Title

Statistical Techniques for Analyzing Irregular and Sparse Cyclical Longitudinal Data with Applications to Bipolar Disorder

Author

Calimlim, Brian Manalo

Publication Date

2014

University of California

Los Angeles

Statistical Techniques for Analyzing Irregular and Sparse Cyclical Longitudinal Data with Applications to Bipolar Disorder

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Public Health

by

Brian Manalo Calimlim

2014

Abstract of the Dissertation

Statistical Techniques for Analyzing Irregular and Sparse Cyclical Longitudinal Data with Applications to Bipolar Disorder

by

Brian Manalo Calimlim

Doctor of Public Health

University of California, Los Angeles, 2014

Professor Catherine A. Sugar, Chair

Bipolar disorder is an illness characterized by abnormal mood swings encompassing both mania and depression, often with irregular longitudinal patterns. The variable and cyclical episodic nature of the disease presents many challenges for statistical analyses. These complex features make it difficult to characterize the data, define disease improvement measures, and develop appropriate statistical models. This is particularly problematic among rapid cycling bipolar disorder patients whose disease is defined by highly erratic and frequent mood shifts. In this dissertation, I present two approaches to analyzing data from bipolar disorder studies. The first approach focuses on the time spent in various mood states. Using longitudinal mood severity rating scale scores, data are transformed into a sequence of mood states. These sequences are analyzed as a Markov chain and stationary distributions are used to measure within- and between-group differences. The non-parametric bootstrap is employed to test for differences. The second approach focuses on features of the mood episodes. Mood severity rating scale scores are modeled as a longitudinal function of episodes and patient-specific characteristics. Episodes are parameterized by their durations, peak severities (amplitudes), and times of occurrence (locations). This flexible parametric model is fit to the data using a global iterative search algorithm known as Particle Swarm Optimization. To reduce the dimensional space of the search algorithm, an episode detection method is proposed. Estimates are derived for each patient and are used as inputs in secondary statistical models. These approaches are applied to a three-arm randomized trial of rapid cycling bipolar disorder patients. Mechanisms of these approaches are tailored to address sparsity and small sample size issues present in the data. Simulations are used to assess the statistical performance and agreement of these approaches, and recommendations for clinical application are presented.

The dissertation of Brian Manalo Calimlim is approved.

Patricia D. Walshaw

Peter C. Whybrow

Donatello Telesca

Gang Li

Catherine A. Sugar, Committee Chair

University of California, Los Angeles

2014

To the Lord who establishes my steps and straightens my path.

Table of Contents

List of Figures

List of Tables

Vita

2003	B.S. (Cybernetics), UCLA, Los Angeles, California.
2005–2009	Biostatistician, Cerner LifeSciences, Beverly Hills, CA.
2006	M.S. (Biostatistics), UCLA, Los Angeles, California.
2006–2012	Teaching Assistant, Department of Biostatistics, UCLA, Los Angeles, CA.
2006–2008	Graduate Student Researcher, Southern California Injury Prevention Research Center, UCLA, Los Angeles, CA.
2008–2009	Graduate Student Researcher, Department of Urology, UCLA, Los Angeles, CA.
2010–2012	Graduate Student Researcher, Department of Community Health Sciences, UCLA, Los Angeles, CA.
2012–2013	Graduate Student Researcher, Semel Institute Biostatistics Core, UCLA, Los Angeles, CA.
2013–Present	Health Economist, ICON plc, El Segundo, CA.

Publications and Presentations

B.B. Dean, **B.M. Calimlim**, P. Sacco, D. Aguilar, R. Maykut, and D. Tinkelman. Uncontrolled asthma among children: impairment in social functioning and sleep. *The Journal of Asthma* 2010 June; 47(5):539-44.

B.B. Dean, **B.M. Calimlim**, P. Sacco, D. Aguilar, R. Maykut, and D. Tinkelman. Uncontrolled asthma: assessing quality of life and productivity of children and their caregivers using a cross-sectional Internet-based survey. *Health and Quality of Life Outcomes* 2010 September 8; 8:96.

B.M. Calimlim, C.A. Sugar, P.D. Walshaw, and P.C. Whybrow. Thyroid hormone treatment efficacy in bipolar disorder with irregular sparse cyclical longitudinal data. Presented at the Joint Statistical Meetings, 2011 July.

B.M. Calimlim, P.M. Massey, M. Prelip, A. Afifi, E.S. Quiter, S. Nessim, N. Wongvipat-Kalev, D. Osuna, and D.C. Glik. Examining health literacy among a low-income adolescent population in California. Presented at the American Public Health Association Annual Meeting, 2012 October.

P.M. Massey, M. Prelip, **B.M. Calimlim**, E.S. Quiter, and D.C. Glik. Contextualizing an expanded definition of health literacy among adolescents in the health care setting. *Health Education Research* 2012 December; 27(6):961-74.

P.M. Massey, **B.M. Calimlim**, A. Afifi, E.S. Quiter, S. Nessim, N. Wongvipat-Kalev, and D.C. Glik. Exploratory findings toward a multidimensional measure of adolescent health literacy. *American Journal of Health Behavior* 2013 May; 37(3):342-50.

