patients from the normal population, the most common of which is the Inventory of Complicated Grief (ICG), a 19 item self-report questionnaire [Prigerson *et al.*, 1995]. In fact, both the [Prigerson *et al.*, 2009] Criteria Set and the [Shear *et al.*, 2011] Criteria Set were derived in part based on data from the ICG. Further, ICG is often used as one of the main inclusion criteria for CG treatment studies [Shear *et al.*, 2005; Shear *et al.*, 2014]. The ICG is a well validated self-report measure of CG symptom severity

TABLE 2. Proposed criteria for complicated grief

A. The person has been bereaved, i.e. experienced the death of a loved one, for at least 6 months

B. At least one of the following symptoms of persistent intense acute grief has been present for a period longer than is expected by others in the person's social or cultural environment
   1. Persistent intense yearning or longing for the person who died
   2. Frequent intense feelings of loneliness or like life is empty or meaningless without the person who died
   3. Recurrent thoughts that it is unfair, meaningless, or unbearable to have to live when a loved one has died, or a recurrent urge to die in order to find or to join the deceased
   4. Frequent preoccupying thoughts about the person who died, e.g. thoughts or images of the person intrude on usual activities or interfere with functioning

C. At least two of the following symptoms are present for at least a month:
   1. Frequent troubling rumination about circumstances or consequences of the death, e.g. concerns about how or why the person died, or about not being able to manage without their loved one, thoughts of having let the deceased person down, etc.
   2. Recurrent feeling of disbelief or inability to accept the death, like the person cannot believe or accept that their loved one is really gone
   3. Persistent feeling of being shocked, stunned, dazed or emotionally numb since the death
   4. Recurrent feelings of anger or bitterness related to the death
   5. Persistent difficulty trusting or caring about other people or feeling intensely envious of others who have not experienced a similar loss
   6. Frequently experiencing pain or other symptoms that the deceased person had, or hearing the voice or seeing the deceased person
   7. Experiencing intense emotional or physiological reactivity to memories of the person who died or to reminders of the loss
   8. Change in behavior due to excessive avoidance or the opposite, excessive proximity seeking, e.g. refraining from going places, doing things, or having contact with things that are reminders of the loss, or feeling drawn to reminders of the person, such as wanting to see, touch, hear or smell things to feel close to the person who died. (Note: sometimes people experience both of these seemingly contradictory symptoms.)

D. The duration of symptoms and impairment is at least 1 month

E. The symptoms cause clinically significant distress or impairment in social, occupational or other important areas of functioning, where impairment is not better explained as a culturally appropriate response

Figure 1.2: Criteria for Complicated Grief (CG) proposed by [Shear *et al.*, 2011].

A: "The individual experience death of a someone who has been in a close relationship"

B: Since the death, at least 1 symptom experienced on more days than not and persisted to at least 12 months after death in adults and 6 months in children.

1. Persistent Yearning/longing for deceased. In children, yearning expressed in play and behavior including behaviors that reflect being separated from and reuniting with, a caregiver or other attachment figure.
2. Intense sorrow and emotional pain in response to death.
3. Preoccupation with the deceased
4. Preoccupation with the circumstances of the death. In children, this preoccupation with the deceased may be expressed through themes of play and behavior and may extend to preoccupation with possible death of other close to them.

C: Since the death, at least 6 of following symptoms experienced more days than not, and have persisted for at least 12 months after death in adults and 6 months in children:

Reactive distress to the death:

1. Marked difficulty accepting death. In children this is dependent on the child's capacity to comprehend the meaning and permanence of the death.
2. Experiencing disbelief or emotional numbness over the loss.
3. Difficulty with positive reminiscing about the deceased.
4. Bitterness of anger related to death
5. Maladaptive appraisals about oneself in relation to the deceased or the death (e.g. self-blame)
6. Excessive avoidance of reminders of the loss (e.g., avoidance of individuals, places, or situations associated with the deceased; in children, this may include avoidance of thoughts and feelings regarding the deceased.

Social/identity disruption

1. A desire to die in order to be with the deceased
2. Difficulty trusting other individuals since the death
3. Feeling alone or detached from other individuals since the death
4. Feeling that life is meaningless or empty without the deceased, or th belief that one cannot function without the deceased
5. Confusion about one's role in life, or a diminished sense of one's identity (e.g., feeling that a part of oneself died with the deceased
6. Difficulty or reluctance to pursue interests since the loss or to plan for the future (e.g., friends and family)

D: The disturbance causes clinically significant distress or impairment in social, occupational, or other important areas of functioning.

E: The bereavement reaction is out of proportion to or inconsistent with cultural, religious, or age-appropriate norms.

Figure 1.3: Criteria for Persistent Complex Bereavement Disorder (PCBD) proposed by [American Psychiatric Association, 2013].

with prior evidence for high internal consistency (Cronbachs $\alpha = 0.94$) and test-retest reliability (0.80) [Prigerson *et al.*, 1995]. The ICG assesses a range of CG symptoms including preoccupation with the person who died, intrusive and distressing thoughts related to the death, avoidance of reminders of the person who died, feelings of yearning for the person who died, loneliness, and feelings of bitterness, anger and/or disbelief regarding the death. Each item is rated on a 5-point likert scale, with responses ranging from 0="not at all" to 4="severe."

Another measure of disease severity is the Structured Clinical Interview of Complicated Grief (SCI-CG), a 32 item scale administered by a clinician [Bui *et al.*, 2015]. This is a diagnostic instrument designed to capture 30 complicated grief symptoms that are present in at least one of the three proposed criteria sets mentioned above. The SCI-CG is scored 1="Absent," 2="Unsure or Equivocal," and 3="Present." It also includes an item on functional impairment. The SCI-CG has been shown to have good internal consistency (Cronbachs $\alpha = 0.76$), test-retest reliability (ICC= 0.68), and inter-rater reliability (ICC = 0.95) [Bui *et al.*, 2015].

Both the PGD and CG criteria sets were empirically derived, but the study samples and methodologies differed. The criteria proposed for prolonged grief disorder [Prigerson *et al.*, 2009] was based on a community sample of 291 bereaved individuals, 28 of whom were judged to have PGD. The method for deriving the criteria set included several phases that were preceded by prior instrument development and modification. In phase 1 of the final criteria development, the authors employed the Inventory of Complicated Grief - Revised (ICG-R) and used Item Response Theory (IRT) and differential item functioning (DIF) to derive a set of informative symptoms of CG. The ICG-R is a modification of the Inventory of Complicated Grief (ICG) [Prigerson *et al.*, 1995], which was designed to assess putative CG symptoms. The ICG-R included additional symptoms proposed by an expert panel. Phase 1 analyses resulted in 12 symptoms that were deemed informative, unbiased symptoms of CG and therefore were to be considered in the diagnostic algorithm. In

Phase 2, the authors ordered the individuals in terms of CG symptom severity based on the scores from the IRT analysis from the first phase. In addition, for each individual, a rater separately determined whether the person had CG or not. They then used a cut-off score that maximized the agreement between the rater and score-based diagnosis as their CG criterion standard. In Phase 3, the authors identified an optimal diagnostic algorithm for PGD. Since yearning was the most common and informative of the symptoms in the analyses, they decided that it would be considered a mandatory symptom. The authors then sought to determine using combinatorics the number and combination of the remaining 11 items that would yield the most efficient diagnosis for CG with respect to their criterion standard. The net result was a criteria set with two symptom clusters, Domain B (separation distress) and Domain C (cognitive, emotional, and behavioral symptoms). The authors proposed that one symptom out of one is required from domain B and at least five out of 12 from domain C.

Numerous criticisms of the [Prigerson *et al.*, 2009] study are discussed by [Shear *et al.*, 2011]. [Shear *et al.*, 2011] point out that the sample used by [Prigerson *et al.*, 2009] consisted of primarily older, white widows and was not necessarily generalizable to the CG population as a whole. In addition, the sample size was relatively small (n=291) with only 28 study participants judged to have CG. In addition, their criteria set was developed using IRT, which relies on the initial assumption that the trait being measured is best represented as a single factor. Further, no justification is given on why only 22 of the 39 possible symptoms of the ICG-R were included in the initial analysis. In addition, their final criteria set requires that the patient has been bereaved for at least six months, yet the data that this analysis was based on included individuals who were bereaved for less than six months. Lastly, and most importantly, the methodology they used is suspect due to its circular nature; they used the 12 items and clinical impression to select a cut point score for caseness and then used that cut point to determine which combination of those same items were important in

making a diagnosis.

The criteria proposed by [Shear *et al.*, 2011], with detailed methodology presented in a companion paper [Simon *et al.*, 2011], was based on a sample of bereaved healthy controls (n=95), patients diagnosed as having either a mood or anxiety disorder (n=369) and patients presenting for treatment of CG (n=318). Among those presenting for treatment of CG, only those who scored at least a 30 on the ICG and were also diagnosed with CG on clinical interview were considered to have the condition of interest (n=288). The authors performed an exploratory factor analysis of the baseline ICG from these cases and found a clear six-factor solution. These six factors were viewed as symptom clusters and then used to guide the development of the final criteria set. In later research, we validated this six factor structure on a much larger clinical sample using confirmatory factor analyses, and further showed it to be consistent across three distinct study samples [Mauro *et al.*, 2015b]. To be consistent with the PGD criteria, Shear et al. maintained the division of CG symptoms into separation distress (domain B) and associated symptoms (domain C). However, the factor analysis in the clinical CG sample indicated that yearning was a part of a cluster of four items that comprised one of the six factors discovered. As a result, the authors included all four symptoms together in Domain B (separation distress), rather than making yearning itself necessary. They then collapsed the remaining five factors to obtain domain C. Further, suicidal thinking and behavior was added to domain B based on strong research evidence of the association between CG and suicidality and judgment of its clinical significance. Once the domain structure was determined, they then used the combined sample (cases and controls) to compute the sensitivity and specificity for varying thresholds for each each of the two domains, choosing the best ones. The final proposed criteria set required at least one out of four symptoms from domain B and at least two out of eight from domain C.

In return, the [Shear *et al.*, 2011] criteria set received some criticisms from [Boelen and Prigerson,

2012]. They argued that the analyses were based solely on the ICG, but some symptoms included in the final criteria set were not actually tapped by this measure, resulting in a criteria set that was not completely empirically derived. Further, some of the criteria are broadly formulated allowing for too many ways to qualify for a CG diagnosis. Another concern they had was that the majority of the sample (73%) had at least one secondary diagnosis, or comorbidity. Of most interest statistically, they argued that it was not entirely clear how they moved from factor analysis results to the final criteria set. The factor analysis had six factors, while the final criteria set only had two domains. It is this specific criticism that we are targeting with our proposed methods.

The most recent criteria for CG was given in the newest edition of the DSM. Here, provisional criteria is given for Persistent Complex Bereavement Disorder (PCBD) in section III, "Emerging Measures and Model". Essentially these are disorders which require further study before being moved to the main part of the manual and being recognized as an official diagnosis. In many ways, the PCBD criteria set appears to be a compromise between [Prigerson *et al.*, 2009] and [Shear *et al.*, 2011] criteria sets. Like the CG criteria set, yearning is not considered a necessary symptom for PCBD, where one out of four possible symptoms is required from Domain B. However, similar to the PGD criteria set, a majority of symptoms is required from Domain C (six or more out of 12). In order to get CG moved into the main part of the DSM manual, it is absolutely crucial that this disagreement among the experts is resolved. As statisticians, we hope to help resolve this problem by constructing a criteria set that is as empirically based and objectively derived as possible.

Recall, one of the major criticisms the [Shear *et al.*, 2011] criteria set received was that there was not a direct link between the factor analyses and the final criteria set. With our proposed methods, we will attempt to provide that link. In this context, factors are often thought of as symptom clusters or domains; these symptom clusters or domains could also be the result of a conceptual model for the disease. We will assume that this structure is already known. Further,

one of the major differences among the criteria sets is the number of items required from each of the domains in order to have a positive diagnosis. More specifically,

- [Prigerson *et al.*, 2009]: 1 out of 1 item required from Domain A, 5 or more items out of 9 required for Domain B.
- [Shear *et al.*, 2011]: 1 or more out of 4 items required from Domain A, 2 or more items out of 8 required for Domain B.
- DSM-5: 1 or more out of 4 items required from Domain A, 6 or more items out of 12 required for Domain B.

As our result, our goal is to empirically **determine the minimum number of symptoms to require from each domain** in order to make a positive diagnosis. Essentially, this is a binary classification problem, for which we rely heavily on support vector machine (SVM) and logistic regression methods.

The data on which we will apply and evaluate our proposed methods come from three NIHM funded randomized clinical trials for CG treatment. The first, which we will refer to as the Pittsburgh study, recruited participants to a university-based clinic [Shear *et al.*, 2005]. Participants were either assigned to Complicated Grief Therapy (CGT) or to Interpersonal Psychotherapy (IPT), a standard therapy for depression. Those randomized to CGT were significantly more likely to improve (51% vs. 28%, $p = 0.02$). The second trial, referred to as CGTOA, also examined the performance of CGT versus IPT at a university-based clinic, but instead looked at a sample of only older adults (aged 60 or older). Those randomly assigned to CGT also show significantly more improvement when compared to those randomized to IPT (70.5% vs. 32.0%, $p < 0.001$), [Shear *et al.*, 2014]. The last trial, HEAL, is still ongoing. This trial is recruiting patients to four different university-based clinics to assess the efficacy of an antidepressant medication as compared

to placebo. In addition to a medication assignment, patients are also randomized to CGT or no therapy, resulting in a four arm trial. With this design, in addition to assessing for a medication effect, the possibility of a medication therapy interaction will also be evaluated. We will be using baseline data from all three of these trials, with data from the Pittsburgh study and part of CGTOA serving as our training sample and data from HEAL and another part of CGTOA as our validation sample.

1.3 Introduction to Statistical Methods

1.3.1 Binary Classification and Support Vector Machines

In a normal classification problem, a predictor function takes on a discrete number of values representing each level of the outcome. In the case of binary prediction, which is the focus of this paper, the predictor function can take only one of two values (for example diseased or not diseased). Because of this property, we can always divide the input space into a collection of regions labeled according to the classification rule. The boundaries that separate these regions, or *decision boundaries*, can either be rough or smooth depending on the function that is used for prediction. For some classification procedures this boundary is linear. For the familiar case of logistic regression, the linear decision boundary is the set of all points for which the log odds are zero, or more specifically, the hyperplane defined by $\{x|\beta_0 + \beta^T x = 0\}$. Those with log odds greater than zero are classified as diseased and those less than zero are classified as not diseased.

Another approach to classification is to explicitly model the boundary between two classes as linear. In the case of a two-class problem in a p-dimensional input space, this results in modeling the decision boundary as a hyperplane [Hastie *et al.*, 2009]. One such method that looks for this separating hyperplane is Support Vector Machines (SVM). SVMs, in the form of linear separating hyperplanes, were first discussed by Vladimir N. Vapnik in 1996. The SVM seeks to find an

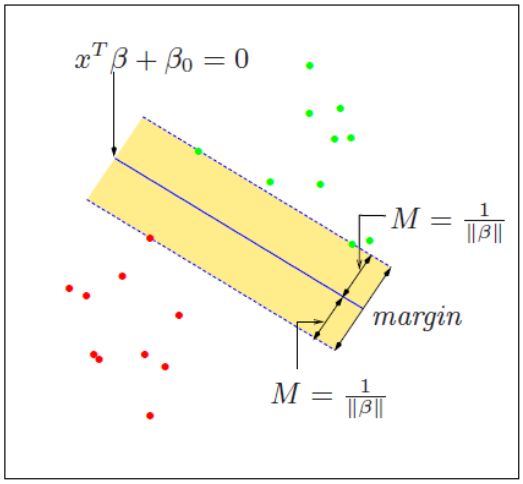

Figure 1.4: SVM illustration, linearly separable case [Hastie *et al.*, 2009].

optimally separating hyperplane if one exists, where an optimal separating hyperplane separates the two classes and maximizes the distance to the closest point from either class, or the margin [Vapnik, 1996]. In the case that one does not exist, the SVM was extended to find a hyperplane that minimizes some measure of overlap between the classes in the training data or by allowing for a non-linear decision boundary. First, attention will be focused on the case that an optimally separating hyperplane exists, or, that is, the classes do not overlap in the feature space. Next, the extension of SVM to the case when the data are not linearly separable will be discussed. The next subsections are heavily reliant on Chapter 12 of *"The Elements of Statistical Learning"* [Hastie *et al.*, 2009].

1.3.1.1 Separable Case

Consider the training data of n pairs $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, where $x_i \in \mathbb{R}^p$ and $y_i \in \{-1, 1\}$. Define a hyperplane by

$$\{x : f(x) = x^T\beta + \beta_0 = 0\},$$

where β is a unit vector: $\|\beta\| = \sqrt{\beta_1^2 + \ldots + \beta_p^2} = 1$. A classification rule induced by $f(x)$ is

$$G(x) = \text{sign}[f(x)] = \text{sign}[x^T\beta + \beta_0].$$

In fact, $f(x)$ gives the signed distance from a point x to the separating hyperplane. Since the classes are separable, the sign of y_i and $f(x_i)$ will always be the same, as no points lie on the wrong side of the hyperplane in the case that the data are linearly separable. Therefore, it is possible to find $f(x)$ such that $y_i f(x_i) > 0 \ \forall i$. This then allows us to find the hyperplane that creates the biggest margin between the training points from each class. By selecting this specific separating hyperplane, we are maximizing the SVM's ability to predict the correct classification of new data sets [Noble, 2006]. Mathematically, this results in the following optimization problem:

$$\begin{aligned} \max_{\beta,\beta_0,\|\beta\|=1} \quad & M \\ \text{subject to } y_i(x_i^T\beta + \beta_0) \quad & \geq M, i = 1, \ldots, n. \end{aligned} \tag{1.1}$$

This set of conditions ensures that all of the training points are a least a signed distance M from the decision boundary defined by β and β_0. However, the $\|\beta\| = 1$ constraint is not desirable due to the fact that this makes it a non-convex optimization problem. Therefore, the problem needs to be further manipulated. It turns out that the conditions in Equation 1.1 can be replaced with the following condition since $\|\beta\| = 1$:

$$\frac{1}{\|\beta\|} y_i(x_i^T\beta + \beta_0) \geq M$$

or

$$y_i(x_i^T\beta + \beta_0) \geq M\|\beta\|.$$

Since for any β and β_0 satisfying this inequality, any positively scaled multiple satisfies it too, we can arbitrarily set $\|\beta\| = 1/M$. This results in the following optimization problem:

$$\min_{\beta,\beta_0} \|\beta\|^2 \tag{1.2}$$
$$\text{subject to } y_i(x_i^T\beta + \beta_0) \geq 1, i = 1, \ldots, n.$$

Because this is a convex optimization problem with a quadratic criterion and linear inequality constraints, finding a solution is routine. The problem is presented graphically in Figure 1.4.

1.3.1.2 Non-separable Case

One option in the case that the data are not linearly separable is to enlarge the space of the data using kernel functions. In general a kernel function will project data from a low-dimensional space to a space of higher dimension. If one chooses a good kernel function, data that was not linearly separable in its original space might become linearly separable in a higher dimensional space. For example, see Figure 1.5. Here the original data is not linearly separable, but when the square of the original variable is included as an additional input, the data become linearly separable in two dimensions. When the linear decision boundary in the enlarged space is projected back to the original feature space, the decision boundary is no longer linear.

However, in some cases, even this method will not work. For example, if observations from two different classes share exactly the same inputs, no matter how the data is transformed, it will not be perfectly separable. In addition, projecting into very high-dimensional space can be problematic due to the *curse of dimensionality*. This refers to the fact that as the number of variables in the model increases, the number of possible solutions also increases dramatically, making it hard for the SVM to select the correct model [Noble, 2006]. In addition, if a very high dimensional kernel function is used, the boundary that results will be very specific to the examples in the training data. Therefore, a better method that can handle the non-separable case is necessary.