CHAPTER 1

Introduction

The purpose of this dissertation is to develop statistical methods for the analysis of data from studies of bipolar disorder, an illness characterized by abnormal and irregular mood swings. Its highly variable and cyclical nature presents challenges for statistical analyses, making it difficult to characterize the data and develop appropriate models. The multidimensional nature of the disease, encompassing both mania and depression, complicates the definition of disease improvement measures. Moreover, as the nature of the disease is episodic, data are generally longitudinal which adds another dimension of complexity, especially as the patterns of mood changes across time are not regular. These challenges present many barriers in data analysis but also provide avenues for developing innovative approaches.

The approaches developed here are motivated by data from a three-arm randomized trial comparing two different thyroid hormone treatments against a placebo in a treatment-refractory population. The patients in the trial exhibit an increased frequency of mood swings, and are thus classified as rapid cycling. This adds an additional layer of complexity by making it difficult to distinguish between random fluctuation and the true underlying data mechanism. Furthermore, the study suffers from challenges in sample size and data sparsity. A total of 32 patients are observed at approximately two-week intervals which may be long relative to their cycling rate. The aim of this dissertation is to address these challenges by developing unique, clinically-based analytical approaches. Because this is a treatment trial, efforts are focused on detecting improvements in various dimensions of disease severity, such as episode severity, frequency, and duration, and comparing the effects among groups.

Chapter 2 provides an introduction to bipolar disorder and the clinical setting of application. Basic mechanisms of the disease, diagnosis, and available treatments are discussed, as well as instruments used to quantitatively measure symptom severity. Issues concerning the statistical analysis of bipolar disorder data are also discussed in the context of defining and measuring treatment effects. Chapter 3 presents the details of the three-arm randomized trial. Results from the originally proposed analysis are given and the shortcomings of the approach are discussed. Chapter 4 presents a state-based approach to the problem by adopting a perspective focused on the mood states of bipolar disorder. Data are formulated as sequences of mood episodes and are analyzed as a Markov chain. This is combined with a bootstrap procedure to facilitate hypothesis testing. Chapter 5 presents an approach that focuses on episode characteristics. Longitudinal data are modeled as episode combinations characterized by severity and duration. A grid-based search is adopted to detect episodic regions and data are fit to the model using particle swarm optimization, an iterative global search algorithm. Resulting estimates are analyzed in a repeated measures framework to assess group differences. Chapter 6 presents simulation studies of these two approaches. Chapter 7 outlines areas of future work.

CHAPTER 2

Literature Review

2.1 Bipolar Disorder

2.1.1 Overview

Bipolar disorder is a psychiatric illness characterized by abnormal mood shifts and behavior. These mood swings result in inconsistent and unpredictable energy, activity, and productivity levels. While it is normal for people to experience mood shifts in daily life, bipolar disorder patients exhibit mood severities and irregularities to a degree that obstructs normal living, functionality, and social interaction. Mood can generally be categorized by four states: euthymia, mania, depression, and mixed. A period of elevated mood and energy level is known as a manic episode and the person is said to be in a manic state. A period of lowered mood and interest level is known as a depressive episode and the person is said to be in the depressed state. A period with a mood that is reflective of both the manic and depressed states is known as a mixed episode and the person is said to be in the mixed state. Times of normal mood are referred to as euthymia and the person is said to be in the euthymic state. Formal definitions and criteria of manic, depressive, and mixed episodes do exist are described later.

United States population statistics suggest a persistent prevalence of bipolar disorder. Based on data collected from 1980 to 1985, the lifetime prevalence of bipolar disorder in the adult United States population was estimated to be approximately 1.3%[1,2], while data collected from 2000 to 2003 suggest a prevalence of ranging from 1% to 4%[3,4]. Data from the later survey also indicated a median age of onset of 25 years; the mean age varied by bipolar disorder subtypes with younger ages of onset being observed in more severe subtypes. Treatment effectiveness remains an issue. Based on a United States national survey, only 48.8% of bipolar disorder patients received treatment in the previous twelve months while 18.8% received adequate treatment (two or more months of pharmacotherapy plus four or more physician visits in the past year)[5]. Furthermore, the economic burden of the disease is substantial. For 2009, the total direct and indirect cost of bipolar disorder in the United States was estimated to be $151 billion[6]. Variation in symptom severity profiles among bipolar disorder patients present further challenges in effective disease treatment and management. These statistics underscore the continuing need for treatment research in bipolar disorder.

2.1.2 Diagnosis

Bipolar disorder is diagnosed by examining a patient's mood history and identifying episodes suggestive of the disease. Guidelines defining and characterizing mood episodes are outlined in The Diagnostic and Statistical Manual for Mental Disorders (DSM), currently in its fifth edition[7]. The DSM classifies episodes according to abnormal mood severity and duration, associated psychological symptoms, and functional impairment. Episodes are generally classified as manic, major depressive, and mixed, although subtypes do exist. Episode severity is determined by the number of associated symptoms, the degree of impairment, and whether psychotic features are present. Symptoms must not be the result of substance-related physiological effects.

Manic episodes are characterized by an abnormally elevated euphoric or irritable mood that persists for at least a week (or for any duration if associated with a hospitalization). To be considered a manic episode, the DSM requires that at least three associated symptoms (four if the mood is only irritable) persist during this period and be present to a significant degree. These associated symptoms are: inflated self-esteem, decreased need for sleep, talkativeness, racing thoughts, distractibility, increased goal-directed activity, and excessive involvement in pleasure-oriented activities without regard for consequences. Additionally, these symptoms must be severe enough to cause a distinct impairment in work activity or social functioning, require hospitalization, or have psychotic features.

Also used to diagnose bipolar disorder are hypomanic episodes, which share the same characterizations and symptoms as manic episodes. Although symptom severities required for this categorization are not great enough to cause significant impairment or lead to a hospitalization, mood disturbance and changes in functionality should be great enough to be observable to others and considered to be uncharacteristic of the patient in the absence of symptoms. Moreover, the DSM requires a duration of at least four days.

Major depressive episodes are also a defining feature of bipolar disorder and are characterized by a depressed mood and an overall lack of interest or pleasure (anhedonia). The DSM considers a major depressive episode as lasting for at least two weeks and requires the presence of at least five associated symptoms. These associated symptoms are: depressed mood for most of the day, anhedonia in nearly all activities for most of the day, significant weight loss or change in appetite, insomnia or hypersomnia, observable psychomotor agitation or retardation, fatigue, feelings of worthlessness, diminished concentration abilities, and recurrent suicidal thoughts. All symptoms must present nearly every day during the mood disturbance (with the exception of the suicidal thoughts), and symptoms must exhibit severities great enough to cause clinically significant distress or impairment in social and occupational functionality.

Periods with both manic and major depressive symptoms are classified as mixed episodes. To meet the DSM criteria for a mixed episode, the symptom criteria of manic and major depressive episodes must be met simultaneously and the associated symptoms must persist nearly every day for at least a week. Similarly, symptom severity must be great enough to cause marked impairment in occupational and social functioning, require hospitalization, or have psychotic features.

Patients are diagnosed according to their lifetime episode history and are further classified into subtypes of the disease. The two major classifications of bipolar disorder are Bipolar I Disorder and Bipolar II Disorder. Patients with a clinical history of at least one manic or mixed episode are diagnosed as having Bipolar I Disorder. Although it is not included in the criteria for diagnosis, it is common for Bipolar I Disorder patients to have a clinical history that includes major depressive episodes. Patients with a history of at least one major depressive and one hypomanic episode are diagnosed as having Bipolar II Disorder. It is possible for a patient diagnosed with Bipolar II Disorder to progress into Bipolar I Disorder by later experiencing a manic or mixed episode: In a longitudinal study examining the conversion of bipolar disorders, 17.4% of Bipolar II Disorder patients progressed into Bipolar I Disorder[8]. Other classifications exist, such as Cyclothymic Disorder and Bipolar Disorder Not Otherwise Specified, though these classes are not central to the approaches described here.

First specified in the fourth edition of the DSM is a feature of bipolar disorder known as rapid cyling, which occurs in a subset of both Bipolar I Disorder and Bipolar II Disorder. Rapid cycling bipolar disorder is characterized by shorter and more frequent mood swings. To formally be a rapid cycler according to the DSM, a subject must meet the Dunner-Fieve criterion[9] of four or more mood episodes in a twelve month period. On average, bipolar subjects who are non-rapid cyclers experience one manic/hypomanic and one depressive episode in a twelve month period, while rapid cyclers experience approximately eight times as many[10,11]. Studies suggest that approximately 15% to 20% of bipolar disorder patients meet the Dunner-Fieve criterion[9,12,10].

2.1.3 Disease Mechanisms

The mechanisms that underlie mood-related illnesses such as bipolar disorder encompass a broad spectrum, including environmental stressors, genetic predispositions, and cellular biology. Treatment research has focused on the neurobiology underlying bipolar disorder, particularly biochemical abnormalities and complex pathways involved in regulatory processes related to mood. This includes the study of neurotransmitters such as serotonin and dopamine, thyroid hormones triiodothyronine (T_3) and thyroxine (T_4), and associated signaling networks.

Neurobiological research in mood disorders has primarily focused on neurotransmitter systems. Neurotransmitters are chemicals within nerve cells that, when released, trigger a response within a neighboring cell. These target cells have neurotransmitter-specific receptors and are involved in other processes. Neurotransmitter release is caused by a cascade of stimuli. This creates a signaling framework that allows neurons to regulate systems within the body. Of particular interest in the study of mood disorders is the limbic system, a set of brain structures associated with the regulation of behavioral, emotional, and cognitive functions related to mood. Neurotransmitter systems modulate this regulation. Attention has largely focused on the neurotransmitters norepinephrine, dopamine, and serotonin. Because increased levels of norepinephrine and dopamine are positively correlated with alertness, arousal, and motivation, research suggests that low levels of these transmitters are associated with depression, while high levels are tied to mania. Research has also implicated serotonin transport systems with depression and mania which are synergistic with norepinephrine and dopamine regulation. However, neurotransmitter level is only one of several influential factors influencing the signaling pathway. Other factors of influence include post-synaptic receptor sensitivity, intracellular signaling pathways, cellular plasticity, and genetic predisposition[13,14]. Bipolar disorder and its symptom manifestations are believed to be driven by such a cascade of neurobiological systems[15,16].

2.1.4 The Role of Thyroid Hormones

Biochemical abnormalities in thyroid hormones and dysfunction of the hypothalamo-pituitary-thyroid (HPT) axis have also received attention in bipolar disorder. Thyroid hormones are involved in the signaling and regulation of many processes throughout the body such as metabolism, protein synthesis, and neural development. The production of thyroid hormones is controlled by the HPT axis, which consists of the hypothalamus, pituitary gland, and thyroid gland. The pituitary gland – located adjacent to the hypothalamus – manages a multitude of biological mechanisms through hormone production and serves as a link between the nervous system and the endocrine system, which includes the thyroid gland. To increase levels of thyroid hormone, the hypothalamus produces thyrotropin releasing hormone (TRH). Increased levels of TRH cause the pituitary gland to produce thyrotropin stimulating hormone (TSH), which in turn triggers the thyroid gland to produce triiodothyronine (T_3) and thyroxine (T_4). Serum levels of T_3 and T_4 serve as a negative feedback signal to the hypothalamus and regulate TRH production.