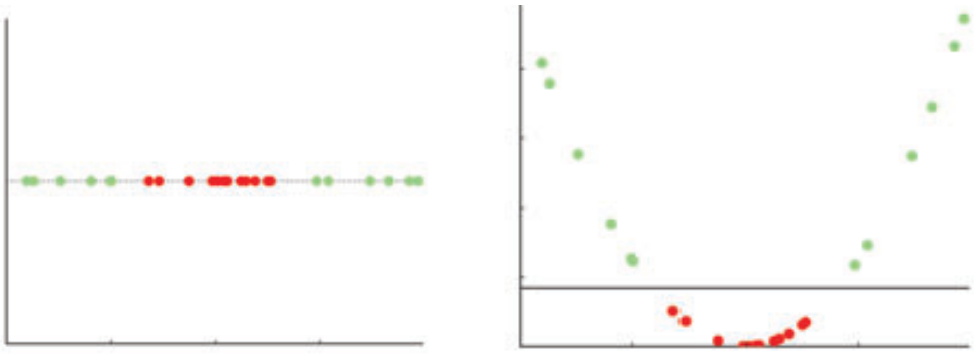

Figure 1.5: Two classes not linearly separable in one dimension, but linearly separable in two dimensions by including the square of the original values as an input [Noble, 2006].

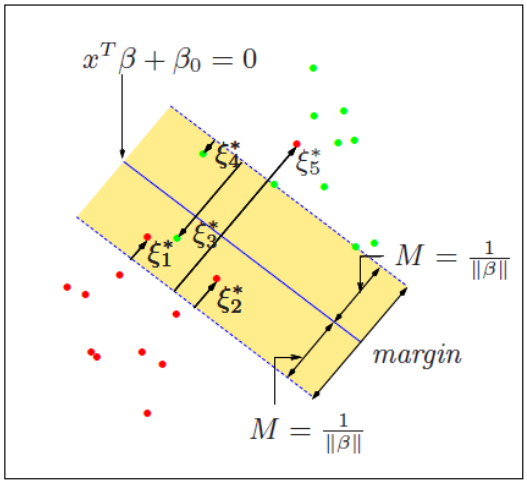

Figure 1.6: SVM illustration, non-linearly separable case where $\xi_i^* = M\xi_i$ [Hastie *et al.*, 2009].

Suppose that the classes overlap in the feature space. One way to deal with this overlap is still to maximize M, the margin, but allow for some points to be on the wrong side of the margin. Define the slack variables, a measure of the overlap for each point, as $\xi = (\xi_1, \xi_2, \ldots, \xi_N)$. Then there are two ways to modify the constraint in Equation 1.1:

$$y_i(x_i^T\beta + \beta_0) \geq M - \xi_i, \tag{1.3}$$

or

$$y_i(x_i^T\beta + \beta_0) \geq M(1 - \xi_i), \tag{1.4}$$

$\forall i,\ \xi_i \geq 0,\ \sum_{i=1}^{N} \xi_i \leq$ constant. However, both of these constraints lead to different solutions. Geometrically, constraint 1.3 measures the overlap in actual distance from the margin, while constraint 1.4 measures the overlap in relative distance, which changes with the width of the margin M. Although the former might seem like the more natural choice, it results in a nonconvex optimization problem, while the latter does not. For this reason, the standard SVM uses constraint 1.4.

The value ξ_i in the constraint $y_i(x_i^T\beta + \beta_0) \geq M(1 - \xi_i)$ is the proportional amount that the predicted value $f(x_i) = x_i^T\beta + \beta_0$ is on the wrong side of its margin. Therefore by bounding $\sum \xi_i$, we are actually putting a bound on the total proportional amount by which the predictions fall on the wrong side of their margin. Misclassifications will occur when $\xi_i > 1$, as the distance from the margin will then be a negative value. When $0 < \xi_i < 1$, the observation is on the correct side of the boundary but falls within the margin. When $\xi_i = 0$ no misclassification has been made and the point is far enough away from the decision boundary. By putting a bound on $\sum \xi_i$, call it K, we are actually limiting the number of misclassifications in the training data to K.

By dropping the norm constraint on β and letting $M = 1/\|\beta\|$ as we did before, we end up with

the following optimization problem:

$$\min_{\beta,\beta_0} \|\beta\|^2 \tag{1.5}$$
$$\text{subject to} \quad y_i(x_i^T\beta + \beta_0) \geq 1 - \xi_i \, \forall i, \quad \xi_i \geq 0, \quad \sum \xi_i \leq K.$$

This is illustrated graphically in Figure 1.6.

1.3.1.3 SVMs from a Decision Theory Perspective

From a computational prospective it turns out that it is easier to represent (1.5) in the following form:

$$\min_{\beta,\beta_0} \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{N} \xi_i \tag{1.6}$$
$$\text{subject to} \quad \xi_i \geq 0, \quad y_i(x_i^T\beta + \beta_0) \geq 1 - \xi_i \, \forall i,$$

where C, known as the cost parameter, has replaced K in the equation. Here C is a tuning parameter that controls the trade-off between wanting to maximize the margin and minimize the error bound.

By letting $\lambda = 1/C$ and noticing that $\xi_i = \max[0, 1 - y_i f(x_i)]$, Equation 1.6 can be re-expressed as:

$$\min_{\beta,\beta_0} \sum_{i=1}^{n} [1 - y_i f(x_i)]_+ + \frac{\lambda}{2}\|\beta\|^2, \tag{1.7}$$

where $f(x) = x^T\beta + \beta_0$ and $(a)_+ = a$ if $a > 0, 0$ otherwise.

This has the form *loss + penalty* which is a very familiar principle in the function estimation literature. From this perspective, the SVM is clearly minimizing a loss function, more specifically the "hinge" loss function, subject to a penalty that shrinks the coefficients (excluding the intercept) towards zero. The tradeoff between the two is controlled by the tuning parameter λ. The "hinge" loss function is defined as $L(y, f) = [1 - yf]_+$ where $(a)_+ = a$ if $a > 0, 0$ otherwise. The loss

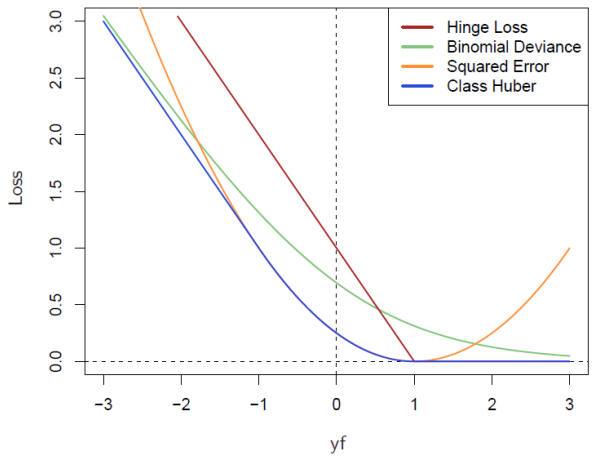

Figure 1.7: Illustration of the hinge loss function, as well some other commonly used loss functions for binary classification [Hastie *et al.*, 2009].

function will be zero only if $y_i f(x_i) \geq 1$, which is true if the points are well inside their margin. For points that are misclassified or too close to the decision boundary, a linear penalty is paid. The hinge loss function, as well as some other commonly used loss functions for binary classification, are presented graphically in Figure 1.7.

In 2008, Zou, Zhu, and Hastie presented a unified statistical view of the binary margin-based classifier, which includes the SVM. Let $y \in \mathfrak{C}$ where y is the class label and $\mathfrak{C} = \{-1, 1\}$. Consider a margin-based loss function $\phi(y, f) = \phi(yf)$, where the quantity yf is called the margin. The empirical ϕ risk is defined as $\text{EMR}_n(\phi, f) = \frac{1}{n}\sum_{i=1}^{n} \phi(y_i f(x_i))$. Then a binary margin-based ϕ classifier is obtained by solving

$$\hat{f}^{(n)} = \arg\min_{f \in \mathfrak{F}_n} \text{EMR}_n(\phi, f),$$

where $\mathfrak{F}_n$ denotes a regularized functional space. The classifier is given by $\text{sign}(\hat{f}^{(n)}(\mathbf{x}))$. In the case of SVM, ϕ is clearly the hinge loss and $\mathfrak{F}_n$ is the collection of penalized kernel estimators.

It turns out that the loss function plays a fundamental role in the success of the margin-based classification problem such as the SVM. Lin 2002 showed that in the SVM case, the population

minimizer of the hinge loss is exactly the Bayes rule. The Bayes rule is the optimal classification rule if the underlying distribution of the data is known. Thus the SVM directly approximates the Bayes rule without needing to estimate the conditional class probabilities, $P(y = c_i|x)$. Lin 2004 later extended this to the idea of Fisher-consistent loss functions for classification problems. In the traditional estimation problem, an estimator is Fisher consistent if when the estimator is calculated using the entire population, rather than the sample, the true value of the estimator is obtained. More explicitly, suppose we have a random sample $X_1, X_2, ..., X_n$ where each X_i follows a cumulative distribution F_θ. Let $\hat{\theta} = T(\hat{F}_n)$. Then the estimator is Fisher-consistent if $T(F_\theta) = \theta$. In the binary classification problem, a loss function ϕ is said to be Fisher-consistent if

$$\hat{f}(x) = \arg\min_{f(x)}[\phi(f(x))p(y = 1|x) + \phi(-f(x))p(y = -1|x)]$$

has a unique solution $\hat{f(x)}$ and

$$\text{sign}(\hat{f}(x)) = \text{sign}(p(y = 1|x) - 1/2).$$

This condition basically states that with infinite samples, you can exactly recover the Bayes rule by minimizing the ϕ loss. Lin 2004 showed that under very general conditions, margin-based loss functions are Fisher consistent and any Fisher consistent loss can be used to construct a binary-margin classifier. Since the SVM was originally developed from the perspective of simply maximizing a margin, this connection to loss function theory helps to explain the success of the SVM as a classification tool.

In general, Fisher consistency is a necessary condition for a loss function to give reasonable performance. It means that the loss function has the correct target function, but does not guarantee that the procedure converges to this target function quickly. However, in function estimation problems, Fisher consistency usually leads to consistency and rate of convergence results under some mild conditions if the function space is large enough. Since this nice property is known for estimation

problems, Lin 2004 makes a connection between consistency in classification and consistency in function estimation. He shows that consistency in classification follows from consistency in function estimation. By making this connection, researchers can rely on the well-established framework for establishing asymptotic results for loss function-based methods in function estimation problems rather than dealing with the consistency of the classifier directly.

1.3.1.4 Primal/Dual Form of the SVM

The optimization problem given in (1.6) is quadratic with linear inequality constraints, so is therefore convex and can be solved using Lagrange multipliers. The Lagrange (primal) function corresponding to this optimization problem is given by:

$$L_p = \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}\alpha_i[y_i(x_i^T\beta + \beta_0) - (1-\xi_i)] - \sum_{i=1}^{N}\mu_i\xi_i, \tag{1.8}$$

which we minimize with respect to β, β_0, and ξ_i. Setting the derivatives to zero we get,

$$\beta = \sum_{i=1}^{N}\alpha_i y_i x_i \tag{1.9}$$

$$0 = \sum_{i=1}^{N}\alpha_i y_i \tag{1.10}$$

$$\alpha_i = C - \mu_i, \forall i, \tag{1.11}$$

as well as the constraints $\alpha_i, \mu_i, \xi_i \geq 0 \forall i$.

By substituting (1.9) – (1.11) into (1.8), we obtain the Lagrangian (Wolfe) dual objective function:

$$L_D = \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{i'=1}^{N}\alpha_i\alpha_{i'}y_i y_{i'}x_i^T x_{i'}, \tag{1.12}$$

which gives a lower bound on the objective function (1.6) for any feasible point. We maximize L_D subject to $0 \leq \alpha_i \leq C$ and $\sum_{i=1}^{N}\alpha_i y_i = 0$. In addition to (1.9) – (1.11) the Karush-Kuhn-Tucker

(KKT) conditions include the constraints:

$$\alpha_i[y_i(x_i^T\beta + \beta_0) - (1 - \xi_i)] = 0 \quad (1.13)$$

$$\mu_i\xi_i = 0 \quad (1.14)$$

$$y_i(x_i^T\beta + \beta_0) - (1 - \xi_i) \geq 0, \quad (1.15)$$

for $i = 1, \ldots, N$. Together, (1.9) – (1.15) uniquely characterize the solution to the primal and dual problem. It turns out that maximizing the dual (1.12) is a simpler convex quadratic programming problem than the primal (1.8).

1.3.1.5 The Kernel Trick

Consider a transformation of the original feature variables, $h(x_i)$. By making this transformation, we will have a non-linear boundary in the original input space. In this case, the Lagrange dual function (1.12) has the form:

$$L_D = \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{i'=1}^{N} \alpha_i\alpha_{i'}y_iy_{i'}\langle h(x_i), h(x_{i'})\rangle, \quad (1.16)$$

From (1.9) we see that the solution $f(x)$ can be written as,

$$\begin{aligned} f(x) &= h(x)^T\beta + \beta_0 \\ &= \sum_{i=1}^{N} \alpha_i y_i \langle h(x_i), h(x_{i'})\rangle + \beta_0 \end{aligned} \quad (1.17)$$

So both (1.16) and (1.17) only involve $h(x)$ through inner products. It turns out that we do not need to specify the transformation $h(x)$ at all, but only need to know the kernel function

$$K(x, x') = \langle h(x), h(x')\rangle \quad (1.18)$$

that computes inner products in the transformed space. So one can either explicitly map the data with $h(x)$ and take the inner product, or take any kernel and use it right away, without knowing

or caring what $h(x)$ looks like. It turns out for particular choices of h, these inner products can be computed very cheaply when compared to the dimensionality of h.

1.3.1.6 Choosing the tuning parameter

Minimizing the objective function of the SVM, given by (1.7), requires choosing a value for the tuning parameter λ. The most common way to do this is using K-Fold Cross-Validation. This works by splitting the data into K roughly equal parts. For each $k = 1, 2, \ldots K$, we fit the model with a given value of parameter λ to the other $K - 1$ parts. We then apply this fitted model to the kth part and compute the misclassification rate. Since we will do this process for each of the K parts, we will end up with K estimates for the misclassification rate that we then average to get the cross-validation error. We repeat this process for many values of λ, choosing the λ that minimizes this cross-validation error.

1.3.2 Exploratory Factor Analysis (EFA)

Before being able to to determine the minimum number of symptoms to require from each domain, we first need to know the number of domains and the items they contain. For this we will employ the methodology of [Simon *et al.*, 2011], by performing an exploratory factor analysis (EFA) on the measure of interest, in this case the ICG, on confirmed cases. Heuristically, an exploratory factor analysis will use the correlation matrix of the items of interest and derive factors, which are weighted combinations of all of the variables. For example, if we were looking at two factors derived from the correlation matrix of 10 variables, our first two factors would look like:

$$F_1 = w_{1,1}X_1 + w_{1,2}X_2 + \ldots w_{1,10}X_{10}$$

$$F_2 = w_{2,1}X_1 + w_{2,2}X_2 + \ldots w_{2,10}X_{10},$$

where the F are factors, X are the items, and w are the weights.

As many factors as there are items is possible. The factors are extracted following specific rules. The weights for the first factor are chosen so that it explains the maximum amount of variability among the scores across all of the subjects. The second factor is derived so that it explains the maximum amount of variance that remains and is uncorrelated, or orthogonal, to the first factor. All remaining factors are derived in the same way. If we wanted to completely capture all of the variance, we would need to use all 10 factors. However, if we are okay with only explaining some percentage of the variance, then we can replace the 10 items with a reduced number of factors. There are a number of criteria to help in choosing the number of factors including percent of cumulative variation explained, number of eigenvalues greater than 1 (Kaiser's rule), scree plots, and interpretability of the factors.

Often, the factors will be rotated to aid in the overall interpretation of the factors. By examining which items load onto which factors, we can learn which items tend to be correlated with one another. These groupings may also allow us to "see" what underlying constructs our scale might be capturing. When the items are continuous, the factors are derived using Pearson's correlation. However, when we have dichotomous or ordinal items, we instead need to use polychoric correlations. Further, the assumption that the factors are uncorrelated (or orthogonal) can be relaxed, by using an oblique rotation. This allows for the factors to be correlated with one another, a realistic property when examining most mental health scales [Streiner and Norman, 2008].

1.4 Summary of Introduction

In this dissertation, we propose an empirically based algorithm for disease classification. Suppose that we have $j = 1, \cdots, p$ variables, each belonging to one of $k = 1, \cdots, G$ domains. We are going to assume we already know which domain (or symptom cluster) each variable belongs to. Our goal is to learn a classification rule based on this group structure to classify subjects as diseased

or non-diseased. Further, the form of this classification rule needs to be consistent with criteria sets found in the DSM. Our focus will be Complicated Grief, a new psychiatric disorder, but our method can also address many of the emerging issues in disease classification discussed above.

Chapter 2

Statistical Methodologies

2.1 Statistical Framework of Clinical Research Problem

Suppose there are $j = 1, \cdots, p$ variables in $k = 1, \cdots, G$ groups, or symptom domains. The number of domains, and the symptoms they contain, is usually determined by psychometric analysis or based on a conceptual model for the disease. For simplicity, we assume this structure is given a priori and present methodology on how to obtain this structure if it is unknown in Section 4.1.1. Using this grouping structure among the variables, we aim to develop a DSM-like criteria set in order to diagnose, or classify, whether a subject has a certain psychiatric disorder. Most DSM criteria sets require c_k number of symptoms to be present in each of the domains in order to receive a positive diagnosis [American Psychiatric Association, 2013]. Generally, c_k is unknown and needs to be estimated.

Let $i = 1, \cdots, n$ index subjects and $y_i \in \{0, 1\}$ be our binary outcome denoting a subject's disease status (with one representing diseased, and zero non-diseased). Let

$$x_{ik} = \text{number of symptoms in domain } k \text{ for the } i\text{th subject.}$$

Let $x_{i\cdot} = (x_{i1}, \cdots, x_{iG})$. For the general problem of estimating a decision rule $h(x_{\cdot})$ mapping a

subject's symptoms or other biological measures to his or her disease status y, define a loss function associated with h as $L[h(x_{.}), y]$. The optimal decision rule under $L(\cdot, \cdot)$ is defined as the one that minimizes the expected value of the loss function. That is, the optimal rule is defined by

$$h^*(x_{.}) = \text{argmin} E\left\{L[h(X_{i.}), Y_i]\right\}. \tag{2.1}$$

Unfortunately the expected loss function involves the joint distribution of $X_{i.}$ and Y_i and cannot be directly computed. In practice, the optimal decision rule is obtained by minimizing the empirical loss, that is,

$$\text{argmin} \sum_{i=1}^{n} \{L[h(x_{i.}), y_i]\}. \tag{2.2}$$

Evaluating the expected loss (2.1) and its empirical version (2.2) provides a theoretical basis for our methods development and comparison.