T_4 is a precursor of the biologically active T_3, which enter cells and affect cellular mechanisms through binding with nuclear receptors[17]. The binding of T_3 can impact nuclear transcription, protein synthesis, and other cellular pathways[18,19]. These changes then influence other functions throughout the body. T_4 is predominantly the hormone of transport in the blood and is converted to the active hormone T_3. Conversion from T_4 to T_3 principally occurs outside the thyroid gland and provides a means of maintaining a localized level of active thyroid hormone. Converted T_3 makes up about 80% of the T_3 present in the blood[20]. The remaining 20% are produced by the thyroid gland. T_4, on the other hand, can only be produced by the thyroid gland.

The relationship between thyroid hormone production, T_3 cellular transporters, T_3 nuclear receptors, T_4 to T_3 conversion, and the feedback system within the HPT axis create a complex thyroid regulation system. Breakpoints along this cascade can lead to imbalances and cause biological irregularities. Of particular interest is the association of thyroid function with mood and behavioral disturbances[21,22]. Much like the neurotransmitter systems, thyroid hormone receptors are widely distributed in brain regions and limbic system structures associated with mood disorders. This suggests that thyroid hormones may influence mood regulation and interact with neurotransmitter systems. This relationship is supported by mood-related neurotransmitter interactions with thyroid hormones, such as the increased serotonin system responsiveness associated with T_3 replacement therapy in hypothyroid patients[23]. Additional evidence linking thyroid hormones and mood include animal studies where T_3-receptor knocked-out mice were found to exhibit behavior suggestive of cognitive dysfunction and anxiety[24,25]. Further evidence includes mood disruptions in hyper- and hypothyroid patients[26] that link hypothyroidism with depression and hyperthyroidism with anxiety and irritability. These findings suggest that thyroid hormone regulation is involved with mechanisms related to mood and its symptomatic manifestations.

Patient studies implicate thyroid hormone dysfunction as a possible underlying mechanism of mood disorders. In major depressive patients, research suggests a correlation between episode recurrence and decreases in serum T_3[27]. Low T_4 levels have been associated with greater frequency of episodes among bipolar disorder patients[28] and longer hospitalizations in individuals with mood disorders[29]. T_4 levels also reflect a positive correlation with depressive episode onset and remission[30]. Additional evidence associating HPT axis dysregulation with mood include a diminished TSH response to TRH stimulation in bipolar[31] and unipolar depression patients[32], and a diminished response to T_4 treatment (measured by thyroid hormone serum levels) among depressed patients compared to healthy controls[33]. These findings implicate thyroid hormone dysfunction as a risk factor in mood disorders, though the inherent mechanisms connecting HPT dysfunction and mood disorders have yet to be fully elucidated.

2.1.5 Pharmacotherapies in Bipolar Disorder

There is currently no cure for bipolar disorder, although medical treatments are available for managing and controlling mood symptoms. The American Psychiatric Association (APA) practice guidelines[34] – and the more recently updated University of South Florida (USF) practice guidelines[35] – outline treatment protocols according to episode type and phase. These recommendations specify strategies classified by acute and maintenance treatments. Acute treatments are prescribed when a patient is in a mood episode. The objective of acute treatments is to rapidly reduce symptoms and return the patient to the euthymic state. Maintenance treatments are prescribed when a patient is not in a mood episode, with the intention of preventing episode recurrence, eliminating symptoms or reducing severity increases during the non-episode period, and improving mood stability.

Pharmacological treatments currently approved by the Food and Drug Administration (FDA) for the treatment of bipolar disorder can be categorized into two classes: mood stabilizers and antipsychotics. Mood stabilizers are treatments that aim to reduce symptoms without triggering a mood episode. Lithium was the first mood stabilizer to receive FDA approval in 1970. Other FDA-approved mood stabilizers include anticonvulsants, which are antiepileptic drugs that aim to reduce hyperactive brain activity. These are divalproex, carbamazepine, and lamotrigine. Research investigating the mechanisms behind mood stabilizers have focused on their impact on various signaling pathways within and between cells[13].

Antipsychotics are drugs designed to combat the symptoms associated with mental cognition that sometimes accompany bipolar disorder. Antipsychotics approved by the FDA are chlorpromazine, olanzapine, risperidone, quetiapine, ziprasidone, aripiprazole, and asenapine. Also approved is the combination treatment of olanzapine and fluoxetine, a selective serotonin reuptake inhibitor (SSRI) which is used to treat depression. In general, these antipsychotics aim to inhibit dopamine activity by blocking dopamine receptors. and, in some cases, certain serotonin receptors as well[13].

Currently, there are only eleven pharmacological agents currently approved by the FDA for the treatment of bipolar disorder. These drugs are approved for specific episode types, phase (acute treatment versus maintenance treatment), and treatment combinations (monotherapy versus adjunctive therapy). In practice, treatment protocols involving these drugs may include adjunctive treatments with other pharmacotherapies, or be used in a different phase.

Treatment guidelines from the APA and USF suggest that treatment for acute manic and mixed episodes begin with mood stabilizer monotherapy, commonly lithium or divalproex. For more severe cases, specific antipsychotics may be added to the regimen. If the treatment is still ineffective, other treatment combinations are considered and dosages may be increased. Nearly all FDA-approved treatments have been indicated for the treatment of acute mania. For acute depressive episodes, only two treatments have received FDA approval: quetiapine monotherapy and olanzapine taken in combination with fluoxetine. USF guidelines suggest these treatments as first-line therapy. If the treatment is ineffective, the USF recommends treatment with mood stabilizers, potentially in combination with other SSRI's. Both the APA and USF guidelines do not recommend antidepressant monotherapy for the treatment of bipolar disorder depression due to the risk of potentially triggering a mood shift into mania.

Maintenance treatments indicated by the FDA are limited to lithium, lamotrigine, olanzapine, and aripiprazole. Quetiapine has also been indicated for long term management, though only as an adjunctive treatment. For long term management, the APA guidelines recommend treatment with lithium or divalproex. USF guidelines make different recommendations based on remission status. If full remission has been achieved, the USF guidelines recommend the continuation of effective and well-tolerated acute treatments. If only partial remission is achieved, treatment recommendations are dependent on whether the residual symptoms are reflective of mania or depression, although both recommendations include mood stabilizers in conjunction with antipsychotics.

While these pharmacological agents have been found to improve symptoms and assist with disease management, some patients – such as rapid cyclers – have a malignant form of bipolar disorder and are refractory to traditional treatments. There have been no treatments formally approved specifically for rapid cycling. While the APA guidelines suggest the use of mood stabilizers as an initial treatment, they also indicate that combination treatments may be necessary and do not make any specific recommendations. While some treatment efficacy has been demonstrated with mood stabilizers and antipsychotics, treatment response in general appears to be lower in rapid cycling patients compared to non-rapid cyclers[36]. In particular, studies have observed persistent depressive symptoms among rapid cyclers, so much so as to implicate treatment resistant depression as a hallmark feature of rapid cycling[37]. While antidepressants may be used to address this issue, they are thought to trigger a mood switch into mania which is especially problematic in a population already susceptible to cycling.

2.1.6 Thyroid Metabolism and Rapid Cycling Bipolar Disorder

To address the unmet therapeutic needs in the rapid cycling population, treatment research has focused on systems interacting with neurotransmitter systems, particularly the HPT axis and thyroid hormones[21,38]. While the literature examining adjunctive T_3 and T_4 treatments is not as extensive as that on mood stabilizers and antipsychotics, findings suggest that thyroid hormone treatments may provide therapeutic benefits.

One of the first studies of T_4 found that supraphysiological doses induced remission in 5 out of 7 female rapid cycling patients[39]. Positive effects of T_4 treatment were observed in 12 case reports[40] and suggested that the addition of T_4 to mood stabilizers prevented rapid cycling. A later study further confirmed the effects of adjunctive T_4 treatment and observed a response in 10 of 11 rapid cycling patients[41]. This included a decreased severity and frequency in both depressive and manic symptoms. Of the four patients who later participated in single or double-blind placebo substitution, three relapsed, further supporting the effectiveness of adjunctive T_4 treatment. Other studies implicate adjunctive T_4 treatment as an effective acute treatment for major depressive episodes among a treatment resistant population, and may be an effective option among treatment refractory rapid cyclers[41,42].

Favorable effects of T_3 treatment have also been reported. A meta-analysis that included studies with both unipolar depression and bipolar patients implicated T_3 as a viable adjunctive treatment among patients resistant to tricyclic antidepressants[43,44]. Other studies suggest that T_3 may accelerate the response to tricyclic antidepressants[45] and augment the effect of sertraline[46].

The relationship between thyroid metabolism and treatment refractory rapid cycling bipolar disorder is further discussed in Chapter 3.

2.1.7 Measures

There is vast literature concerning the appropriate measurement instruments of bipolar disorder for the purposes of phenotype assessment and clinical study[47,48,49]. The most common quantitative traits in bipolar disorder are measures of symptom severities associated with particular episode types. These scales aim to characterize the global impression of mood and can potentially be used to classify subjects into clinically meaningful mood states. These measures are also helpful in evaluating treatment response and clinical remission. Due to the dual nature of bipolar disorder, separate scales are often used to evaluate depression and mania.

One of the most widely used scales for depression is the Hamilton Depression Rating Scale (HDRS)[50]. The HDRS consists of items that assess the severity of depressive symptoms. These include: depressed mood, feelings of guilt, suicide, insomnia, anhedonia, impaired speech and concentration, agitation, anxiety, loss of appetite, fatigue, hypochondria, weight loss, and cognizance of the illness. The original version of the HDRS was comprised of 17 items which were rated by a clinician. Additional items have been added in subsequent versions of the HDRS, to be used for subtyping depression. Each item consists of ordered categorical responses, scored on a numeric scale (ranging from either zero to four or zero to two); the sum of item scores is used as the overall measure of depression severity. HDRS scores range from 0 to 52. Guidelines suggest that scores less than 8 indicate euthymia/normal mood[51,52,53], although there is some variation in the cutoffs for subthreshold depression and higher severity categories[54]. The APA[51] classifies scores between 8 and 13 as mild, 14 to 18 as moderate, 19 to 23 as severe, and 23 and above as very severe. Other instruments used to measure depression severity include the Beck Depression Inventory (BDI) and the Montgomery Asberg Depression Rating Scale (MADRS).