Based on the above notation, a DSM-like criteria expresses the decision rule as

$$h(x_{i.}) = I(x_{i1} \geq c_1, x_{i2} \geq c_2, \ldots, x_{iG} \geq c_G) \tag{2.3}$$

with unknown parameters $c_1, \cdots, c_G$. An estimated decision rule should be of the form

$$I(x_{i1} \geq \hat{c}_1, x_{i2} \geq \hat{c}_2, \ldots, x_{iG} \geq \hat{c}_G).$$

A commonly used loss function for binary classification problems encountered in the DSM criteria set development is the zero-one loss, or the misclassification error. Thus, substituting zero-one loss to the general problem in (2.1), the optimal DSM-like rule we aim to obtain is given by

$$\begin{aligned} h^*(x_{.}) &= I(x_1 \geq c_1^*, x_2 \geq c_2^*, \cdots, x_G \geq c_G^*) \\ &= \text{argmin} E\{I(Y_i \neq h(X_{i.,}))\} \\ &= \text{argmin} P\left\{Y_i \neq I(X_{i1} \geq c_1, X_{i2} \geq c_2, \ldots, X_{iG} \geq c_G)\right\}. \end{aligned}$$

The optimal rule defined in this sense minimizes the expected misclassification error rate, or the expected misdiagnosis rate, between the true disease status and the identified diagnosis rule. In practice, the probability measure in the above minimization problem is unknown, so we seek to minimize the empirical loss function:

$$\sum_{i=1}^{n} I\Big(y_i \neq I(x_{i1} \geq c_1, x_{i2} \geq c_2, \ldots, x_{iG} \geq c_G)\Big). \tag{2.4}$$

2.2 Proposed Statistical Methods

2.2.1 Direct Optimization Method: Exhaustive Search

The existing DSM criteria sets for most psychiatric disorders contain a small number of domains (i.e., G is small) and a moderate number of symptoms in each domain. Therefore c_k can only take a moderate number of possible values. Due to the discrete nature of this problem, the empirical loss (2.4) can be directly minimized by assessing the misclassification rate for all possible tuples, $(l_1, l_2, \ldots, l_G)$, where $l_k \in (0, p_k)$ and p_k is the maximum number of variables in domain k, and choosing the tuple that minimizes the overall misclassification rate. In the case that two or more tuples both minimize the misclassification rate, one will be chosen at random. This method will be referred to as the Exhaustive Search. Using this approach is reasonable when the number of domains being assessed is small and the number of items within each domain is also small. As either of these become larger, or if one of the domains is no longer a count of symptoms, but rather a continuous biomarker [First and Zimmerman, 2006], the Exhaustive Search becomes infeasible and another solution is necessary. We introduce a few alternatives in the next sections and compare pros and cons of the Exhaustive Search with these other methods in Section 2.4.

2.2.2 Linear Discriminant Rules

Let the binary variable z_{kj} denote the presence or absence of the jth symptom in the kth domain, and let p_k denote the total number of symptoms in this domain. The current system used in DSM criteria sets uses diagnosis decision rules based on the total counts of symptoms in each domain, that is, $x_k = \sum_{j=1}^{p_k} z_{kj}$. However, a symptom-specific weight can also be used to build diagnosis scoring rules and assist clinical decisions (e.g., Framingham risk score). Let $z_{.k} = (z_{1k}, \cdots, z_{p_k k})$. Consider the linear discriminant rule for the kth domain as

$$f_k(z_{.k}) = \beta_{0k} + \sum_{j=1}^{p_k} \beta_{jk} z_{jk}, \quad k = 1, \cdots, G, j = 1, \cdots, p_k,$$

and, if $f_k(z_{.k}) > 0$, then the criteria in the kth domain is met. Note that when $\beta_{1k} = \beta_{2k}, \cdots, = \beta_{p_k k}$, this weighted rule reduces to the unweighted symptom counts based rule. To estimate linear discriminant rules with unknown weights, the Exhaustive Search does not apply.

In the situation where there is only one variable per domain, such as the symptoms count, the linear discriminant rule within each domain takes the form,

$$f_k(x_k) = \beta_{0k} + \beta_{1k} x_k, \quad k = 1, \cdots, G,$$

and if $f_k(x_k) > 0$ then the criteria in the kth domain is considered met. The overall DSM-diagnosis rule is to meet all the criteria in each domain, that is, a certain number of symptoms is present in each of the domains. The overall rule therefore takes the "AND" form. That is, a positive diagnosis will be issued only if the prediction in each domain is positive. In the situation of one variable per domain, the linear discriminant rule is equivalent to the previously introduced existing DSM decision rule $I(x_1 \geq c_1, \cdots, x_G \geq c_G)$ in (2.3), and thus covers it as a special case.

For the ease of notation and presentation, the rest of this section is developed for the one variable per domain scenario. The results automatically carry over to the more general case of

multiple variables per domain. Based on the DSM structure, the overall decision rule is the sign of the minimum of all domain-specific rules, which is defined as

$$h(x_{\cdot}) = \text{sign}\{\min(f_1(x_1), \cdots, f_G(x_G))\}.$$

Substituting this decision rule and zero-one loss into (2.2), the optimal decision functions we aim to obtain is given by

$$\begin{aligned}(f_1^*(x_1), ..., f_G^*(x_G)) &= \text{argmin} E\{L[h(X_{\cdot}), Y]\} \\ &= \text{argmin} P\left\{Y \neq \text{sign}(\min(f_1(X_1), ..., f_G(X_G)))\right\},\end{aligned}$$

where Y_i is the binary outcome of disease status, but now coded $\{-1, 1\}$. The final optimal decision rule (or diagnosis rule) is given by

$$h^*(x_{\cdot}) = \text{sign}(\min(f_1^*(x_1), ..., f_G^*(x_G))).$$

This optimal rule aims to minimize the misclassification error rate (or the misdiagnosis rate) as in (2.1), only with the decision function replaced by several domain-specific linear discriminant functions. Corresponding to (2.2), the empirical loss function to be minimized here is

$$\sum_{i=1}^{n} I\left\{y_i \neq \text{sign}(\min(f_1(x_1), ..., f_G(x_G)))\right\}. \tag{2.5}$$

2.2.3 Iterative Optimization Method

2.2.3.1 Overview

It is well known that the loss function in (2.5) is difficult to minimize due to discontinuity and non-convexity of the zero-one loss [Hastie *et al.*, 2009]. To tackle this problem, we consider two iterative algorithms. The first common approach is to replace the zero-one loss by some convex surrogate loss and develop a computationally tractable procedure [Steinwart, 2005]. For example,

we can fit this model by a large-margin based classifier, in particular, the support vector machine (SVM), which replaces the zero-one loss by a regularized hinge loss function as a surrogate. Here we approximate the loss simultaneously using all domain-specific decision functions and all subjects. Our first algorithm then iteratively updates the approximated surrogate loss function. We introduce details of this algorithm in Section 2.2.3.2, and refer to it as iteratively optimizing a simultaneous approximation of the original empirical zero-one loss function.

Our second iterative algorithm is motivated from an observation that the expected population loss function can be re-expressed in an alternative form, and thus the optimal rule also has an alternative conditional expression. An iterative algorithm can then be applied to fit each domain-specific rule in turn using a surrogate loss on a subset of subjects conditioning on the other domains. We introduce details of this second algorithm in Section 2.2.3.3, and refer to it as optimizing a surrogate of a conditional optimal rule on a subsample.

2.2.3.2 Simultaneous algorithm:

The hinge loss function, a convex approximation to the zero-one loss function, is defined as $L(y, f) = [1 - yf]_+$ where $(a)_+ = a$ if $a > 0$, and 0 otherwise. Replacing the original loss function (2.5) by a regularized hinge loss, the resulting optimization problem is thus given by:

$$\min_{\beta_{0k},\beta_{1k},k=1,\cdots,G} \quad \{1 - y_i \min(f_1(x_{i1}), ..., f_G(x_{iG}))\}_+ + \lambda \sum_{k=1}^{G} \frac{1}{2}\|\beta_{1k}\|^2. \tag{2.6}$$

By letting $\lambda = 1/C$ and noticing that $\xi_i = \max[0, 1 - y_i \min(f_1(x_{i1}, ..., f_G(x_{iG}))]$, the optimization problem (2.6) can be reparameterized into:

$$\min_{\boldsymbol{\beta}_k,\xi_i} \frac{1}{2}(\beta_{11}^2 + \cdots + \beta_{1G}^2) + C \sum_{i=1}^{n} \xi_i, \tag{2.7}$$

subject to the following constraint

$$y_i \min(f_1(x_{i1}), ..., f_G(x_{iG})) \geq 1 - \xi_i, \quad \xi_i \geq 0. \tag{2.8}$$

For diseased subjects, $y_i = 1$ and $f_k > 0, \forall k$, so constraint (2.8) actually requires $y_i f_k \geq 1 - \xi_i, \forall k$. For non-diseased subjects, $y_i = -1$ and only requires at least one f_k to be negative enough for (2.8) to hold. As a result, the optimization problem can be rewritten as follows:

$$\min_{\boldsymbol{\beta}_k, \xi_i} \frac{1}{2}(\beta_{11}^2 + \cdots + \beta_{1G}^2) + C \sum_{i=1}^{n} \xi_i$$

subject to two separate sets of constraints for diseased and non-diseased subjects:

$$\text{for } y_i = 1 : y_i(\beta_{0k} + \beta_{1k} x_{ik}) \geq 1 - \xi_i, k = 1, \cdots, G, \quad \xi_i \geq 0 \tag{2.9}$$

$$\text{for } y_i = -1 : y_i \min(\beta_{01} + \beta_{11} x_{i1}, \cdots, \beta_{0G} + \beta_{1G} x_{iG}) \geq 1 - \xi_i, \quad \xi_i \geq 0. \tag{2.10}$$

For the constraints (2.9) placed on **diseased** subjects, there is a penalty controlled by non-zero slack variables ξ_i if the prediction in **any** of the G domains is negative. For the constraints (2.10) placed on **non-diseased** subjects, there is a penalty only if **all** of the predictions in each domain yields a positive sign, i.e., $\min(\beta_{01} + \beta_{1k} x_{i1}, \cdots, \beta_{0G} + \beta_{1k} x_{iG}) > 0$. Otherwise, if at least one domain yields a negative sign, then $\min(\beta_{01} + \beta_{1k} x_{i1}, \cdots, \beta_{0G} + \beta_{1k} x_{iG}) < 0$, and the penalty is zero. The constraint (2.10) involves taking a minimum as an operation.

To minimize the empirical loss, we encourage the fitted overall decision rule to have the same sign as the outcome. In order to have a a positive classification (or diagnosis), the criteria in each domain would need to be met. In contrast, in order to receive a negative classification (or diagnosis), only the criteria in one of the domains is required not to be met. It is straightforward to show that the loss function corresponding to this optimization problem (2.6) is

$$\begin{aligned} \min_{\beta_{0k}, \beta_{1k}, k=1,\cdots,G} \quad & \Big\{ \sum_{i=1}^{n} \max_{k=1,\cdots,G} I(y_i = 1)[1 - (\beta_{0k} + \beta_{1k} x_{ik})]_+ \\ & + \sum_{i=1}^{n} \min_{k=1,\cdots,G} I(y_i = -1)[1 + (\beta_{0k} + \beta_{1k} x_{ik})]_+ \\ & + \frac{1}{2} \lambda_n (||\beta_{11}||^2 + \cdots + ||\beta_{1G}||^2) \Big\}. \end{aligned}$$

The key point is that the loss function for the positive and negative classes is different. The "max" operation corresponds to L_∞ norm of G hinge loss functions.

Once we have estimated our linear discriminant functions, the parameters c_k, i.e., the number of symptoms to be required from each domain can be estimated as

$$\hat{c}_k = [x^* : \text{such that } \hat{\beta}_{0k} + \hat{\beta}_{1k} x^* > 0],$$

where $[x]$ denotes the smallest integer a such that $a \geq x$. In other words, $\hat{c}_k$ is the smallest integer leading to a positive classification in the kth domain. If $\hat{c}_k$ is greater than the maximum number of variables in domain k denoted as p_k, we then let $\hat{c}_k = p_k$. If $\hat{c}_k < 0$, we let $\hat{c}_k = 0$. In this case, the kth domain is not needed for diagnosis and we achieve domain level variable selection. For a domain negatively related to disease status (for example, less number of beneficial factors indicates disease status), we reverse code these factors to count non-presence of beneficial factors. By this reverse coding, missing more beneficial factors (more non-present beneficial factors) indicates a higher likelihood of disease.

Recall our minimization problem is:

$$\min \quad \sum_{i=1}^{n} \{1 - y_i \min(f_1(x_{i1}), ..., f_G(x_{iG}))\}_+ + \lambda_n \sum_{k=1}^{G} \frac{1}{2} \|\beta_{1k}\|^2.$$

Unfortunately, due to the minimization inside the objective function, $\min(f_1, \cdots, f_G)$, this is not a convex minimization problem. However, the above objective function can be rewritten as

$$\begin{aligned} \sum_{i=1}^{n} I(y_i = 1) \max \{(1 - f_1(x_{i1}))_+, ..., (1 - f_G(x_{iG}))_+\} + \lambda_n \sum_{k=1}^{G} \frac{1}{2} \|\beta_{1k}\|^2 \\ - \sum_{i=1}^{n} I(y_i = -1) \max \{(-1 - f_1(x_{i1}))_+, ..., (-1 - f_G(x_{iG}))_+\}, \end{aligned} \tag{2.11}$$

which is a difference of two convex functions that allows us to use a global optimization technique called the difference convex (DC) alogorithms [An and Tao, 1997].

Here we briefly introduce the DC algorithm. Let $s = s_1 + s_2$ where s_1 is convex and s_2 is concave. The basic idea of the DCA is to construct a sequence of subproblems defined by the affine minorization of s_w,

$$s_1(w) + s_2(w^l) + \langle \nabla s_2(w^l), w - w^l \rangle$$

and solve them iteratively, where $\nabla s_2(w)$ is the subgradient of $s_w(w^l)$ at w^l. Given the solution of the lth subproblem, the $(l+1)$th subproblem can be solved by minimizing $s_1(w) + \langle \nabla s_2(w^l), w \rangle$ with respect to w. By concavity of s_2, DCA yields a sequence of nonincreasing convex upper approximations $s_1(w) + s_2(w^l) + \langle \nabla s_2(w^l), w - w^l \rangle$ to s(w). This process is iterated until convergence is established.

Algorithm 1: Simultaneous Approximation of Loss Function

Based on formulation (2.11), we propose the following algorithm for the optimization:

Step 1. Estimate f_1, ..., f_G separately using standard SVMs and treat these as initial estimates.

Step 2. Iteratively apply the DC (difference of convex functions) algorithm to update $f_1, ..., f_G$. This requires updating only one f_k at a time, replacing all others with their current estimate. At each iteration, it is a quadratic programming optimization.

This iteration algorithm above is based on the cyclic coordinate descent algorithm [Luenberger, 1984]. Cyclic coordinate descent begins by setting all variables to some initial value. It then sets the first variable to a value that minimizes the objective function, holding all other variables constant. This is a one-dimensional optimization problem. The algorithm then finds the minimizing value of a second variable, while holding all other values constant (including the new value of the first variable). Then the third variable is optimized and so on. When all variables have been passed, the algorithm restarts. Multiple passes are made until some convergence criterion is met [Genkin *et al.*,

2007]. The convergence properties of coordinate descent in convex problems are well-established [Tseng, 2001]. For non-convex problems, global convergence is not guaranteed and a reasonable choice of initial values is required. For each proposed method, we provide appropriate initial values. Using these values, we found satisfactory convergence performance in our simulations studies. This algorithm has been successfully applied to various types of outcomes (e.g., continuous outcomes, survival outcomes) and applications [Genkin *et al.*, 2007].

Details of DC algorithm for step 2:

In order to apply the DC algorithm to our problem, we first need to recognize that the function in (2.11) is equivalent to

$$\sum_{i=1}^{n} \max\{(1-y_i f_1(x_{i1}))_+, ..., (1-y_i f_G(x_{iG}))_+\} + \lambda_n \sum_{k=1}^{G} \frac{1}{2}\|\beta_{1k}\|^2$$
$$- \sum_{i=1}^{n} I(y_i=-1)\Big[\max\{(-1-f_1(x_{i1}))_+, ..., (-1-f_G(x_{iG}))_+\}$$
$$+ \max\{(1+f_1(x_{i1}))_+, ..., (1+f_G(x_{iG}))_+\}\Big].$$

To minimize this function, we start with f_1, fixing $f_2, \dots f_G$ as done in a cyclic coordinate descent algorithm.

Consider updating f_1. Let

$$a_i = \max\{(1-y_i f_2(x_{i2}))_+, ..., (1-y_i f_G(x_{iG}))_+\},$$

$$b_i = \max\{(-1-f_2(x_{i2}))_+, ..., (-1-f_G(x_{iG}))_+\}$$

and

$$c_i = \max\{(1+f_2(x_{i2}))_+, ..., (1+f_G(x_{iG}))_+\}.$$

We aim to minimize

$$\frac{1}{2}\lambda_n\|\beta_{11}\|^2 + \sum_{i=1}^n \max\left\{(1-y_i f_1(x_{i1}))_+, a_i\right\}$$

$$-\sum_{i=1}^n I(y_i=-1)\left[\max\left\{(1+f_1(x_{i1}))_-, b_i\right\} + \max\left\{(1+f_1(x_{i1}))_+, c_i\right\}\right].$$

A DC algorithm is another iteration procedure: at the kth iteration, let the current f_1 be

$$f_1^{(k)}(x_{i1}) \equiv \beta_{01}^{(k)} + \beta_{11}^{(k)} x_{i1}.$$

To obtain the updated β_{01}, β_{11}'s, we minimize

$$\begin{aligned} & \sum_{i=1}^n \max\left\{(1-y_i f_1(x_{i1}))_+, a_i\right\} \\ - & \sum_{i=1}^n I(y_i=-1)\left[I(1+f_1^{(k)}(x_{i1}) \le -b_i) + I(1+f_1^{(k)}(x_{i1}) > c_i)\right] f_1(x_{i1}) \\ + & \frac{1}{2}\lambda_n\|\beta_{11}\|^2. \end{aligned} \tag{2.12}$$

This is now a convex minimization so we can obtain its dual problem and solve it by quadratic programming.

As an aside, for some criteria sets in the DSM manual, the relationship between the domains is "OR", rather than "AND". One example is ADHD, where children must have at least six symptoms from either the inattention group or the hyperactivity and impulsivity criteria [American Psychiatric Association, 2013]. To fit an "OR" relationship, our classification rule now takes on the form:

$$\text{sign}\left\{\max(f_1, \cdots, f_G)\right\}.$$

As long an individual meets the criteria for at least one domain, i.e. f_k is positive for at least one k, that individual will be classified as diseased. The objective function is still the same as (2.7), however, the constraints will be modified as

$$\text{for } y_i = 1 : y_i \max(\beta_{01} + \beta_{11}x_{i1}, \cdots, \beta_{0G_1} + \beta_{1G_1}x_{iG_1}) \ge 1 - \xi_i, \quad \xi_i \ge 0,$$

$$\text{for } y_i = -1 : y_i(\beta_{0k} + \beta_{1k}x_{ik}) \ge 1 - \xi_i, k = 1, \cdots, G_1, \quad \xi_i \ge 0$$

Using linear constraints to replace the 'max' operation in the above display, we obtain

$$\text{for } y_i = 1 : \beta_{01} + \beta_{11} x_{i1} \leq \zeta_i, \cdots, \beta_{0G_1} + \beta_{1G_1} x_{iG_1} \leq \zeta_i, \quad y_i \zeta_i \geq 1 - \xi_i, \quad \xi_i \geq 0.$$

Also, note that the "OR" rule can be fit by switching labels for y_i, i.e., replacing "1" by "-1" and vice-versa, and applying the same algorithm for the "AND" rule presented above.

2.2.3.3 Conditional algorithm

As an alternative to Algorithm 1 proposed above, we propose a second algorithm, referred to as the Conditional Approach, that is motivated by solving for the optimal rule for our decision function iteratively within each domain. This approach is based on a useful observation that re-expresses the expected population loss function in an alternative form, and thus reveals an alternative conditional form of the optimal rule on a sub-sample of subjects given decision functions for the other domains. We can then apply an iterative algorithm to estimate this optimal rule by updating each domain sequentially.

Theoretical motivation for Algorithm 2:

Still let X_k denote the number of symptoms for a subject (or more generally, a feature variable) in each domain k for $k = 1, ..., G$. Recall that the optimal decision functions we aim to obtain is

$$(f_1^*(x_1), ..., f_G^*(x_G)) = \text{argmin} P\left\{Y \neq \text{sign}(\min(f_1(X_1), ..., f_G(X_G)))\right\}.$$

Note

$$\begin{aligned} & P\left\{Y \neq \text{sign}(\min(f_1(X_1), ..., f_G(X_G)))\right\} \\ = \; & E\left[I(f_2(X_2) > 0, ..., f_G(X_G) > 0) I(Y \neq \text{sign}(f_1(X_1)))\right] \\ & + E\left[I(\text{at least one of } f_2(X_2), ..., f_G(X_2) \leq 0) I(Y \neq -1)\right]. \end{aligned}$$

Therefore, given $f_2^*, ..., f_G^*$, f_1^* minimizes

$$E\left[I(f_2^*(X_2) > 0, ..., f_G^*(X_G) > 0)I(Y \neq \text{sign}(f_1(X_1))\right]$$

$$= \int_{x_1} \{P(f_2^*(X_2) > 0, ..., f_G^*(X_G) > 0, Y = 1|X_1 = x_1)I(\text{sign}(f_1(x_1)) = -1)$$

$$+P(f_2^*(X_2) > 0, ..., f_G^*(X_G) > 0, Y = -1|X_1 = x_1)I(\text{sign}(f_1(x_1)) = 1)\} \, dP_1(x_1),$$

where $P_1(x_1)$ denotes the distribution of X_1. Therefore, it is clear that the optimal Bayes classifier for f_1^* is

$$\text{sign}(f_1^*(x_1)) = \text{sign}\left[P(f_2^*(X_2) > 0, ..., f_G^*(X_G) > 0, Y = 1|X_1 = x_1)\right.$$

$$\left.-P(f_2^*(X_2) > 0, ..., f_G^*(X_G) > 0, Y = -1|X_1 = x_1)\right]$$

$$= \text{sign}\left[P(Y = 1|X_1 = x_1, A_1^*) - P(Y = -1|X_1 = x_1, A_1^*)\right],$$

where

$$A_1^* = \{f_2^*(X_2) > 0, ..., f_G^*(X_G) > 0\}.$$

Similarly, we obtain that the Bayes classifier $f_k^*(x_k), k = 1, ..., G$, is given as

$$\text{sign}(f_k^*(x_k)) = \text{sign}\left[P(Y = 1|X_k = x_k, A_k^*) - P(Y = -1|X_k = x_k, A_k^*)\right],$$

where

$$A_k^* = \left\{f_1^*(X_1) > 0, ..., f_{k-1}^*(X_{k-1}) > 0, f_{k+1}^*(X_{k+1}) > 0, ..., f_G^*(X_G) > 0\right\}.$$

Algorithm 2: Conditional Optimal Rule

From the above derivation, it motivates us to develop the following algorithm. Suppose A_k^* to be known. Then to estimate f_k^*, it would be ideal to minimize the empirical risk

$$\sum_{i=1}^{n} I(\text{subject } i \in A_k^*)I(y_i f_k(x_{ik}) \leq 0)).$$

However, the minimization of this empirical risk is not feasible. Thus, using the formulation of SVM, we estimate f_k by replacing the second indicator function by its surrogate loss and instead minimize the following regularized hinge-loss:

$$\sum_{i=1}^{n} I(\text{subject } i \in A_k^*) \left(\{1 - y_i f_k(x_{ik})\}_+ + \lambda_n \|f_k\|^2 \right),$$

where $(1-x)_+ = \max(x, 0)$ and $\|f_k\|$ is some reproducing kernel Hilbert space (RKHS) norm for f_k. For the linear discriminant rule in Section 2.2.2, $\|f_k\| = \sum_j \beta_{jk}^2$. This minimization can be easily carried out using existing SVM software packages.

From this perspective, the algorithm is not specific to the SVM. In fact, any appropriate loss function could be used. For example, we could also estimate f_k by minimizing the negative log-likelihood loss (binomial deviance) for logistic regression:

$$\sum_{i=1}^{n} I(\text{subject } i \in A_k^*) \log[1 + e^{-y_i f_k(x_{ik})}].$$

Here y_i would need to be coded as $(0, 1)$ instead of $(-1, 1)$. Again, this minimization can be easily carried out using existing logistic regression software packages.

Since A_k^* is not known, we then propose the iterative procedure as follows:

Step 1. We estimate $f_1, ..., f_G$ independently using their corresponding feature variables in each domain and treat them as initial classifiers;

Step 2. For $k = 1, ..., G$, we define A_k the same way as A_k^* but replace those f's by the updated f's. We then apply the SVM or logistic regression to estimate f_k using the subjects who belong to A_k.

Step 3. We iterate Step 2 until convergence.

When using the formulation of the SVM, the method will be referred to as **SVM Iterative**. When using the logistic loss function, the method will be referred to as **Logistic Iterative**. Note

that like Algorithm 1, this algorithm also relies on cyclic coordinate descent, i.e. estimating one parameter while holding all of the other parameters constant.

For clarity's sake, let us assume that we only have two domains and we are using the SVM loss function. Then the algorithm would proceed as follows:

1. Fit a SVM using only the domain 1 variable x_{i1} to obtain decision function $\hat{f}_1(x_{i1})$.

2. Restrict the full sample to all subjects who meet the domain 1 criterion, i.e., the sub-sample $\mathcal{S}_1 = \{i : \hat{f}_1(x_{i1}) > 0\}$.

3. Using only the subjects in $\mathcal{S}_1$ and the domain 2 variable, obtain decision function $\hat{f}_2(x_{i2})$.

4. Now restrict the full sample to all subjects who meet the domain 2 criterion, i.e., $\mathcal{S}_2 = \{i : \hat{f}_2(x_{i2}) > 0\}$, and use the domain 1 variable x_{i1} to re-fit decision function $\hat{f}_1(x_{i1})$.

5. Iterate steps 2 and 3 until convergence. The convergence criterion will be met when the parameter estimates are no longer changing (below some small threshold). The final decision rule is to classify a subject as diseased if $\hat{f}_1(x_{i1}) > 0$ and $\hat{f}_2(x_{i2}) > 0$.

2.3 Theoretical Considerations

In this section, we examine the asymptotic properties of the Exhaustive Search and the iterative method.

2.3.1 Asymptotic properties of exhaustive search algorithm

Let $Z_i = (X_{11}, \cdots, X_{1G}, Y_1), \cdots, Z_n = (X_{n1}, \cdots, X_{nG}, Y_n)$ be i.i.d. random variables in a measurable space $(\mathcal{X}, \mathcal{A})$ with probability law P, and for a measurable function $h : \mathcal{X} \to \mathbb{R}$. Let the

expectation, empirical measure and empirical process at f be denoted by

$$Ph = \int h dP, \quad \mathbb{P}_n h = \frac{1}{n}\sum_{i=1}^{n} h(Z_i), \quad \mathbb{G}_n h = \sqrt{n}(\mathbb{P}_n - P)h.$$

Define a specific functional $h(\cdot)$ as $h(Z_i) = I\Big(Y_i \neq I(X_{i1} \geq c_1, \ldots, X_{iG} \geq c_G)\Big)$. Recall the solution for exhaustive search method, $\hat{c}_n = (\hat{c}_1, \cdots, \hat{c}_G)$, is obtained by minimizing the empirical objective function defined in (2.4),

$$\mathbb{P}_n h \;\; = \;\; \frac{1}{n}\sum_{i=1}^{n} h(Z_i) = \frac{1}{n}\sum_{i=1}^{n} I\Big(Y_i \neq I(X_{i1} \geq c_1, \ldots, X_{iG} \geq c_G)\Big).$$

Let $c^* = (c_1^*, \cdots, c_G^*)$ denote the true optimal cut points that minimize the expected loss,

$$Ph = P\left\{Y_i \neq I(X_{i1} \geq c_1, \ldots, X_{iG} \geq c_G)\right\}$$

and we assume that such a maximum is unique. We show the following theorem hold for $\hat{c}_n$ by empirical processes theory [van der Vaart and Wellner, 1996].

Theorem 1 *Under conditions (a) c^* exists and is unique; and (b) $P(Y = 1|X = x)$ and the joint density of X is continuous in the whole space $[-\infty, \infty] \times \cdots \times [-\infty, \infty]$. Then it holds that with probability one,*

$$\hat{c}_n \to c^*.$$

Proof. Indicator functions $I(x_1 \geq c_1, \ldots, x_G \geq c_G)$ are cadlag processes which are bounded in total variation and belong to the Vapnik-Červonencis class. Thus they are bounded in uniform entropy integral with square-integrable envelope. It follows that they belongs to a Donsker class, and hence Glivenko-Cantelli. Therefore $\{h \in \mathcal{F} : I\Big(Y_i \neq I(X_{i1} \geq c_1, \ldots, X_{iG} \geq c_G), 0 \leq c_1 \leq p_1, \cdots, 0 \leq c_G \leq p_G\Big)\}$ is Glivenko-Cantelli. By the Glivenko-Cantelli theorem, with probability one,

$$\sup\Big|(\mathbb{P}_n - P)I\Big(Y_i \neq I(X_{i1} \geq c_1, \ldots, X_{iG} \geq c_G)\Big)\Big| \to 0.$$

For any convergent subsequence in $\hat{c}_n = (\hat{c}_1, ..., \hat{c}_G)$ which converges to $(\tilde{c}_1, ..., \tilde{c}_G)$ ($\tilde{c}$'s can be infinity), we have

$$\begin{aligned}
&\mathbb{P}_n I\Big(Y_i \neq I(X_{i1} \geq c_1^*, \ldots, X_{iG} \geq c_G^*)\Big) \\
&\geq \mathbb{P}_n I\Big(Y_i \neq I(X_{i1} \geq \hat{c}_1, \ldots, X_{iG} \geq \hat{c}_G)\Big) \\
&= (\mathbb{P}_n - P) I\Big(Y_i \neq I(X_{i1} \geq \hat{c}_1, \ldots, X_{iG} \geq \hat{c}_G)\Big) + P I\Big(Y_i \neq I(X_{i1} \geq \hat{c}_1, \ldots, X_{iG} \geq \hat{c}_G)\Big) \\
&\geq P I\Big(Y_i \neq I(X_{i1} \geq \hat{c}_1, \ldots, X_{iG} \geq \hat{c}_G)\Big) - \Big|(\mathbb{P}_n - P) I\Big(Y_i \neq I(X_{i1} \geq \hat{c}_1, \ldots, X_{iG} \geq \hat{c}_G)\Big)\Big|
\end{aligned}$$

If we take the limit of both the left-hand side (first line) and the right-hand side (last line) and note that the second term of the right-hand side converges to zero, we can conclude based on condition (b),

$$P I\Big(Y_i \neq I(X_{i1} \geq c_1^*, \ldots, X_{iG} \geq c_G^*)\Big) \geq P I\Big(Y_i \neq I(X_{i1} \geq \tilde{c}_1, \ldots, X_{iG} \geq \tilde{c}_G)\Big).$$

Since c^* is the unique minimum by condition (a), it yields that $\tilde{c} = c^*$. In other words, any convergent subsequence in $\hat{c}_n$ must converge to c^*. Therefore, the whole sequence $\hat{c}_n$ converges to c^* almost surely.

Furthermore, if we assume that $P I\Big(Y_i \neq I(X_{i1} \geq c_1, \ldots, X_{iG} \geq c_G)\Big)$ is twice continuously differentiable in a neighborhood of c^* and its second derivative is strictly negative, then the asymptotic normality of the exhaustive search solution can be obtained from the standard M-theorem (c.f. Thorem 3.2.16, [van der Vaart and Wellner, 1996]). In other words, $\sqrt{n}(\hat{c}_n - c^*)$ converges to a normal distribution with mean zero.

2.3.2 Fisher consistency of minimizing hinge-loss

Recall, the iterative method minimizes a hinge-loss instead of a zero-one loss and it considers a more general class of diagnosis rules that includes the DSM-like rule. We show Fisher consistency of

using hinge-loss to replace zero-one loss. Let $f_1^*, ..., f_G^*$ be the limit of the optimal decision functions minimizing the population hinge loss function,

$$L(f_1, ..., f_G) = E\left[\{1 - Y \min(f_1(X_1), ..., f_G(X_G))\}_+\right].$$

Thus, by simple algebra for any $X = (X_1, ..., X_G)$, $f_k^*, k = 1, ..., G$ should minimize

$$P(Y = 1|X)\{1 - \min(f_1(X_1), ..., f_G(X_G))\}_+$$
$$+P(Y = -1|X)\{1 + \min(f_1(X_1), ..., f_G(X_G))\}_+.$$

Due to the nonparametric choice of $(f_1, ..., f_G)$, $Z = \min(f_1(X_1), ..., f_G(X_G))$ can be chosen to be any real number, so we conclude that the optimal value for Z must satisfy $\text{sign}(Z) = \text{sign}(2P(Y = 1|X) - 1)$. In other words, the sign of $\min(f_1^*, ..., f_G^*)$ is the same as the Bayes rule. We thus obtain the Fisher consistency for replacing the zero-one loss function by the hinge loss function.

As a remark, although theoretically, we require f's to be fully nonparametric to obtain the Fisher consistency of using the hinge-loss, our empirical experience shows that linear rules for f's are often sufficient in terms of prediction performance (e.g., sensitivity and specificity).

2.4 Choosing between Different Approaches

The Exhaustive Search is certainly the simplest of the methods presented, and the easiest to implement. Based on the statistical theory, we also expect that it will perform quite well under suitable conditions. However, there are many scenarios where the Exhaustive Search becomes infeasible, and another option, such as SVM Iterative or Logistic Iterative becomes necessary. If the number of domains becomes quite large and/or the number of symptoms within a domain becomes large, carrying out the Exhaustive Search will become very time intensive. This is especially true if one of the domains is replaced with a continuous biomarker. In this scenario, how to even carry out an exhaustive search is not clear.

Another advantage of the iterative methods compared to the Exhaustive Search is their flexibility to incorporate item-specific weights, or to empirically determine a scoring system. The current DSM criteria sets give each symptom in a domain the same weight and simply counts the total number of symptoms in each domain. Another advantage lies in its great flexibility to include non-linear decision rules and a large number of variables within domains through the kernel trick for SVM.

As for the two iterative algorithms, both rely on cyclic coordinate descent [Luenberger, 1984], in that one set of parameters are updated while holding all of the others constant. Algorithm 1 is flexible in the sense that it can easily incorporate 'AND' relationships as well as 'OR' relationships (see 5.2). However, our current clinical application only requires 'AND' relationship. Algorithm 2 is flexible in the sense that in each iteration any appropriate classifier (e.g., SVM, logistic regression, random forest) can be used. In Algorithm 1, the solving for f_k directly depended on the current values for the other discriminant functions, $\hat{f}_1, \ldots, \hat{f}_{k-1}, \hat{f}_{k+1} \ldots, \hat{f}_G$. This required using another algorithm, the DC algorithm, to handle the non-convexity problem that results. On the other hand, in the Algorithm 2 approach, the estimate for f_k is only indirectly related to the current values for the other domains. This is through the subset of data that is being used to estimate f_k, A_k. As a result, this approach does not have a non-convexity problem since it chooses to focus within a single domain only, rather than across the domains.

Although both algorithms are reasonable approaches to solving this problem, the first algorithm is computationally much more intensive. This is because it is an iterative algorithm that contains another iterative algorithm (DC algorithm). Further, the second algorithm can be carried out using software packages for SVMs and logistic regression, while the first one requires using a quadratic programming solver. We have shown that if we know A_k^*, then this second algorithm is the Bayes classifier. For these reasons, we will only examine the performance properties of Algorithm 2 (SVM

Iterative and Logistic Iterative) in the remainder of this dissertation.

Chapter 3

Simulations

3.1 Initial Data Simulation

All of the data used to evaluate the Exhaustive Search and the iterative methods was simulated to reflect the actual data as closely as possible.

In the initial simulation setting (Setting A), we randomly selected $n = 300$ vectors of length 17 from a multivariate normal distribution with mean zero. The 17 Dimensions reflect 17 symptoms within two domains. Domain A has five items, Domain B, 12. Items within the same domain had a correlation of 0.8, items across domains a correlation of 0.65, and the variance for each item was set to one. Each item was then dichotomized, with the cut point chosen to reflect the prevalence of each item in the real data. The dichotomized items were then summed within each domain to determine CountA and CountB, the number of symptoms present in each domain.

Next, we assigned a case status to each individual; if CountA$\geq$ 1 and CountB $\geq$ 3, then the individual was classified as diseased, else as not diseased. Therefore, the true $c_1 = 1$ and $c_2 = 3$. We added random noise to either the case status, the counts, or both.

- Error Structure 1 (case status only): Randomly switched the case status for 25% of the observations that fell near the decision threshold ($0 \leq$ countA ≤ 2 and $2 \leq$ countB ≤ 4).

This updated case status was used in the simulations.

- Error Structure 2 (counts only): Randomly added one to CountA with 10% probability or randomly subtracted one to CountA with 10% probability. If the updated CountA was -1 or 6 (i.e. outside the range of Domain A), CountA was reverted to 0 or 5 respectively. The same process was repeated for CountB. Updated counts were used in the simulations.

- Error Structure 3 (counts only): Randomly added one to CountA with 15% probability or randomly subtracted one to CountA with 15% probability. If the updated CountA was -1 or 6, CountA was reverted to 0 or 5 respectively. The same process was repeated for CountB. Updated counts were used in the simulations.

- Error Structure 4 (case status and counts): Case status was manipulated in the same way as Simulation 1. Further, the counts were manipulated as in Simulation 2.

- Error Structure 5 (case status and counts): Case status was manipulated in the same way as Simulation 1, except case status was only switched 15% of the time. Further, the counts were manipulated as in Simulation 2.

1000 data sets were simulated under each of these error structures and used to train each of the competing methods. Each resulting decision rule was then evaluated on an independent test set of size $n = 10000$ with the same error structure as the original data set. It is well known that using the training data to evaluate model performance is not appropriate, as the training error does not properly account for model complexity. Training error tends to decrease whenever we increase model complexity. However, with too much fitting, the model adapts itself too closely to the training data and will not generalize well [Hastie *et al.*, 2009]. For this reason, having an independent test set is recommended. Using this test set, sensitivity, specificity, and misclassification rate were computed

for each estimated decision rule produced by each of the 1000 training sets and then averaged. In addition, the proportion of times (c_1, c_2) was correctly chosen was also computed.

3.2 Other Simulation Settings

In addition to Simulation Setting A described above, the competing methods were also evaluated under several other varying data structures (two and three factors), sample sizes ($n = 150$ and $n = 300$), correlation strengths (strong and moderate), data types (counts as well as continuous biomarkers), and with and without model misspecification. Details of each setting are given below.

Setting A: *Two factors, strong correlation among factors,* $n = 300$. See above.

Setting B: *Two factors, strong correlation among factors,* $n = 150$. Same as Setting A, except each training data set was of size 150 instead of 300.

Setting C: *Two factors, moderate correlation among factors,* $n = 300$. As in Setting A, we randomly selected $n = 300$ vectors of length 17 from a multivariate normal distribution with mean zero, except that items within the same domain had a correlation of 0.65 (rather than 0.8), and items across domains a correlation of 0.5 (instead of 0.65). The remainder of the data simulation was as in Setting A.