The rating scale most wide used to measure mania severity is the Young Mania Rating Scale (YMRS)[55]. Like the HDRS, the YMRS is composed of items which assess various symptoms on an ordinal categorical scale. The instrument is comprised of eleven items covering the following domains: elevated mood, increased motor activity, increased sexual interest, decreased sleep, irritability, talkativeness, racing thoughts, grandiose ideas, aggressive behavior, unkempt appearance, and cognizance of the illness. Items are rated by a clinician and the sum of item scores are used as a summary measure of mania severity. Scores for the YMRS range from 0 to 60. The original paper/scale developers suggested cut-off scores of 13, 20, 26, and 30 to categorize euthymia, hypomanic, manic, and severely manic, respectively, although the literature reflects variability in classification. For example, previous studies have used more conservative scores of 8[56,57] and 6[58,59] as the threshold for euthymia. Other instruments used to measure mania severity include the Bech-Rafaelsen Mania Assessment Scalse (MAS) and the Clinician-Administered Rating Scale for Mania (CARS-M).

The HDRS and YMRS must be administered by a clinician. This has the drawback of making it impractical to track rapid shifts in mood as data collection requires scheduled clinical appointments. The ChronoRecord[60] was developed to capture daily fluctuations in mood to finely chart the course of disease. The ChronoRecord is a self-report measure of a patient's overall mood during a 24-hour time period using a 100-point visual analog scale. The lowest point is anchored to the most severe depression level the patient has experienced, while the highest point is anchored to the most severe mania level the patient has experienced. While the ChronoRecord can be measured frequently, a potential drawback is that it is a subjective measure anchored by a patient's disease experience. In contrast, the HDRS and YMRS offer an objective clinical rating based on the entire spectrum of the disease. There is some evidence that ChronoRecord scores correlate well with HDRS and YMRS scores[61], and that the instrument is able to discriminate between mania and hypomania[62].

2.2 Statistical Considerations

The complex nature of bipolar disorder presents many challenges for data analysis. These challenges must be considered to appropriately characterize the course of the disease, and determine the efficacy of an intervention. These considerations affect study design, modeling, and inference choices. Below I present several considerations which will be the focus of this dissertation.

2.2.1 Defining and Measuring a Treatment Effect

Improvements in mood symptom stabilization can be characterized in multiple dimensions. This makes it difficult to define a treatment effect that is both comprehensive and simple. Investigators must choose specific characteristics when defining a treatment response. One characteristic is episode severity. This "amplitude" may be the worst score within an episode experienced, the average episode severity, variability of episode severity, or trends in episode severity over time. Another dimension is temporal episode patterns, including frequency (rate) and duration (length). As with severity, there are multiple formulations including the longest episode experienced, the average episode duration, variability of episode duration, proportion of time spent in a mood state, trends in episode duration, and time between episodes. Treatment effects may also be multidimensional. For example, it is unclear whether a decreases in episode severity and increases in duration are favored over an unchanged severity with decreases in frequency. Although a multidimensional definition may be more comprehensive, it may require more complex statistical methods and models. The interpretation of the treatment effect must also consider the heterogeneity within the study population, as the disease course can vary widely from patient to patient. For example, two patients may be characterized as rapid cyclers, though their episode frequencies, severities, and types may differ. Moreover, this presents complications in defining a single quantitative measure of treatment effect. While scales capturing both depressive and manic symptom severities provide a comprehensive picture of mood and mood patterns, methods for combining this information – and developing metrics for measurement – remain unclear.

2.2.2 Irregular Patterns

The complications of defining and measuring a treatment effect are further exacerbated by the irregular mood patterns present in bipolar disorder. The disease is episodic and described as cyclic in nature, yet the collection of episodes themselves are not necessarily constrained to a predictable, steady pattern. This presents challenges in developing a mathematical model that provides enough structure to capture key components of the disease, yet is flexible enough to allow for irregularities. Additionally, statistical models must also distinguish a true signal from the surrounding noise, which is also complicated by these irregularities. This affects the precision of parameter estimates and reduces statistical power.

2.3 Motivating Study: Three-Arm Randomized Trial

The next chapter presents my motivating example, a three-arm randomized thyroid hormone treatment trial aimed at assessing treatment effects of T_3 and T_4 compared to a placebo. This study reflects the challenges outlined above as well as trial-specific issues. The study's proposed analysis plan is presented and its inadequacies are highlighted and discussed. Chapter 4 presents a state-based approach to the data and outlines a procedure for analyzing the data as mood state sequences using the bootstrap resampling technique to test for treatment effects. Chapter 5 presents a flexible mood-based approach for describing mood trajectories using key episode characteristics, and fits the model using an iterative optimization technique known as particle swarm optimization. Chapter 6 evaluates the statistical properties of these two methods through simulation studies. Future work is outlined in Chapter 7.

CHAPTER 3

Three-Arm Randomized Trial

3.1 Background

Though advances have been made in the pharmacological therapies for bipolar disorder, a subgroup of treatment-refractory rapid cycling bipolar disorder patients persists. Previous research has identified a disproportionate amount of hypothyroidism among rapid cyclers compared to non-rapid cyclers[63,64]. This association poses a unique dilemma as lithium treatment can induce hypothyroidism[65], suggesting that lithium may inadvertently obstruct mood stabilization. Rapid cycling has also been associated with gender, with rapid cyclers more common in women than men[12,10]. This may partially explain the association between hypothyroidism and rapid cycling, as women have an increased risk of developing hypothyroidism compared to men[66]. While these findings suggest thyroid dysfunction as a contributing factor to treatment-refractory rapid cycling, it is unclear whether the association is confounded by gender effects or lithium treatment.