Setting D: *Two factors, strong correlation among factors,* $n = 300$ *with model misspecification.* Data was simulated exactly as in Setting A. However, rather than using the symptom counts based on the true data structure in the model training, counts based on an incorrect data structure were used. To be more specific, one item that truly loaded on Domain B was treated as though it belonged to Domain A. To make this as realistic as possible, the item chosen to be misspecified had a moderate cross-loading on the other domain in the real data example and therefore could have been mistakenly placed on the wrong factor. Based on this misspecification, countA could now range from 0-6 and countB from 0-11. Error structures two through five were all applied to

these misspecified counts, rather than the true counts.

Setting E: *Two factors (one count, one continuous), strong correlation among factors,* $n = 300$. Here, we randomly selected $n = 300$ vectors of length 13 from a multivariate normal distribution with mean zero. The first 12 dimensions reflect a symptom cluster with 12 symptoms and the last dimension, a continuous biomarker. Items within the symptom cluster had a correlation of 0.8. The correlation between each symptom and the continuous biomarker was 0.65. The variance for each symptom and the continuous biomarker was set to one. Each item within the symptom cluster was then dichotomized, with the cut point chosen to reflect the prevalence of each item in the real data. The dichotomized items were then summed within the domain to determine CountA. A case status was then assigned to each individual based on the number of symptoms they had (CountA) and their continuous biomarker level(totalB); if CountA≥ 2 and totalB ≥ -0.5, then the individual was classified as diseased, else as not diseased. Therefore, the true $c_1 = 2$ and $c_2 = -0.5$.

The error structures discussed above needed to be adapted to handle the continuous biomarker. In structure 1, case status was randomly switched for 25% of the observations that fell near the decision threshold ($1 \leq$ countA ≤ 3 and $-1 \leq$ totalB ≤ 0). For error structure 2, the noise added to CountA remained the same. For totalB, 20% of the time an observation from a N(0, 0.2) distribution was added to it. In error structure 3, this percentage was upped to 30%. Error structure 4 remained a combination of error structure 1 and 2, and error structure 5 a combination of error structure 1 (with a reduced probablility of switching) and error structure 2.

In addition to evaluating the sensitivity, specificity, and misclassification rate in this setting, the mean squared error was also computed based on the true value of the threshold for TotalB.

Setting F: *Three domains, strong correlation among factors,* $n = 300$. We randomly selected $n = 300$ vectors of length 17 from a multivariate normal distribution with mean zero. The 17 Dimensions reflect 17 symptoms within three domains. Domain A has nine items, Domain B six,

Domain C two. Items within the same domain had a correlation of 0.8, items across domains a correlation of 0.65, and the variance for each item was set to one. Each item was then dichotomized, with the cut point chosen to reflect the prevalence of each item in the real data. The dichotomized items were then summed within each domain to determine CountA, CountB, and CountC. Next, a case status was assigned to each individual; if CountA$\geq$ 2, CountB $\geq$ 1 and CountC $\geq$ 0, then the individual was classified as diseased, else as not diseased. Therefore, the true $c_1 = 2$, $c_2 = 1$, and $c_3 = 0$. The same error structures described above were extended to handle three counts rather than two.

Setting G: *Three domains, strong correlation among factors,* $n = 300$ *with model misspecification.* Data was simulated as in Setting F. However, rather than using the symptom counts based on the true data structure in the model training, counts based on an incorrect data structure were used. To be more specific, one item that truly loaded on Domain A was treated as though it belonged to Domain C. This item had a strong cross-loading on Domain C in the real data and therefore could have been realistically placed on Domain C by mistake. Based on this misspecification, countA could now range from 0-8, countB from 0-6, and countC from 0-3. Error structures were all applied to these misspecified counts, rather than the true counts.

Setting H: *Truth is two domains, but one domain is used instead* Here, the data are simulated exactly as in Setting A, with all of the same error structures applied. However, instead of estimating the true decision rule based on two domains, a decision rule based on one domain is estimated instead. The count used for this single domain model is just the sum of the counts observed for Domain A and Domain B (countA+countB). This represents a more severe form of model misspecication than previously examined. Previously, only one item was incorrectly placed in the wrong domain; now, the overall grouping structure is wrong. In the one domain scenario, both Logistic Iterative and SVM Iterative reduce to just performing a single regression to determine the

threshold, as there are no other domains to iterate across.

3.3 Methods Evaluated

In addition to evaluating the performance of our three newly proposed methods, Exhaustive Search, SVM Iterative, Logistic Iterative, we also evaluated four other methods: SVM Naïve, Logistic Naïve, SVM Linear, and Logistic Linear.

SVM Naïve: In the two factor settings, a linear SVM of the form $f(x_1) = b_{01} + b_{11} * x_1$, where x_1 is the symptoms count for Domain A, was fit on the full data to determine c_1. A separate linear SVM of the form $f(x_2) = b_{02} + b_{12} * x_2$, where x_2 is the symptoms count for Domain B, was fit on the full data to determine c_2. The classification rule was taken to be the intersection of these two decisions rules, (c_1, c_2).

Logistic Naïve: In the two factor settings, a logistic model of the form $\log(\frac{p}{1-p}) = b_{01} + b_{11} * x_1$, where x_1 is the symptoms count for Domain A, was fit on the full data to determine c_1. A separate logistic model of the form $\log(\frac{p}{1-p}) = b_{02} + b_{12} * x_2$, where x_2 is the symptoms count for Domain B, was fit on the full data to determine c_2. The classification rule was taken to be the intersection of these two decisions rules, (c_1, c_2).

SVM Linear: In the two factor settings, a linear model of the form $f(x) = b_0 + b_1 x_1 + b_2 x_2$ was fit using an SVM. If $f(x) \geq 0$, then the person is classified as diseased, otherwise, not diseased.

Logistic Linear: In the two factor settings, a logistic model of the form $\log(\frac{p}{1-p}) = b_0 + b_1 x_1 + b_2 x_2$ was fit. If $f(x) \geq 0$, then the person is classified as diseased, otherwise, not diseased.

Unfortunately both SVM Linear and Logistic Linear do not provide classification rules that are consistent with the logic structure of the DSM. They are explored here purely out of statistical interest.

After evaluating the performance of SVM Naïve, Logistic Naïve, SVM Linear and Logistic

Linear under all of the two factor scenarios (A-E), we decided to eliminate them as alternatives due to poor performance. For this reason, they were not evaluated in the three factor settings.

3.4 Simulation Results

In all of the tables of simulation results, there is a column labeled Oracle. This is the average sensitivity, specificity, and misclassification rate of the true classification rule when applied to the test data and can be thought of as a gold standard by which to judge the competing methods. Table 3.1 presents the results of Simulation A under each of the error structures. Under all of the error structures, the Exhaustive Search, Logistic Iterative, and SVM Iterative perform similarly in terms of average sensitivity, specificity, and misclassification rates, with rates very close to the Oracle. In terms of selecting the correct (c_1, c_2), Exhaustive Search has the highest chance, followed by Logistic Iterative, and then by SVM Iterative. The gaps between our proposed methods and the Oracle diverge more as the error structure becomes more severe, but does not seem to drastically impact the overall diagnostic measures. Both of the linear rules, which were included out of statistical interest and not clinical relevance, perform moderately worse than the previous three methods, with misclassification rates that are somewhat worse than the Oracle method. Lastly, the two naïve methods perform the worst of all, with misclassification rates greater than 10% in all cases. Neither of the naïve methods ever select the correct (c_1, c_2), always overestimating them both.

In Simulation Setting B, presented in Table 3.2, the training sets were of size $n = 150$ instead of $n = 300$. The overall pattern of results remains similar, with the Exhaustive Search, Logistic Iterative, and SVM Iterative performing well, the linear rules performing slightly worse, and the naïve methods performing the worst of all. Unlike in Setting A, where the Exhaustive Search outperformed all methods, under Error Structure 2, SVM Iterative and Logistic Iterative actually

Table 3.1: Simulation Setting A; Two domains, Strong Correlation, $n = 300$

Error		Oracle	Exhaustive Search	Linear Logistic	Naïve Logistic	Iter Logistic	Linear SVM	Naïve SVM	Iter SVM
1	(c_1, c_2)	N/A	99.4%	N/A	0.0%	99.0%	N/A	0.0%	96.3%
	avg sens	0.927	0.927	0.855	0.609	0.926	0.855	0.659	0.924
	avg spec	0.986	0.985	0.948	1.000	0.986	0.946	0.999	0.985
	avg misclass	0.035	0.035	0.084	0.137	0.035	0.085	0.121	0.036
2	(c_1, c_2)	N/A	99.8%	N/A	0.0%	95.3%	N/A	0.0%	90.9%
	avg sens	0.981	0.972	0.884	0.625	0.966	0.889	0.649	0.962
	avg spec	0.972	0.981	0.955	1.000	0.982	0.953	1.000	0.982
	avg misclass	0.022	0.022	0.069	0.126	0.023	0.068	0.118	0.025
3	(c_1, c_2)	N/A	96.9%	N/A	0.0%	85.2%	N/A	0.0%	79.2%
	avg sens	0.959	0.956	0.878	0.623	0.942	0.883	0.644	0.938
	avg spec	0.973	0.974	0.955	1.000	0.976	0.952	1.000	0.975
	avg misclass	0.031	0.032	0.071	0.126	0.035	0.071	0.120	0.037
4	(c_1, c_2)	N/A	96.6%	N/A	0.0%	87.2%	N/A	0.0%	76.5%
	avg sens	0.908	0.905	0.848	0.606	0.895	0.847	0.626	0.889
	avg spec	0.970	0.970	0.946	1.000	0.972	0.944	0.999	0.969
	avg misclass	0.052	0.052	0.088	0.139	0.055	0.089	0.132	0.059
5	(c_1, c_2)	N/A	98.3%	N/A	0.0%	91.6%	N/A	0.0%	82.3%
	avg sens	0.934	0.933	0.862	0.615	0.925	0.863	0.632	0.919
	avg spec	0.976	0.976	0.950	1.000	0.977	0.948	0.999	0.975
	avg misclass	0.039	0.039	0.081	0.133	0.041	0.081	0.127	0.044

Table 3.2: Simulation Setting B; Two domains, Strong Correlation, $n = 150$

Error		Oracle	Exhaustive Search	Linear Logistic	Naïve Logistic	Iter Logistic	Linear SVM	Naïve SVM	Iter SVM
1	(c_1, c_2)	N/A	91.8%	N/A	0.0%	89.0%	N/A	0.0%	79.4%
	avg sens	0.927	0.923	0.854	0.607	0.915	0.850	0.664	0.915
	avg spec	0.986	0.981	0.946	1.000	0.985	0.945	0.998	0.976
	avg misclass	0.035	0.039	0.086	0.138	0.039	0.087	0.119	0.046
2	(c_1, c_2)	N/A	91.5%	N/A	0.0%	83.6%	N/A	0.0%	78.1%
	avg sens	0.972	0.965	0.883	0.620	0.953	0.883	0.654	0.950
	avg spec	0.981	0.965	0.952	1.000	0.982	0.951	0.999	0.973
	avg misclass	0.022	0.0351	0.071	0.127	0.027	0.071	0.116	0.0345
3	(c_1, c_2)	N/A	82.9%	N/A	0.0%	68.8%	N/A	0.0%	70.7%
	avg sens	0.959	0.942	0.878	0.618	0.924	0.873	0.650	0.932
	avg spec	0.973	0.970	0.952	1.000	0.977	0.950	0.999	0.967
	avg misclass	0.031	0.039	0.073	0.128	0.041	0.073	0.118	0.045
4	(c_1, c_2)	N/A	78.8%	N/A	0.0%	70.1%	N/A	0.0%	64.2%
	avg sens	0.908	0.894	0.846	0.601	0.878	0.845	0.635	0.889
	avg spec	0.970	0.968	0.944	0.999	0.972	0.942	0.998	0.957
	avg misclass	0.052	0.058	0.090	0.141	0.061	0.091	0.130	0.067
5	(c_1, c_2)	N//A	87.4%	N/A	0.0%	76.6%	N/A	0.0%	70.2%
	avg sens	0.934	0.924	0.861	0.609	0.909	0.857	0.640	0.913
	avg spec	0.976	0.974	0.947	1.000	0.977	0.945	0.999	0.965
	avg misclass	0.039	0.043	0.082	0.135	0.047	0.082	0.125	0.053

do slightly better here. In addition, with a smaller sample size, the gap between the three methods of interest and the Oracle is slightly wider than it was under the larger sample size. For example, under error structure 5, where noise has been added to both the counts and the case statuses, the Oracle rule yields a misclassification rate of 0.039 on the independent test set. When the sample size is $n = 300$, the average misclassification rates for the Exhaustive Search, Logistic Iterative, and SVM Iterative are $0.039, 0.041$, and 0.044 respectively. When the sample size decreases to $n = 150$, the rates are slightly higher at $0.043, 0.047, 0.053$.

Table 3.3 presents the results if the domains are only moderately correlated, as opposed to strongly correlated in Table 3.1. Results under this scenario are nearly identical to when the correlation was strong, and therefore will not be discussed in detail.

In Simulation Setting D, the model has been misspecified. The true classification rule is based

Table 3.3: Simulation Setting C; Two domains, Moderate Correlation, $n = 300$

Error		Oracle	Exhaustive Search	Linear Logistic	Naïve Logistic	Iter Logistic	Linear SVM	Naïve SVM	Iter SVM
1	(c_1, c_2)	N/A	99.9%	N/A	0.0%	99.2%	N/A	0.0%	96.1%
	avg sens	0.915	0.915	0.837	0.527	0.914	0.842	0.587	0.911
	avg spec	0.979	0.979	0.938	1.000	0.979	0.936	0.998	0.978
	avg misclass	0.045	0.045	0.100	0.177	0.045	0.099	0.156	0.047
2	(c_1, c_2)	N/A	100.0%	N/A	0.0%	98.7%	N/A	0.0%	94.9%
	avg sens	0.961	0.961	0.875	0.544	0.959	0.885	0.582	0.954
	avg spec	0.981	0.981	0.945	1.000	0.981	0.942	0.999	0.981
	avg misclass	0.026	0.026	0.080	0.162	0.027	0.077	0.149	0.029
3	(c_1, c_2)	N/A	99.5%	N/A	0.0%	94.3%	N/A	0.0%	86.7%
	avg sens	0.949	0.948	0.871	0.542	0.941	0.880	0.572	0.932
	avg spec	0.972	0.972	0.944	1.000	0.973	0.941	0.999	0.972
	avg misclass	0.036	0.036	0.082	0.163	0.039	0.080	0.153	0.042
4	(c_1, c_2)	N/A	98.0%	N/A	0.0%	91.9%	N/A	0.0%	75.3%
	avg sens	0.887	0.886	0.832	0.521	0.879	0.835	0.554	0.869
	avg spec	0.966	0.965	0.934	0.999	0.966	0.931	0.997	0.961
	avg misclass	0.064	0.064	0.105	0.180	0.067	0.105	0.168	0.074
5	(c_1, c_2)	N/A	99.2%	N/A	0.0%	95.2%	N//A	0.0%	82.5%
	avg sens	0.918	0.917	0.854	0.533	0.912	0.858	0.565	0.899
	avg spec	0.969	0.969	0.937	0.999	0.970	0.935	0.998	0.967
	avg misclass	0.049	0.050	0.093	0.170	0.051	0.093	0.160	0.057

Table 3.4: Simulation Setting D; Two domains, Strong Correlation, $n = 300$, model misspecification

Error		Oracle	Exhaustive Search	Linear Logistic	Naïve Logistic	Iter Logistic	Linear SVM	Naïve SVM	Iter SVM
1	avg sens	0.927	0.904	0.856	0.646	0.890	0.855	0.668	0.966
	avg spec	0.986	0.971	0.948	1.000	0.972	0.947	1.000	0.966
	avg misclass	0.035	0.053	0.084	0.124	0.057	0.085	0.117	0.057
2	avg sens	0.981	0.952	0.884	0.661	0.929	0.887	0.679	0.941
	avg spec	0.972	0.971	0.955	1.000	0.975	0.954	1.000	0.968
	avg misclass	0.022	0.036	0.069	0.114	0.040	0.068	0.108	0.041
3	avg sens	0.959	0.936	0.878	0.655	0.910	0.881	0.673	0.929
	avg spec	0.973	0.967	0.955	1.000	0.973	0.953	0.999	0.964
	avg misclass	0.031	0.043	0.071	0.115	0.0481	0.071	0.110	0.0478
4	avg sens	0.908	0.882	0.848	0.640	0.865	0.848	0.651	0.879
	avg spec	0.970	0.963	0.946	0.999	0.965	0.944	0.999	0.957
	avg misclass	0.052	0.066	0.088	0.127	0.070	0.089	0.124	0.070
5	avg sens	0.934	0.910	0.862	0.648	0.892	0.865	0.659	0.908
	avg spec	0.976	0.967	0.950	1.000	0.970	0.948	0.999	0.961
	avg misclass	0.039	0.053	0.080	0.122	0.057	0.080	0.118	0.057

on two domains, one domain with five items, the other with 12. Instead of using the true counts to perform the classification, misspecified counts were used instead, where one item from Domain B was included in Domain A instead. Results are presented in Table 3.4. As there no longer exists a "true" (c_1, c_2) based on the modified counts, that row was removed from the table. The same general pattern exists with the Exhaustive Search, Logistic Iterative, and SVM Iterative all doing reasonably well, followed by the linear rules which perform okay, and then by the naïve rules which do not perform well at all. Under all scenarios, the Exhaustive Search performs the best of all, with the two iterative methods not very far behind. In addition, under all of the error structures, we see a small gap between the misclassification rate under the Oracle Rule and the Exhaustive Search for the first time. The differences in the misclassification rates under the Oracle rule and the Exhaustive Search range from 1.4 to 1.8 percentage points. In previous settings, it was generally under half a percentage point. Despite this, all three methods still perform quite well in the presence of some model misspecification.

Table 3.5 presents the results if instead of two counts, we have one count and one continuous biomarker to perform classification. Under this scenario, the Exhaustive Search becomes infeasible and is therefore not evaluated. Here, Logistic Iterative seems to perform the best, closely followed by SVM Iterative, in terms of diagnostic performance measures. Its average misclassification rates are very close to the Oracle, with rates that are just about a half a percentage point higher. In terms of selecting the correct c_1, Logistic Iterative also performs the best, correctly getting c_1 over 96% of the time. Since the second domain is now a continuous measure, mean squared error was used to determine how close each of the methods comes to the true threshold value, again with Logistic Iterative performing the best of the evaluated methods.

In Setting F, Table 3.6, we move to a three domain structure. Under all of the error structures, we see the Exhaustive Search performing closest to the Oracle rule, with almost no discrepancies. We do see a small gap between Logistic Iterative and the Exhaustive Search, and a slightly larger gap between SVM Iterative and the Exhaustive Search than was previously seen under the two factor simulation settings. When we move to a misspecified three factor model, as is the case in Table 3.7, we see the Exhaustive Search and Logistic Iterative performing similarly, with the SVM Iterative performing slightly worse. As was the case with the misspecified two factor model (Table 3.4), there is a larger gap in the misclassification rates for the Exhaustive Search compared to the Oracle rule than was seen when using the correctly specified model. Here the difference ranges from 2.3 percentage points to 2.7 percentage points, where previously with the two factor model the gap was between 1.4 to 1.8 percentage points. In general, the misclassification rates (both of the Oracle rule and the evaluated methods) were higher for the three factor settings than they were for the two factor settings, which makes sense as the three factor structure is more complicated.

In the last simulation setting, Setting H (Table 3.8), the true decision rule is based on two domains (as was the case in Setting A). However, all three methods are applied to come up with

Table 3.5: Simulation Setting E; Two domains (one count, one continuous), Strong Correlation, $n = 300$

Error		Oracle	Linear Logistic	Naïve Logistic	Iter Logistic	Linear SVM	Naïve SVM	Iter SVM
1	(c_1)	N/A	N/A	15.4%	99.8%	N/A	55.5%	98.2%
	MSE	N/A	N/A	0.483	0.005	N/A	0.444	0.025
	avg sens	0.963	0.906	0.677	0.959	0.915	0.722	0.957
	avg spec	0.984	0.951	1.000	0.980	0.949	1.000	0.976
	avg misclass	0.025	0.069	0.141	0.029	0.066	0.122	0.033
2	(c_1)	N/A	N/A	5.4%	98.3%	N/A	51.6%	95.5%
	MSE	N/A	N/A	0.520	0.008	N/A	0.471	0.028
	avg sens	0.978	0.906	0.667	0.972	0.918	0.714	0.969
	avg spec	0.978	0.951	0.999	0.972	0.948	0.995	0.969
	avg misclass	0.022	0.068	0.144	0.028	0.065	0.126	0.031
3	(c_1)	N/A	N/A	3.8%	97.1%	N/A	44.4%	92.5%
	MSE	N/A	N/A	0.522	0.011	N/A	0.472	0.033
	avg sens	0.967	0.901	0.665	0.960	0.911	0.707	0.950
	avg spec	0.970	0.949	0.999	0.965	0.946	0.994	0.961
	avg misclass	0.031	0.071	0.144	0.037	0.069	0.130	0.044
4	(c_1)	N/A	N/A	8.4%	96.9%	N/A	46.8%	90.9%
	MSE	N/A	N/A	0.486	0.012	N/A	0.449	0.051
	avg sens	0.946	0.892	0.667	0.940	0.898	0.705	0.924
	avg spec	0.966	0.945	0.999	0.960	0.943	0.995	0.955
	avg misclass	0.043	0.078	0.146	0.049	0.077	0.132	0.059
5	(c_1)	N/A	N/A	7.2%	98.1%	N/A	47.0%	92.5%
	MSE	N/A	N/A	0.500	0.010	N/A	0.457	0.026
	avg sens	0.958	0.896	0.666	0.952	0.904	0.707	0.942
	avg spec	0.971	0.948	0.999	0.965	0.945	0.995	0.961
	avg misclass	0.035	0.075	0.146	0.040	0.073	0.131	0.047

Table 3.6: Simulation Setting F; Three domains, Strong Correlation, $n = 300$

Error Structure		Oracle	Exhaustive Search	Iterative Logistic	Iterative SVM
1	(c_1, c_2, c_3)	N/A	99.9%	98.6%	95.3%
	avg sens	0.904	0.904	0.901	0.893
	avg spec	0.979	0.979	0.979	0.976
	avg misclass	0.050	0.050	0.052	0.057
2	(c_1, c_2, c_3)	N/A	99.7%	92.1%	86.1%
	avg sens	0.969	0.969	0.948	0.934
	avg spec	0.976	0.976	0.977	0.963
	avg misclass	0.027	0.027	0.034	0.048
3	(c_1, c_2, c_3)	N/A	96.1%	74.8%	79.2%
	avg sens	0.954	0.950	0.904	0.915
	avg spec	0.965	0.966	0.970	0.951
	avg misclass	0.039	0.040	0.054	0.062
4	(c_1, c_2, c_3)	N/A	97.7%	88.7%	78.0%
	avg sens	0.887	0.884	0.866	0.854
	avg spec	0.962	0.962	0.964	0.950
	avg misclass	0.068	0.069	0.075	0.088
5	(c_1, c_2, c_3)	N/A	98.7%	90.2%	81.5%
	avg sens	0.923	0.922	0.904	0.888
	avg spec	0.967	0.967	0.969	0.956
	avg misclass	0.050	0.050	0.056	0.070

Table 3.7: Simulation Setting G; Three domains, Strong Correlation, $n = 300$, model misspecification

Error Structure		Oracle	Exhaustive Search	Iterative Logistic	Iterative SVM
1	avg sens	0.904	0.862	0.845	0.874
	avg spec	0.979	0.965	0.973	0.946
	avg misclass	0.050	0.076	0.078	0.083
2	avg sens	0.969	0.902	0.885	0.909
	avg spec	0.976	0.973	0.976	0.946
	avg misclass	0.027	0.053	0.057	0.068
3	avg sens	0.954	0.884	0.868	0.886
	avg spec	0.965	0.969	0.969	0.937
	avg misclass	0.039	0.062	0.068	0.082
4	avg sens	0.887	0.833	0.816	0.848
	avg spec	0.962	0.959	0.966	0.922
	avg misclass	0.068	0.091	0.093	0.107
5	avg sens	0.923	0.864	0.845	0.881
	avg spec	0.967	0.965	0.970	0.929
	avg misclass	0.050	0.074	0.078	0.090

a threshold based on a single domain. The count used for this single domain model is just the sum of the counts observed for Domain A and Domain B in the true model. This represents a more severe form of model misspecication than previously examined. Previously, only one item was incorrectly placed in the wrong domain; now, the overall grouping structure is wrong. Under all three scenarios, the methods perform nearly identical to one another. This simulation, when compared to Simulation A, also illustrates that there is much to be gained by knowing the true number of domains. In Simulation A, when the number of domains is correct, all three methods come very close to using the oracle decision rule. When one factor is used incorrectly, there is a discrepancy between what the methods are able to obtain in terms of diagnostic performance when compared to the oracle. Despite this, under the wrong domain structure, all of the misclassification errors remain under 10%, suggesting some robustness to choosing the wrong number of domains.

Table 3.8: Simulation Setting H; Model misspecification - One domain considered when the truth is two domains

Error Structure		Oracle	Exhaustive Search	Logistic	SVM
1	avg sens	0.927	0.827	0.852	0.845
	avg spec	0.986	0.958	0.947	0.948
	avg misclass	0.035	0.088	0.086	0.088
2	avg sens	0.981	0.882	0.883	0.884
	avg spec	0.972	0.949	0.952	0.951
	avg misclass	0.022	0.073	0.071	0.071
3	avg sens	0.959	0.877	0.877	0.877
	avg spec	0.973	0.949	0.953	0.953
	avg misclass	0.031	0.075	0.073	0.073
4	avg sens	0.908	0.823	0.845	0.839
	avg spec	0.970	0.952	0.943	0.945
	avg misclass	0.052	0.093	0.091	0.093
5	avg sens	0.934	0.846	0.860	0.856
	avg spec	0.976	0.951	0.947	0.948
	avg misclass	0.039	0.085	0.083	0.084

3.5 Practical Considerations

In implementing the above algorithm on the simulated data, a problem arose especially when the sample size was small or the number of factors increased. At some points in the iterative process, the number of controls would become very small or even zero. In these scenarios, both the SVM and logistic regression failed to converge or could not be estimated at all. From the perspective of the clinical problem, this essentially meant that once we knew the symptom count(s) for the other domain(s), the current domain was not necessary in terms of classification. Therefore, the threshold for this domain should be set to zero. To incorporate this into the algorithm, at each step in the iteration, if the number of controls was below five, instead of running an SVM or logistic regression, the slope and intercept for that domain at that iteration was set to zero and one respectively and it moved to the next step in the procedure. If the number of controls was greater than five, then

the algorithm proceeded as normal.

In terms of time needed to execute each of the methods, Logistic Iterative was always the fastest, followed by the Exhaustive Search. In the more complicated three-factor structures, Logistic Iterative was almost always twice as quick as the Exhaustive Search. SVM Iterative took longer than Logistic Iterative since it required tuning a cost parameter (5-fold cross-validation) at each step. However, we have found that the time necessary to carry out SVM Iterative can be drastically reduced without any effect on the diagnostic performance. Instead of tuning at each iteration, tuning only needs to be carried out in the first iteration. The tuning parameters selected in the first step are then used in the remainder of the iterations. As a result, computing time for SVM Iterative was reduced $55\% - 78\%$.

Chapter 4

Real Data Application

4.1 Applications to Pittsburgh and CGTOA Studies: Model Training

The data to which each of the three methods, Exhaustive Search, SVM Iterative and Logistic Iterative, were applied is the same data set as was used in [Simon *et al.*, 2011]. This sample was comprised of bereaved healthy controls (n=95), patients diagnosed as having either a mood or anxiety disorder (n=369), and patients presenting for treatment of CG (n=318). All participants completed the 19-item Inventory of Complicated Grief [Prigerson *et al.*, 1995], a well validated self-report measure of CG symptom severity with prior evidence for high internal consistency (Cronbachs $\alpha = 0.94$) and test-retest reliability (0.80). The ICG assesses a range of CG symptoms including preoccupation with the person who died, intrusive and distressing thoughts related to the death, avoidance of reminders of the person who died, feelings of yearning for the person who died, loneliness, and feelings of bitterness, anger and/or disbelief regarding the death. Each item is rated on a 5-point scale, with responses rated as occurring either 0=never, 1=rarely, 2=sometimes, 3=often, or 4=always. Among those presenting for treatment of CG, only those who scored at least a 30 on the ICG and were also diagnosed with CG on clinical interview were considered to be cases (n=288).

These cases participated in the Pittsburgh and CGTOA studies, discussed in more detail in the Introduction. Noncases were defined as bereaved individuals who did not present with CG as their primary diagnosis and scored less than a 25 on the ICG (n=377). This resulted in a total sample size of n=665.

4.1.1 Determining the Domain Structure

The same methodology that was used to determine the six factor structure of the ICG in [Simon *et al.*, 2011] was used here, with the exception that we instead examined two and three factor structures in order to determine our symptom domains. First, all ICG items were dichotomized with "often" or "always" being treated as the symptom was "present" and other categories as "absent". Next a two factor and a three factor exploratory factor analysis model was fit on the data for **cases only**. Robust weighted least squares (WLSMV in Mplus version 7, [Muthën and Muthën, 2012]) and geomin orthogonal rotation were used for both models. An item was considered as loading on a factor if its loading was 0.3 or greater in magnitude. If an item loaded on multiple factors, it was assigned to the factor with the highest loading or to the factor with which it made the most conceptual sense. Results of the two factor model are presented in Table 4.1 and the three factor model in Table 4.2.

In the two factor model, items 3, 6, 7, 8, and 17 loaded together into one symptom domain and items 1, 2, 4, 9, 10, 11, 13, 14, 15, 16, 18, 19 loaded together into another symptom domain. Items 5 (drawn to places/things) and 12 (avoid reminders) did not load on either factor and therefore were dropped from the analyses. To summarize, the first domain has five symptoms and the second, twelve. Domain 1 seems to represent reactions to the death (denial, anger, disbelief, stunned, bitter, etc.) while Domain 2 represents other CG symptoms. The goal is now to estimate how many items from Domain 1 (out of five) and how many items from Domain 2 (out of 12) should be required in

Table 4.1: Two Factor EFA Model

Item	Description	Factor 1	Factor 2
item 1	think about person	0.189	**0.513**
item 2	Memories upset me	0.080	**0.432**
item 3	cannot accept death	**0.572**	0.205
item 4	longing for person	0.131	**0.546**
item 5	drawn to places/things	0.139	0.268
item 6	feeling angry	**0.780**	-0.023
item 7	disbelief	**0.927**	-0.011
item 8	stunned or dazed	**0.844**	0.023
item 9	hard to trust	-0.189	**0.471**
item 10	lost ability care about others	-0.222	**0.711**
item 11	pain in the same area	-0.038	**0.390**
item 12	avoid reminders	0.147	0.215
item 13	life is empty	0.006	**0.774**
item 14	hear the voice of the person	0.073	**0.523**
item 15	see the person	0.200	**0.463**
item 16	unfair that I should live	0.259	**0.499**
item 17	bitter over death	**0.687**	0.088
item 18	envious of others	0.040	**0.346**
item 19	lonely a great deal	-0.128	**0.903**

Table 4.2: Three Factor EFA Model

Item	Description	Factor 1	Factor 2	Factor 3
item 1	think about person	0.243	**0.484**	0.019
item 2	Memories upset me	0.193	**0.316**	0.206
item 3	cannot accept death	**0.621**	0.159	-0.298
item 4	longing for person	0.009	**0.732**	-0.349
item 5	drawn to places/things	0.134	**0.301**	-0.093
item 6	feeling angry	**0.949**	-0.348	0.005
item 7	disbelief	**0.912**	-0.002	-0.705
item 8	stunned or dazed	**0.832**	0.008	-0.530
item 9	hard to trust	0.045	0.132	**0.711**
item 10	lost ability care about others	0.006	0.437	**0.771**
item 11	pain in the same area	0.055	0.292	0.269
item 12	avoid reminders	**0.312**	0.001	0.270
item 13	life is empty	-0.015	**0.833**	0.006
item 14	hear the voice of the person	0.078	**0.552**	0.023
item 15	see the person	0.195	**0.519**	-0.134
item 16	unfair that I should live	0.311	**0.467**	-0.028
item 17	bitter over death	**0.821**	-0.204	0.144
item 18	envious of others	0.099	0.286	0.145
item 19	lonely a great deal	-0.090	**0.905**	0.282

order to be diagnosed with CG.

In the three factor model, the item loadings were as follows: *Domain* 1 items 1, 2, 4, 5, 13, 14, 15, 16, 19; *Domain* 2 items 3, 6, 7, 8, 12, and 17; *Domain 3* items 9 and 10. Items 11 (pain in the same area as the deceased) and 18 (envious of others) did not load on either factor and therefore were dropped from the analyses. To summarize, the first domain has nine symptoms, the second six, and the third two. Here Domain 2 is very similar to Domain 1 in the two factor model and

Table 4.3: Diagnostic Performance of Derived Criteria Sets

	2 Domains	3 Domains
	All Methods	All Methods
sensitivity	0.9167	0.9097
specificity	0.9920	0.9894
misclassification rate	0.0406	0.0451

seems to represent reactions to the death. Domain 3, comprised of "9 = hard to trust others" and "10 = lost ability to care about others," seems to represent some sort of social impairment. Finally, Domain 1 seems to represent other CG symptoms. As is the case with two domains, our goal is to now estimate how many items to require from each domain in order to have a positive CG diagnosis.

4.1.2 Applying the proposed methods

Using the domain structure in Section 4.1.1, we can apply each of our methods to our full data (cases and non-cases) to determine our thresholds, c_k. This requires computing the domain counts based on the individual item endorsements. In the case of two domains, all three methods choose the same thresholds, $c_1 = 1, c_2 = 2$. That is, 1 item out of 5 should be required from Domain 1, and 2 items out of 12 from Domain 2 for a positive CG diagnosis. On the training data, this results in a sensitivity of 0.9167, specificity of 0.9920, and misclassification rate of 0.0406.

For the three factor model, again all three methods choose the same thresholds $c_1 = 2, c_2 = 1, c_3 = 0$. The threshold $c_3 = 0$ implies that the third domain (social impairment) is not necessary for disease diagnosis and can therefore be dropped. These thresholds result in a sensitivity of 0.9097, specificity of 0.9894, and misclassification rate of 0.0451. The diagnostic performance of both the derived criteria sets are summarized in Table 4.3.

So far, the diagnostic performance of each of the methods on both of the factor structures is

examined on the same data that was used to fit the model, our training data. In Section 4.2, we will evaluate these derived criteria sets on an independent test set as a validation study. This same data will also provide us with the first opportunity to evaluate and compare some of the previously proposed criteria sets for CG.

The data we used to perform these analyses was essentially collected under a case-control design: cases were collected from two clinical studies of treatment seeking CG population, and controls were collected from a bereaved population. Under this design, the proposed methods may optimize diagnostic performance (e.g., sensitivity and specificity) of criteria sets under a different weighting scheme than if we used data collected from the general population. To illustrate this point, consider the empirical loss function in (2.2) under the zero-one loss and prediction rule $h(x_i.) = \text{sign}(f(x_i.))$, which is

$$\frac{1}{n}\sum_{i=1}^{n}\left\{I(y_i = 1)I(f(x_i.) < 0) + I(y_i = -1)I(f(x_i.) \geq 0)\right\}.$$

Taking the expectation leads to the following working case-control population loss function from which the case-control samples are drawn:

$$E(f(X_i.) < 0|Y_i = 1)P^*(Y_i = 1) + E\{f(X_i.) \geq 0|Y_i = -1\}P^*(Y_i = -1). \tag{4.1}$$

Under the case-control design, the above conditional expectation given case or control status can be consistently estimated from data. However, the sample proportions of cases and controls do not converge to the true population prevalence in the general population. We use $P^*(Y = 1)$ and $P^*(Y = -1)$ to denote the limit of the sample proportions which depend on the case-control design. Viewing $P^*(Y = 1)$ as a cost parameter c for the objective function, it is clear that this objective function (4.1) minimizes a weighted average of false negatives (1-sensitivity) and false positives (1-specificity),

$$cFN + (1 - c)FP. \tag{4.2}$$

If the sample has an approximately 1:1 case-control ratio, $c = 0.5$, sensitivity and specificity will receive equal weights. In our application, $c = 288/(288 + 377) = 0.43$. The thresholds c_1 and c_2 estimated by the proposed methods using this data maximize a weighted average of sensitivity and specificity with a weight of 0.43 given to the former and 0.57 to the latter. In contrast, when using data collected from a general population of subjects with major bereavements to construct criteria sets, the estimated thresholds minimize the empirical version of (4.2) with c equal to the population prevalence of CG, that is, $c = 6.7\%$ [Kersting *et al.*, 2011]. A reasonable choice of the cost parameter depends on the target population to which the criteria set is intended to be applied.

To investigate the Bayes rule under the case-control design, define E^* as the expectation of (X, Y) under the working population. The expected loss in (4.1) is then $E^*\{I(Y_i h(X_\cdot) < 0)\}$. The Bayes rule minimizing this working population loss function is

$$\text{sign}\{P^*(Y = 1|X) - 1/2\}.$$

Since

$$\begin{aligned} P^*(Y = 1|X) &= \left[\int f(x|Y = 1)ds\right] P^*(Y = 1)/f(X) \\ &= \left\{\left[\int f(x|Y = 1)ds\right] P(Y = 1)P^*(Y = 1)/P(Y = 1)\right\}/f(X) \\ &= P(Y = 1|X)\frac{P^*(Y = 1)}{P(Y = 1)}, \end{aligned}$$

the Bayes rule minimizing the working case-control population is

$$\text{sign}\left\{P(Y = 1|X) - \frac{1}{2}\frac{P(Y = 1)}{P^*(Y = 1)}\right\}.$$

Again if the sample has an approximately 1:1 case-control ratio, the Bayes rule is

$$\text{sign}\left\{P(Y = 1|X) - P(Y = 1)\right\}.$$

In other words, the Bayes rule classifies a subject as a case if the conditional probability of being a case given covariates X is greater than the probability of being a case in the general population (prevalence of case in the population).

When the criteria set is intended to be used in a general bereaved population but the data is collected from a case-control design, an adjustment to the objective function is needed to estimate thresholds that will have good performance in the general population. The key step is to modify the empirical loss function being minimized in (2.2) using sampling weights to reflect the case-control design, when an estimated prevalence of CG in the general population is available. The prevalence of CG in individuals with major bereavements was estimated as 6.7% [Kersting *et al.*, 2011]. For Exhaustive Search, the empirical misclassification rate will be adjusted using this CG prevalence instead of the observed CG sample proportion. Logistic Iterative and SVM Iterative will be adjusted by giving sampling weights to each subject. Since there were 288 cases and 377 controls, the sampling proportions for cases and controls were, $288/(6.7\%N)$ and $377/(93.3\%N)$ respectively, with N denoting the total number of subjects in the population. The sampling weights can then be computed as $0.094 = 377/93.3/(288/6.7\%)$ for cases, and one for controls.