To address these uncertainties, an open-label trial of supraphysiological doses of thyroxine (T_4) in treatment-refractory rapid cycling bipolar disorder was conducted[41]. The study found that gender and prior lithium treatment were not enough to account for the association between rapid cycling and hypothyroidism. This identifies a relationship between rapid cycling and thyroid dysfunction and suggests that thyroid hormone availability may need to be restored among rapid cyclers.

Though the open-label trial identifies a link between rapid cycling and thyroid dysfunction, it is unclear whether diminished thyroid levels are due to low thyroid hormone supplies or thyroid hormone uptake obstructions. In the trial, nearly all responders exhibited serum total T_4 level and free T_4 index increases at the time of clinical response. However, increases in triiodothyronine (T_3) were minor. Therefore, it is unclear whether T_4 supplementation induces a response on its own, or if it rectifies a breakdown in the T_3 conversion pathway.

To help identify the possible thyroid hormone dysfunction, a randomized three-arm trial compared the effects of T_3 and T_4 to a placebo (PL). High-level doses of T_4 were administered, while the T_3 dosing regiment was structured to achieve normal levels. It was hypothesized that, if both T_4 and T_3 treatments were effective, then the metabolism dysfunction involves T_4-to-T_3 conversion impairments. However, if T_4 alone elicited an effect, then the impairment is either due to mechanisms that do not involve T_3 or thyroid hormone imbalances within the brain. Details of the trial are presented next.

3.2 Study Description

Treatment refractory rapid cycling bipolar disorder patients participating in a larger study of patient sensitivity to lithium's antithyroid effects were recruited into a three-arm randomized trial that aimed to evaluate the efficacy of T_3 and T_4 as an adjunct to lithium monotherapy. Participants of the larger study were required to meet the bipolar disorder criteria outlined in the third edition of the DSM[67] and be rapid cyclers. Rapid cycling status was determined by the Dunner-Fieve criterion of four or more episodes of depression and/or mania in the previous twelve month period[9]. Immediately prior to enrollment in the larger study, subjects also had to be refractory to lithium as indicated by continued rapid cycling after four weeks of treatment. Prior to randomization, all patients received lithium treatment and underwent an evaluation period in which at least one mood episode was experienced. Patients were randomized to receive one of the following treatments in addition to lithium and their current medication: PL, T_3, or T_4. Mood behavior was longitudinally tracked during the evaluation period and after treatment randomization. Blood chemistry measurements were also collected longitudinally to monitor thyroid hormone and lithium levels. Investigators were primarily interested in the effects of T_4 relative to T_3 and PL.

3.3 Measures

Mood symptoms were tracked using three primary instruments. Depressive symptoms were monitored using the HDRS[50], while mania symptoms were measured using the YMRS[55]. Because these scales required a structured clinical interview, depression and mania data were collected during the same clinical visit. The study protocol outlined aims to collect data in weekly intervals, although due to variability in appointment scheduling, visits occurred approximately every two weeks on average. Overall mood was tracked daily using the ChronoRecord[60].

To monitor thyroid hormone levels, measurements of thyrotropin (TSH) and T_4 hormones were collected. TSH is a hormone released by the pituitary gland which stimulates T_3 and T_4 production by the thyroid gland. Through a negative feedback control mechanism, elevated levels of T_3 and T_4 in the blood signal TSH production decreases and therefore T_3 and T_4 production levels drop. In the randomized trial, a diminished TSH level served as an indicator of chemical response to thyroid hormone supplementation. Because TSH level correlates to both T_3 and T_4 levels, TSH measurements were used to capture the overall thyroid response. T_4 values were also collected to measure the chemical response to T_4 specifically. In the T_3 group, TSH and T_4 levels were expected to decrease due to T_3 supplementation. In the T_4 group, TSH levels were also expected to decrease, but T_4 levels were expected to increase due to T_4 supplementation.

3.4 Proposed Analysis Plan

This study assessed treatment efficacy by measuring HDRS and YMRS scores over three time periods: (1) pre-randomization, (2) treatment stabilization, and (3) post treatment-stabilization. Pre-randomization encompassed the period prior to treatment assignment. The treatment stabilization period began at the time of randomization and ended four weeks after TSH reached a level below 0.1μIU/mL. This allowed treatment effects to fully engage after thyroid hormone supplementation had elicited a chemical response. The post treatment-stabilization period was intended to encompass the period where treatment effects were in full force. Similarly, due to variability in TSH nadir times, the post treatment-stabilization period differed across patients. Because patients were recruited from a larger study, the length of the pre-randomization period varied across patients. In the PL group, a TSH response was not expected and the average post treatment-stabilization start time of the T_3 and T_4 patients was selected as the post treatment-stabilization period for group comparisons.