With these adjustments, the updated results on the training data are reported in the first few columns of Table 4.4. It can be seen that Exhaustive Search and Logistic Iterative are more sensitive to the choice of weights, while the results for SVM Iterative remain the same. Both Exhaustive Search and Logistic Iterative have lower sensitivity and higher specificity compared to the results in Table 4.3, where we did not adjust for the population prevalence, which is as expected. The misclassification error rate estimated in the general bereaved population is low for all three methods. However, it is dominated by the specificity due to the much higher prevalence of non-CG subjects. One of the possible reasons that the estimate for SVM Iterative did not change might be due to the fact that the specificity was already almost one without the re-weighting. The sensitivity on

Table 4.4: Diagnostic Performance of Derived Criteria Sets (weighted by estimated population prevalence), 2 domains

	Training Data					Validation Data
Method	$c1$	$c2$	sens	spec	misclass*	sens
Exhaustive	2	2	0.8403	1.0000	0.0107	0.7640
Logistic Iterative	1	3	0.8750	0.9947	0.0133	0.8539
SVM Iterative	1	2	0.9167	0.9920	0.0130	0.9160

*: Misclassification rates are adjusted by prevalence of CG.

the validation data is discussed in the next section.

4.2 Applications to CGTOA and HEAL Data: A Validation Study

4.2.1 Study Participants

Participants in our validation study were $n = 178$ individuals who were assessed for participation in one of two NIMH-funded treatment studies of CG; these studies were either the CGTOA or HEAL studies, both discussed in the Introduction. Participants scored a 30 or higher on the ICG and were confirmed by the study PI or his or her delegate to have CG on clinical interview. This interview established prolonged acute grief symptoms accompanied by complicating dysfunctional thoughts, feelings or behaviors. In addition, all participants in the current analysis reported on a structured interview, significant functional impairment from grief symptoms and all were bereaved for at least twelve months. Of note, functional impairment is required by all three published criteria sets. Further, the [Prigerson *et al.*, 2009] and [Shear *et al.*, 2011] criteria sets both require the death to be at least six months ago, while the [American Psychiatric Association, 2013] criteria set requires at least a 12 month period. By limiting our sample to just those who have functional impairment and deaths at least 12 months prior, we are ensuring that any differences observed

among the criteria sets in terms of performance are actually due to the differences that exist among the Domain B and Domain C components in the criteria sets. In addition, our proposed methods only target the symptom domains components of the criteria set, so that is our primary interest in these analyses.

4.2.2 Assessments and Methods

All participants in this validation sample completed both the ICG [Prigerson *et al.*, 1995] and the SCI-CG [Bui *et al.*, 2015]. Both of these measures are discussed in more detail in Section 1.2.2. Each of the published criteria sets will be evaluated using the SCI-CG, which was specifically designed to capture symptoms present on any of them. Each of our proposed criteria sets (2 domain model and 3 domain model) will be evaluated using the ICG, the same measure that was used to derive them.

Since data is only available on confirmed CG cases, only sensitivity can be assessed in this validation study. Fortunately, we will soon be able to evaluate specificity as well, as a sample of treatment seeking controls was recently recruited to complete both the ICG and SCI-CG. The data collection process has recently finished and we are currently in the data cleaning process.

In order to assess sensitivity for each of the published criteria sets, each symptom from each of the criteria sets was matched with a corresponding item from the SCI-CG and rated as present (score = 3) or absent (score= 1 or 2) . In the case of two or more matching SCI-CG items for a particular symptom, the symptom was considered present if ANY of the matching SCI-CG items were endorsed. The classification algorithms provided by each of the criteria sets were then used to determine if each individual was considered a “case under the respective criteria set:

- PGD: Yearning endorsed (Domain B, separation distress) and at least 5 symptoms out of 9 from Domain C (Cognitive, emotional and behavioral symptoms)

- CG: At least 1 out of 4 symptoms from Domain B (separation distress) and at least 2 out of 8 from Domain C (other grief symptoms)
- PCBD: At least 1 out of 4 from Domain B and at least 6 out of 12 from Domain C.

The count and percentage of participants endorsing each individual symptom as well as meeting the overall domain criterion will be used to compare and evaluate the criteria sets. Further, sensitivity of each the criteria sets will be computed as the total number of people meeting that criterion's classification rule divided by the total number in the CG sample (n=178).

4.2.3 Results

A list of SCI-CG items used in the matching for each of the three published criteria sets is given in the Appendix.

Tables 4.5 - 4.7 provide the matching between each of the published criteria set symptoms and the SCI-CG. In addition, they presents the count and percentage of participants endorsing each symptom, meeting the domain specific criteria, as well as meeting the overall criteria. As all participants in this sample are cases, the percent meeting the overall criteria is actually the sensitivity.

The sensitivity of the PGD criteria set was 56.2% (95% CI: 48.9% − 63.5%). Yearning, a necessary symptom for this criteria set, was only endorsed by 87.6% of the CG sample, evidence that yearning may not be a necessary symptom. Further, only 64.0% of individuals with CG met the required five or more symptoms from Domain C, suggesting that the threshold of five or more is too high. Individual symptom endorsements ranged from 41.0% to 76.4%.

The sensitivity of the CG criteria set was 99.4% (95% CI: 98.3% − 100.0%). All but one individual had at least one symptom from Domain B (separation distress) and all individuals had

Table 4.5: Diagnostic Performance of PGD Criteria Set

	SCI-CG Item Match	Number Endorsing	Percent Endorsing
OVERALL SENSITIVITY		100	56.20%
B. Separation Distress: The bereaved person experiences yearning (e.g. craving, pining, or longing for the deceased; physical or emotional suffering as a result of the desired but unfulfilled reunion with the deceased) daily or to a disabling degree	2	156	87.60%
C. Cognitive, emotional and behavioral symptoms: The bereaved person must have **five or more** of the following symptoms experienced daily or to a disabling degree:		114	64.00%
1. Confusion about one's role in life or diminished sense of self (i.e. feeling that a part of oneself has died)	30	103	57.90%
2. Difficulty accepting the loss	7	98	55.10%
3. Avoidance of reminders of the reality of the loss	14	124	69.70%
4. Inability to trust others since the loss	24	73	41.00%
5. Bitterness or anger related to the loss	11	136	76.40%
6. Difficulty moving on with life (e.g. making new friends, pursuing new interests)	31	106	59.60%
7. Numbness (absence of emotion) since the loss	9	98	55.10%
8. Feeling that life is unfulfilling, empty or meaningless since the loss	28	104	58.40%
9. Feeling stunned, dazed or shocked by the loss	8	81	45.50%

at least two symptoms from Domain C. Separation distress symptoms ranged in frequency from 24.7% to 87.6%. Symptoms from Domain C ranged in frequency from 24.2% to 91.6%.

The sensitivity of the PCBD criteria set was 67.4% (95% CI: $60.5\% - 74.3\%$). 94.4% of the CG sample had at least one symptom from Domain B, while only 70.8% had at least six symptoms from Domain C. The poor overall sensitivity is possibly due to the threshold for Domain C being too high. Domain B symptoms ranged in frequency from 65.2% to 87.6%. Symptoms from Domain C ranged in frequency from 17.4% to 76.4%.

The performance results of our proposed two domain criteria set are given in Table 4.8. The sensitivity was 91.6% (95% CI: $87.5\% - 95.7\%$). 92.1% of the CG sample had at least one symptom from Domain A, while 98.9% had at least two symptoms from Domain C. Domain A symptoms

Table 4.6: Diagnostic Performance of CG Criteria Set

	SCI-CG Item Match	Number Endorsing	Percent Endorsing
OVERALL SENSITIVITY		177	99.40%
B. **At least one** of the following symptoms of persistent intense acute grief has been present for a period longer than is expected by others in the persons social or cultural environment		177	99.40%
1. Persistent intense yearning or longing for the person who died	2	156	87.60%
2. Frequent intense feelings of loneliness or like life is empty or meaningless without the person who died	26, 28	155	87.10%
3. Recurrent thoughts that it is unfair, meaningless, or unbearable to have to live when a loved one has died, or a recurrent urge to die in order to find or to join the deceased	22, 23	44	24.70%
4. Frequent preoccupying thoughts about the person who died	4, 5	142	79.80%
C: **At least two** of the following symptoms are present for at least a month		178	100.00%
1. Frequent troubling rumination about circumstances or consequences of the death	6, 12, 13	160	89.90%
2. Recurrent feeling of disbelief or inability to accept the death, like the person cannot believe or accept that their loved one is really gone	7	98	55.10%
3. Persistent feeling of being shocked, stunned, dazed or emotionally numb since the death	8, 9	129	72.50%
4. Recurrent feelings of anger or bitterness related to the death	11	136	76.40%
5. Persistent difficulty trusting or caring about other people or feeling intensely envious of others who have not experienced a similar loss	24, 25, 27	146	82.00%
6. Frequently experiencing pain or other symptoms that the deceased person had, or hearing the voice or seeing the deceased person	20, 21	43	24.20%
7. Experiencing intense emotional or physiological reactivity to memories of the person who died or to reminders of the loss	16, 17	136	76.40%
8. Change in behavior due to excessive avoidance or the opposite, excessive proximity seeking.	14, 15, 18, 19	163	91.60%

Table 4.7: Diagnostic Performance of PCBD Criteria Set

	SCI-CG Item Match	Number Endorsing	Percent Endorsing
OVERALL SENSITIVITY		120	67.40%
B: Since the death, **at least 1** symptom experienced on more days than not and persisted to		168	94.40%
1. Persistent Yearning/longing for deceased	2	156	87.60%
2. Intense sorrow and emotional pain in response to death	3	154	86.50%
3. Preoccupation with the deceased	4	130	73.00%
4. Preoccupation with the circumstances of the death	6	116	65.20%
C: Since the death, **at least 6** of following symptoms experienced more days than not, and have persisted for at least 12 months		126	70.80%
1. Marked difficulty accepting death.	7	98	55.10%
2. Experiencing disbelief or emotional numbness	8, 9	129	72.50%
3. Difficulty with positive reminiscing about the deceased	10	37	20.80%
4. Bitterness of anger related to death	11	136	76.40%
5. Maladaptive appraisals about oneself in relation to the deceased or the death (e.g. self-blame)	12	114	64.00%
6. Excessive avoidance of reminders of the loss	14	124	69.70%
7.A desire to die in order to be with the deceased	22	31	17.40%
8. Difficulty trusting other individuals since the death	24	73	41.00%
9. Feeling alone or detached from other individuals since the death	25	121	68.00%
10. Feeling that life is meaningless or empty without the deceased, or the belief that one cannot function without the deceased	28	104	58.40%
11. Confusion about one's role in life, or a diminished sense of one's identity	30	103	57.90%
12. Difficulty or reluctance to pursue interests since the loss or to plan for the future	31	106	59.60%

ranged in frequency from 53.4% to 66.9%. Symptoms from Domain B ranged in frequency from 10.1% to 83.2%.

The performance results of our proposed three domain criteria set are given in Table 4.9. The sensitivity of our proposed 3 domain criteria set was 87.1% (95% CI: 82.2% − 92.0%). 93.8% of the CG sample had at least two symptoms from Domain A, while 92.1% had at least one symptom from Domain C. Since no symptoms were required from Domain C, it does not impact overall sensitivity, but is included for the sake of completeness. In actuality, this is essentially another two domain criteria set. Domain A symptoms ranged in frequency from 10.1% to 83.2%. Symptoms from Domain B ranged in frequency from 34.3% to 66.9%.

Table 4.10 provides a summary of the sensitivities, with 95% Confidence Intervals, for all of the five criteria sets. The [Prigerson *et al.*, 2009] criteria set has the lowest sensitivity, closely followed by the DSM-5 [American Psychiatric Association, 2013] criteria set. Both of these sensitivities are surprisingly low, considering the subset of patients being examined here. In order to be included in this sample, not only did the individuals need to have an ICG score of 30 or higher, they also were confirmed to have CG by an expert as well as have functional impairment as a direct result of their grief symptoms. Based on this, we would expect to see sensitivities that are almost perfect. From this respect, the [Shear *et al.*, 2011] criteria set performs best of all with an almost perfect sensitivity. Both of our proposed methods do reasonably well, with the 2 domain structure doing slightly better with a sensitivity of 91.6%.

In the previous section, we performed an analysis on the training data when the aim is instead to minimize the misclassification rate in the general bereaved population. Recall this corresponds to minimizing (4.2) with $c = P(CG) = 6.7\%$ in the major bereaved population. This criterion down-weights the role of sensitivity and focuses on specificity. The sensitivity of the criteria set obtained from each of the three methods evaluated on the validation data is summarized in the

Table 4.8: Diagnostic Performance of Proposed 2 Domain Criteria Set

ICG Item	Number Endorsing	Percent Endorsing
OVERALL SENSITIVITY	163	91.57%
Domain A: **At least 1** symptom experienced	164	92.13%
3. I feel I cannot accept the death of "name of loved one"...	98	55.06%
6. I can't help feeling angry about "name of loved one" death...	110	61.80%
7. I feel disbelief over what happened...	119	66.85%
8. I feel stunned or dazed over what happened...	107	60.11%
17. I feel bitter over "name of loved one" death...	95	53.37%
Domain B: **At least 2** symptoms experienced	176	98.88%
1. I think about "name of loved one" so much that its hard for me to do the things I normally do...	104	58.43%
2. Memories of "name of loved one" upset me...	134	75.28%
4. I feel myself longing for "name of loved one"...	147	82.58%
9. Ever since "name of loved one" died it is hard for me to trust people...	80	44.94%
10. Ever since "name of loved one" died I feel like I have lost the ability to care about other people or feel distant from people I care about...	108	60.67%
11. I have pain in the same area of my body or have some of the same symptoms as "name of loved one"...	23	12.92%
13. I feel that life is empty without "name of loved one"...	140	78.65%
14. I hear the voice of "name of loved one" speak to me...	22	12.36%
15. I see "name of loved one" stand before me...	18	10.11%
16. I feel that it is unfair that I should live when "name of loved one" died...	52	29.21%
18. I feel envious of others who have not lost someone close...	79	44.38%
19. I feel lonely a great deal of the time ever since "name of loved one died...	148	83.15%

Table 4.9: Diagnostic Performance of Proposed 3 Domain Criteria Set

ICG Item	Number Endorsing	Percent Endorsing
OVERALL SENSITIVITY	155	87.08%
Domain A: **At least 2** symptoms experienced	167	93.82%
1. I think about "name of loved one" so much that its hard for me to do the things I normally do...	104	58.43%
2. Memories of "name of loved one" upset me...	134	75.28%
4. I feel myself longing for "name of loved one"...	147	82.58%
5. I feel drawn to places and things associated with "name of loved one"...	72	40.45%
13. I feel that life is empty without "name of loved one"...	140	78.65%
14. I hear the voice of "name of loved one" speak to me...	22	12.36%
15. I see "name of loved one" stand before me...	18	10.11%
16. I feel that it is unfair that I should live when "name of loved one" died...	52	29.21%
19. I feel lonely a great deal of the time ever since "name of loved one died...	148	83.15%
Domain B: **At least 1** symptom experienced	164	92.13%
3. I feel I cannot accept the death of "name of loved one"...	98	55.06%
6. I can't help feeling angry about "name of loved one" death...	110	61.80%
7. I feel disbelief over what happened...	119	66.85%
8. I feel stunned or dazed over what happened...	107	60.11%
12. I go out of my way to avoid reminders of "name of loved one"...	61	34.27%
17. I feel bitter over "name of loved one" death...	95	53.37%
Domain C: **NO** symptoms required	178	100.00%
9. Ever since "name of loved one" died it is hard for me to trust people...	80	44.94%
10. Ever since "name of loved one" died I feel like I have lost the ability to care about other people or feel distant from people I care about...	108	60.67%

Table 4.10: Comparison of Criteria Sets

Criteria Set	Sensitivity	95% Low	95% Upp
PGD [Prigerson *et al.*, 2009]	56.2%	48.9%	63.5%
CG [Shear *et al.*, 2011]	99.4%	98.3%	100.0%
PCBD [American Psychiatric Association, 2013]	67.4%	60.5%	74.3%
Proposed 2 Domain	91.6%	87.5%	95.7%
Proposed 3 Domain	87.1%	82.2%	92.0%

last column of Table 4.4. We can see that Exhaustive Search has a sensitivity of 76% and Logistic Iterative has 85% as compared to SVM Iterative with 92% on the validation data.

Based on sensitivity alone, both the [Prigerson *et al.*, 2009] criteria set and the DSM-5 [American Psychiatric Association, 2013] criteria set appear to be missing a significant portion of the CG population. However, in order to fully judge the [Shear *et al.*, 2011] criteria set, as well as our proposed criteria sets, we will also need to examine specificity. Fortunately, our clinical collaborators are currently in the process of collecting the necessary data from a treatment-seeking sample that is deemed free of CG. Once this data is available, we will be able to examine specificity.

Chapter 5

Discussion and Future Research

5.1 Discussion

In this dissertation, we have presented and evaluated several methods to perform disease classification in the presence of the unique logic structure found in DSM-style criteria sets. Based on numerous simulation studies, an Exhaustive Search, the simplest of the techniques, is best applied when the number of domains is small and the predictor variables of interest are counts as it always performs closest to the oracle rule. In these same studies, the newly proposed methods, SVM Iterative and Logistic Iterative, also perform quite well, although not as good as the Exhaustive Search in some cases. However, there are many situations in which the Exhaustive Search becomes infeasible, and hence the need for our two novel approaches, SVM Iterative and Logistic Iterative. As a real example of such a scenario, consider a recent study of major depression and cortisol (stress hormone) levels [Owens *et al.*, 2014]. At baseline, the researchers collected symptoms of depression as well as cortisol levels and used latent class analysis to discover four subgroups. The group with a high number of depressive symptoms and high cortisol levels at baseline was about seven times more likely to go on to develop clinical depression during the three year follow up period as those with low depressive symptoms and low cortisol levels. Based on these results,

it might be desirable to incorporate cortisol levels into future diagnostic criteria for depression. Our proposed methods provide a mechanism to do exactly that. By treating cortisol levels as "Domain 1" in our algorithm and depressive symptom counts as "Domain 2," we could determine the minimum level of cortisol and depressive symptoms necessary to predict a future depression diagnosis. We could then use these thresholds as a screening tool to identify those at high risk of developing clinical depression and offer interventions to prevent its onset or to treat it early to reduce its burden. As was laid out in the Introduction, the need to incorporate biomarker data, similar to this example from the literature, is likely to become more and more common as a result of the transition that medical taxonomy is going through [National Research Council of the National Academies, 2011], and more specifically psychiatric taxonomy [First and Zimmerman, 2006; National Institute of Mental Health, 2011]. We have shown in our simulation studies that when one of our domains is indeed a continuous biomarker, both SVM Iterative and Logistic Iterative perform quite well when compared to using the oracle rule.

One limitation of the SVM Iterative method is that it takes longer to run than either Logistic Iterative or the Exhaustive Search in the simulations and data example scenarios that include a small number of variables per domain. Unlike the other two methods, SVM Iterative requires the tuning of a cost parameter. We have found that rather than performing 5-fold cross validation at each step of the iterative process, we could instead perform tuning only in the first iteration and carry that tuning parameter forward for future iterations. This did not appear to impact the performance of SVM Iterative in any quantifiable way. In addition to its extra computing cost, in a few of the simulation settings (e.g. where there are three domains), SVM Iterative performs a little bit worse than Logistic Iterative, which is not entirely surprising. The advantage of SVM and other machine learning approaches is most evident in high-dimensional settings where the signal is weaker and some noise variables are involved. Although high-dimensional data is not encountered in our

current clinical application, as we discussed earlier, the field is moving towards a more biologically sound diagnostic system using high-dimensional biomarkers and machine learning approaches are being advocated for these applications [Oquendo *et al.*, 2012]. In future work, we plan to explore our approaches in a high-dimensional setting and experiment with other disorders such as depression where biomarkers are currently being collected and assessed.

In our validation study, of which we were only able to evaluate sensitivity, it was clear that both the criteria for PGD [Prigerson *et al.*, 2009] and PCBD (DSM-5) [American Psychiatric Association, 2013] were failing to capture a significant portion of individuals with impairing CG. The CG criteria set [Shear *et al.*, 2011], as well as both the two domain and three domain criteria sets derived based on our proposed algorithms perform reasonably well in terms of sensitivity. As was discussed, data from a sample of bereaved individuals without CG is in the process of being collected. Once available, we will also be able to assess the specificity and overall misclassification rate of the criteria sets. Only then we will be able to assess the full picture. These results are being summarized in [Mauro *et al.*, 2015a].

Several design issues specific to criteria set development have become apparent to us while working on this project. The first is that the data used to develop the criteria set should be sampled from the target population in which it will be used. Otherwise, appropriate adjustments based on the disease prevalence are necessary. We note that the choice of this prevalence most certainly depends on the target population to which the criteria set will be applied. In addition, other evaluations of a criteria set (for example, Positive Predictive Value and Negative Predictive Value) can only be considered under proper design or valid estimation of the disease prevalence. The second issue is that the items on which we build our criteria set should be representative of all of the important domains of the disorder we are measuring. In these analyses, we relied on the ICG, which is known to be missing some key CG symptoms, like suicidal ideation [Shear *et al.*,

2011]. However, another measure that does capture all of the important CG symptoms has already been developed, the SCI-CG. However, we are currently limited to using the ICG because we have the ICG available for a sample of cases and controls, but only have the SCI-CG available on cases. Once we have this new instrument collected on controls, we can easily reapply our techniques and hopefully develop an even better performing criteria set.

Another important part of this process is that we need a valid approach to group symptoms into domains. Our proposed methods assume that the domain structure and the symptoms that belong to each is already known. In the real world, this is not typically the case. To get around that, we used exploratory factor analysis methods to come up with an appropriate grouping structure. However, there may be better ways to do this. Further, for some applications, there may be evidence that using a simple sum score, rather than symptom clusters with thresholds, is a better approach to disease classification. Our framework can easily accommodate this single domain scenario. Lastly, in order to determine the threshold values, an accepted alternative to the gold standard, sometimes referred to as the "LEAD" standard, is necessary. Our methods assume that the case status is known for a subset of the population, despite the fact that a criteria set does not yet exist. In this application, caseness was defined using a well accepted threshold on the ICG as well as a diagnosis by an expert clinician. As a final note, to fully understand the performance of the developed criteria set, antecedent, concurrent and predictive validity set should also be evaluated.

5.2 Future Work

With respect to including other types of data to perform classification, our proposed methods are very flexible. Suppose that instead of domain counts, we had a large number of genes for one domain and imaging measures as our second domain. In this case the Exhaustive Search is certainly not applicable. However, due to the flexibility of the SVM and the kernel "trick," we can easily replace

our linear decision boundary with a nonlinear boundary that uses a Gaussian kernel, for example. This allows us to make a high-dimensional problem feasible. This flexibility also extends to changing how we measure our outcome. Right now we only consider a binary outcome, as is consistent with the DSM. However, it is likely, that in the future, we will instead see a dimensional outcome either in addition to or replacing the current binary diagnoses [Kraemer, 2007; Krueger and Bezdjian, 2009; National Institute of Mental Health, 2011]. The essential idea of our algorithm would remain the same in this scenario; we could replace the SVM in SVM Iterative with Support Vector Regression or logistic regression of Logistic Iterative with linear regression. Lastly, the methods we have evaluated currently assume that all items are equally important in terms of diagnostic ability. As was discussed in the Introduction, this is generally not true [Clark and McKenzie, 1994; Aggen *et al.*, 2005; Cooper *et al.*, 2010]. In section 2.2.2 we laid out the theoretical framework to empirically estimate the relative importance of different symptoms in disease classification. We plan to implement and evaluate this in future work.

In operationalizing our methods, we assume that we know the number of domains and which items they contain. In reality, these are both unknowns and we need to employ statistical methods such as EFAs to derive them. Our current methods do not address this added variability in any way. For our clinical application, we are currently choosing the two factor model over the three factor model because it had a lower misclassification rate on the training data. If we had a full test set, we would use the misclassification rate from that instead to choose between the two competing models. Our simulations settings give us some confidence in using this approach; when the correct number of domains was used, all three methods had lower misclassification errors than when they were applied to the incorrect number of domains (Setting A versus Setting H). However, we are unsure if this is the "best" way to decide between competing structures. Of great interest to us is if there is a way to combine these two processes together; that is, estimate the structure and

the thresholds simultaneously. Right now we are choosing the optimal thresholds based on a given structure. A better approach might be to choose both the structure and thresholds based on some composite performance measures that evaluate fitness of factor model (e.g., through a likelihood) and diagnostic performance (e.g., through misclassification rate) and jointly estimate the structure and thresholds. We plan to examine such a joint approach in future work.

Another strong assumption that we are making is that in our training data, our cases are truly cases and that our controls are truly controls. Because Complicated Grief is a new disorder, and no gold standard exists, it is pretty much impossible for this assumption to hold. In previous work, we proposed a method to perform disease classification in the absence of a gold standard [Wang *et al.*, 2013]. In this approach, our feature variables are known symptoms, such as the ICG items. In order to carry out classification, the method borrows information from auxiliary prognostic markers, such as measures of functional impairment. The task is viewed as statistical learning in the presence of missing data, and introduces a pseudo-EM algorithm to carry out the classification. One limitation of this method is that it results in a linear classification rule based on the feature variables, which is not consistent with criteria sets currently found in the DSM. Eventually, the goal is combine this previous work with our current work to produce criteria sets that are consistent with the DSM logic structure while also accounting for the fact that we do not have a gold standard diagnosis.

Although we did not need the Simultaneous Approach in this application, it is still necessary in other applications such as the "OR" scenario found in the ADHD that was briefly discussed in section 2.2.3.2. Recall, that for ADHD, children must have at least six symptoms from either the inattention group OR the hyperactivity and impulsivity criteria [American Psychiatric Association, 2013]. However, one can imagine more complicated logic structures including some combination of "AND/OR." For example, assume it is known that domains $k = 1, \cdots, G_1$ have an "OR" relationship and domains $k = G_1 + 1, \cdots, G$ have an "AND" relationship. The overall classification

rule is now of the form

$$\text{sign}\left\{\max[f_1, \cdots, f_{G_1}, \min(f_{G_1+1}, \cdots, f_G)]\right\}.$$

Our objective function is the same as in (2.7), but our constraints are modified in the following way:

$$\begin{aligned}
\text{for } y_i = 1: \quad & y_i \max(\beta_{01} + \beta_{11}x_{i1}, \cdots, \beta_{0G_1} + \beta_{1G_1}x_{iG_1}) \geq 1 - \xi_i, \text{and} \\
& y_i(\beta_{0k} + \beta_{1k}x_{ik}) \geq 1 - \xi_i, k = G_1 + 1, \cdots, G, \quad \xi_i \geq 0 \\
\text{for } y_i = -1: \quad & y_i(\beta_{0k} + \beta_{1k}x_{ik}) \geq 1 - \xi_i, k = 1, \cdots, G_1, \quad \xi_i \geq 0 \\
& y_i \min(\beta_{0G_1+1} + \beta_{1G_1+1}x_{iG_1+1}, \cdots, \beta_{0G} + \beta_{1k}x_{iG}) \geq 1 - \xi_i.
\end{aligned}$$

This is also useful in genetic studies when it is known that a path is activated only when all the genes in the same pathway are turned on and a few pathways jointly influence the disease risk additively. In this case, we will know which genes belong to the "AND" group and which belong to the "OR" group. In this case, assuming there are J pathway each with G_j genes, the disease risk can be predicted as

$$\text{sign}\Big\{\max\left[\min(f_1, \cdots, f_{G_1}), \cdots, \min(f_{G_{J-1}+1}, \cdots, f_{G_J})\right]\Big\}.$$

In practice, we may not be known which domains have an "AND" relationship and which have an "OR" relationship. In this case, we can consider taking an iterative step-wise approach. We start with an "AND" rule or an "OR" rule. At each step, there are two operations to build on the existing diagnosis rule: "AND" combination or "OR" combination. We could choose the domain and the operation that will minimize the cross validation misclassification error. This should work when a gene or a domain can only be in an "AND" group or an "OR" group, but not both. For the more complicated situation when a variable can be involved in both, another approach seems necessary. Such an approach would be useful for discovering complicated logic interaction models.

Bibliography

[Aggen *et al.*, 2005] S Aggen, M Neale, and K Kendler. DSM criteria for major depression: Evaluating symptom patterns using latent-trait item response models. *Psychological Medicine*, 35:475–487, 2005.

[American Psychiatric Association, 2010] American Psychiatric Association. APA modifies DSM naming convention to reflect publication changes, 2010.

[American Psychiatric Association, 2013] American Psychiatric Association. *Diagnostic and statistical manual of mental disorders*. American Psychiatric Publishing, Arlington, VA, 5th edition, 2013.

[An and Tao, 1997] L An and P Tao. Solving a class of linearly constrained indefinite quadratic problems by d.c. algorithms. *Journal of Global Optimization*, 11:253–287, 1997.

[Boelen and Prigerson, 2012] P Boelen and H Prigerson. Commentary on the inclusion of persistent complex bereavement-related disorder in dsm-5. *Death Studies*, 36:771–794, 2012.

[Bui *et al.*, 2015] E Bui, C Mauro, D Robinaugh, N Skritskaya, Y Wang, C Gribbin, A Ghesquiere, A Horenstein, N Duan, C Reynolds, S Zisook, N Simon, and M Shear. A psychometric analysis of the structured clinical interview for complicated grief. manuscript, 2015.

[Clark and McKenzie, 1994] L Clark and D McKenzie. A caution on the use of cut-points applied to screening instruments or diagnostic criteria. *Journal of Psychiatric Research*, 28:185–188, 1994.

[Cooper *et al.*, 2010] L Cooper, S Balsis, and M Zimmerman. Challenges associated with a polythetic diagnostic system: Criteria combinations in the personality disorders. *Journal of Abnormal Psychology*, 119(4):886–895, 2010.

[First and Zimmerman, 2006] M First and M Zimmerman. Including laboratory tests in DSM-V diagnostic criteria. *The American Journal of Psychiatry*, 163(12):2041–2, 2006.

[Genkin *et al.*, 2007] A Genkin, D Lewis, and D Madigan. Large-scale bayesian logistic regression for text categorization. *Technometrics*, 49(3):291–304, 2007.

[Hastie *et al.*, 2009] T Hastie, R Tibshirani, and J Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics. Springer, New York, second edition, February 2009.

[Kersting *et al.*, 2011] A Kersting, E Brahler, H Glaesmer, and B Wagner. Prevalence of complicated grief in a representative population-based sample. *Journal of Affective Disorders*, 131:339–343, 2011.

[Kraemer, 2007] H Kraemer. DSM categories and dimensions in clinical and research contexts. *International Journal of Methods in Psychiatric Research*, 16:S8–S15, 2007.

[Krueger and Bezdjian, 2009] R Krueger and S Bezdjian. Enhancing research and treatment of mental disorders with dimensional concepts: toward DSM-V and ICD-11. *World Psychiatry*, 8:3–6, 2009.

[Kupfer *et al.*, 2008] D Kupfer, D Regier, and E Kuhl. On the road to DSM-V and ICD-11. *Eur Arch Psychiatry Clin Neurosci*, 258:2–6, 2008.

[Lin, 2002] Y Lin. Support vector machines and the bayes rule in classification. *Data Mining and Knowledge Discovery*, 6:259–275, 2002.

[Lin, 2004] Y Lin. A note on margin-based loss functions in classification. *Statist. Probab. Lett.*, 68:73–82, 2004.

[Luenberger, 1984] D Luenberger. *Linear and Nonlinear Programming*. Addison-Wesley, Reading, MA, second edition, 1984.

[Mauro *et al.*, 2015a] C Mauro, Y Wang, N Skritskaya, C Reynolds, N Simon, S Zisook, B Lebowitz, N Duan, and M Shear. Comparing the sensitivity of the DSM-5 criteria set for persistent complex bereavement disorder to other previously proposed criteria sets. manuscript, 2015.

[Mauro *et al.*, 2015b] C Mauro, Y Wang, M Wall, N Skritskaya, N Simon, and M Shear. A confirmatory factor analysis of the inventory of complicated grief scale in a large clinical sample of those with confirmed complicated grief. manuscript, 2015.

[Muthën and Muthën, 2012] L Muthën and B Muthën. *Mplus User's Guide*. Muthën And Muthën, Los Angeles, CA, seventh edition, 2012.

[National Institute of Mental Health, 2011] National Institute of Mental Health. NIMH research domain criteria (RDoC), 2011.

[National Research Council of the National Academies, 2011] National Research Council of the National Academies. *Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and and a New Taxonomy of Disease*. The national academies press, Washington, DC, 2011.

[Noble, 2006] W Noble. What is a support vector machine? *Nature Biotechnology*, 24(12):1565–1567, 2006.

[Oquendo *et al.*, 2012] M Oquendo, E Baca-Garcia, A Artes-Rodriguez, F Perez-Cruz, H Galfalvy, H Blasco-Fontecilla, D Madigan, and N Duan. Machine learning and data mining: strategies for hypothesis generation. *Molecular Psychiatry*, 17:956–959, 2012.

[Owens *et al.*, 2014] M Owens, J Herbert, P Jones, B Sahakian, P Wilkinson, V Dunn, T Croudace, and I Goodyer. Elevated morning cortisol is a stratified population-level biomarker for major depression in boys only with high depressive symptoms. *PNAS*, 111(9):3638–3643, 2014.

[Prigerson *et al.*, 1995] H Prigerson, P Maciejewski, C Reynolds, A Bierhals, J Newsom, A Fasiczka, E Frank, J Doman, and M Miller. Inventory of complicated grief: A scale to measure maladaptive symptoms of loss. *Psychiatry Resarch*, 59:65–79, 1995.

[Prigerson *et al.*, 2009] H Prigerson, M Horowitz, S Jacobs, C Parkes, M Aslan, K Goodkin, B Raphael, S Marwit, C Wortman, R Neimeyer, G Bonanno, S Block, D Kissane, P Boelen, A Maercker, B Litz, J Johnson, M First, and P Maciejewski. Prolonged grief disorder: Psychometric validation of criteria proposed for *DSM-V* and *ICD-11*. *PLoS Medicine*, 6(8):e1000121, 2009.

[Shear *et al.*, 2005] MK Shear, E Frank, PR Houck, and CR Reynolds 3rd. Treatment of complicated grief: A randomized controlled trial. *J Am Med Assoc*, 293:2601–2608, 2005.

[Shear *et al.*, 2011] MK Shear, N Simon, M Wall, S Zisook, R Neimeyer, N Duan, C Reynolds, B Lebowitz, S Sung, A Ghesquiere, B Gorscak, P Clayton, M Ito, S Nakajima, T Konishi, N Melhem, K Meert, M Schiff, MF O'Connor, M First, J Sareen, J Bolton, N Skritskaya,

AD Mancini, and A Keshaviah. Complicated grief and related bereavement issues for *DSM-5*. *Depression and Anxiety*, 28(2):103–117, 2011.

[Shear *et al.*, 2014] M Shear, Y Wang, N Skritskaya, N Duan, C Mauro, and A Ghesquiere. Treatment of complicated grief in elderly persons:a randomized controlled trial. *JAMA Psychiatry*, 71(11):1287–1295, 2014.

[Simon *et al.*, 2011] N Simon, M Wall, A Keshaviah, MT Dryman, N LeBlanc, and MK Shear. Informing the symptom profile of complicated grief. *Depresssion and Anxiety*, 28:118–126, 2011.

[Steinwart, 2005] I. Steinwart. Consistency of support vector machines and other regularized kernel classifiers. *IEEE Transactions on Information Theory*, 51:128–142, 2005.

[Streiner and Norman, 2008] D Streiner and G Norman. *Health measurement scales: a practical guide to their development and use*. Oxford University Press, New York, fourth edition, 2008.

[Tseng, 2001] P Tseng. Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of Optimization Theory and Applications*, 109:475494, 2001.

[van der Vaart and Wellner, 1996] A van der Vaart and J Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics. Springer, New York, 1996.

[Vapnik, 1996] V Vapnik. *The Nature of Statistical Learning Theory*. Springer, New York, 1996.

[Wang *et al.*, 2013] Y Wang, H Chen, D Zeng, C Mauro, N Duan, and M Shear. Auxiliary marker-assisted classification in the absence of class identifiers. *Journal of the American Statistical Association*, 108(502):553–565, 2013.

[Zou *et al.*, 2008] H Zou, J Zhu, and T Hastie. New multicategory boosting algorithms based on multicategory fisher-consistent losses. *The Annals of Applied Statistics*, 2(4):1290–1306, 2008.

Appendix

Structured Clinical Interview for Complicated Grief (SCI-CG)

2. Do you often find yourself yearning or longing for "name of loved one" a lot or feel a very strong desire to be with her/him again?

3. Do you often have intense feelings of sorrow or emotional pain because of the death, like pangs of grief?

4. Do you often have thoughts or images of "name of loved one" that keep coming back even when you're focused on other things, for example visualizing "name of loved one" as he/she was just before he/she died or other thoughts or images of her/him?

5. Do you often get lost or absorbed in thoughts or daydreams about "name of loved one"?

6. Do you think or worry a lot about how or why "name of loved one" died?

7. Do you have trouble accepting the idea that "name of loved one" is not coming back, like you can't really believe it, or like you think it should not have happened?

8. Have you felt shocked or stunned since the death?

9. Have you felt emotionally numb, like you couldn't feel anything even if you wanted to?

10. Do you have difficulty having positive memories or thoughts about "name of loved one"?

11. Do you feel bitter or angry about the death, or about something related to the death?

12. Do you have guilty or self-blaming thoughts or beliefs related to the death?

13. Do you worry a lot about not being able to manage without "name of loved one"?

14. Do you avoid anything because it is a reminder of your loss? For example, do you avoid places you went together, activities with other people that you associate with her/him, looking at pictures of her/him or anything else?

15. Do you avoid getting rid of "name of loved one" possessions even if you really need to?

16. When you encounter reminders of the loss, do you often have intense emotional reactions?

17. When you encounter reminders, do you often have physical reactions like feeling nausea or upset stomach, or dizzy or racing heart or trouble breathing or other physical symptoms?

18. Are there things you do or places you go that are special, that help you feel close to "name of loved one" or feel sure you won't forget him/her, like visiting the cemetery/spending time with ashes, reminiscing about him/her, making scrapbooks, or other things like that?

19. Do you often want to see, hear, touch, smell or spend time with things that remind you of "name of loved one"? Like looking at pictures or holding or smelling things that belonged to him/her?

20. Do you often feel pain or think you have other symptoms that "name of loved one" had?

21. Do you often think you are hearing her/his voice or seeing him/her?

22. Do you often have a wish to die in order to find or join "name of loved one"?

23. Do you often have a wish to die because life is not worth living if "name of loved one" is not here?

24. Do you have difficulty trusting other people who haven't experienced a similar loss?

25. Do you find it difficult to care about or feel close to family or friends - like feeling distant or cutoff or alienated from them?

26. Do you often feel very lonely, like you are all alone in the world now that "name of loved one" is gone?

27. Do you often feel intensely envious of others who haven't experienced similar loss?

28. Do you often feel like your life is empty or no longer has purpose or meaning since "name of loved one" died?

29. Do you feel it is very hard for you to experience joy or satisfaction without "name of loved one"?

30. Do you feel confused or uncertain about your role in the world or your identity since "name of loved one" died?

31. Do you find it difficult to pursue interests or plan for the future because you can't share things with "name of loved one" anymore?