As originally proposed by the investigators, treatment efficacy was evaluated by measuring mood scale differences between the pre-randomization and post treatment-stabilization periods. Two different criteria were used to determine a treatment response. Criterion 1 took the most severe scores recorded within each of the two study periods (pre-randomization and post treatment-stabilization), and defined a treatment response as a 50% reduction in those scores on both scales. Additionally, Criterion 1 required that the most severe HDRS and YMRS score during the post treatment-stabilization period to be less than or equal to 10 and 5, respectively. Patients who met both of these conditions were classified as treatment responders. Criterion 2 focused on the average mood scale severity measures, and defined a treatment response as a 50% reduction in the mean mood scores between the two study periods on both scales. Patients who met this condition were classified as treatment responders according to Criterion 2. Fisher's exact test was used to detect statistical differences in rates of treatment response among the groups.

3.5 Results

A total of n=32 participants were recruited into the trial, of whom 9 were randomized to the PL group, 10 to the T_3 group, and 13 to the T_4 group. The durations of the pre-randomization, treatment stabilization, and post treatment-stabilization periods for each group are summarized in Table 3.1. The median pre-randomization period durations were comparable across the three study groups: 80 days, 80 days, and 78 days for the PL, T_3, and T_4 groups, respectively. Median post treatment-stabilization period durations were also comparable: 109 days, 112 days, and 112 days for the PL, T_3, and T_4 groups, respectively. One patient who received T_3 was lost to follow-up before the treatment stabilization period ended and could therefore provide no information about treatment effects. This individual was removed from further analysis. Because TSH levels were expected to remain unchanged in the PL group, the mean treatment stabilization time of non-PL patients (70 days) was used to define the study intervals for all PL patients.

Maximum mood scale score values during the pre-randomization and post treatment-stabilization periods were compared to determine whether the 50%+ reduction condition of Criterion 1 was met for each scale. These findings are summarized in Table 3.2. Improvements in HDRS scores were minimal. Only one patient in the PL group and two patients in each of the T_3 and T_4 groups met the response criteron. The response rate for mania was somewhat higher. Improvements were observed in three, four, and two patients in the PL, T_3, and T_4 groups, respectively. Only one patient in each of the PL and T_3 groups and none in the T_4 group met the improvement and threshold conditions of Criterion 1 for both mania and depression. There were no statistically significant differences between the groups in terms of response.

Results based on the 50%+ reduction in mean scale scores of Criterion 2 are described in Table 3.3. As would be expected, Criterion 2 yielded higher response rates. However, this higher response rate is expected because the maximum-based measures of Criterion 1

Table 3.1: Study Period Durations (days)

group	statistic	pre-randomization	stabilization	post stabilization
PL n = 9	median	80	70	109
	mean	92.89	70	97.67
	SD	64.24	—	44.27
T_3 n = 9	median	80	53	112
	mean	98.70	60.80	119.11
	SD	74.12	24.95	36.30
T_4 n = 13	median	78	70	112
	mean	86.54	75.77	135.54
	SD	44.30	21.30	98.33

Because TSH levels were expected to remain unchanged in the PL group, we imputed a fixed value of 70 days (the mean treatment stabilization duration of non-placebo patients) to define the post treatment period for all placebo patients. One of the ten patients in the T_3 group was lost to follow-up before the treatment stabilization period ended and was removed from the analysis.

Table 3.2: Criterion 1: Maximum Mood Scale Score Improvements by Study Group

Criterion	group	responders		non-responders		p-value
		n	%	n	%	
HDRS	PL	1	11.1	8	88.9	1.0000
	T_3	2	22.2	7	77.8	
	T_4	2	15.4	11	84.6	
YMRS	PL	3	33.3	6	66.7	0.3774
	T_3	4	44.4	5	55.6	
	T_4	2	15.4	11	84.6	
Criterion 1	PL	1	11.1	8	88.9	0.4968
	T_3	1	11.1	8	88.9	
	T_4	0	0.0	13	100.0	

Improvements were defined as a 50%+ reduction in the most severe mood scale scores observed during the post treatment-stabilization period relative to the pre-randomization period. Criterion 1 required an improvement reduction along both scales and the maximum HDRS and YMRS scores during the post treatment-stabilization period had to be ≤ 10 *and* ≤ 5*, respectively. Percentages are given within group by response status. The P-value corresponds to a Fisher's exact test for group differences in response rate.*

Table 3.3: Criterion 2: Mean Mood Scale Score Improvements by Study Group

Criterion	group	responders		non-responders		p-value
		n	%	n	%	
HDRS	PL	1	11.1	8	88.9	0.5985
	T_3	3	33.3	6	66.7	
	T_4	4	30.8	9	69.2	
YMRS	PL	3	33.3	6	66.7	1.0000
	T_3	4	44.4	5	55.6	
	T_4	5	38.5	8	61.5	
Criterion 2	PL	0	0.0	9	100.0	0.5315
	T_3	2	22.2	7	88.9	
	T_4	2	15.4	11	84.6	

Improvements were defined as a 50%+ reduction in the mean mood scale scores observed during the post treatment-stabilization period relative to the pre-randomization period. Criterion 2 required an improvement reduction along both scales. Percentages are given within group by response status. The P-value corresponds to a Fisher's exact test for group differences in response rate.

are sensitive to outliers, while the mean-based measures of Criterion 2 are robust. A 50%+ improvement in mean HDRS score was observed in one, three, and four patients in the PL, T_3, and T_4 groups, respectively. Improvements were again more frequent in mean YMRS scores, with three, four, and five patients meeting the response criteria in the PL, T_3, and T_4 groups, respectively. However, only two patients in each of the T_3 and T_4 groups experienced sufficient improvements along both scales, and Criterion 2 was not satisfied by any patients in the PL group. These observations did not indicate statistically significant differences between groups in response rates.

Investigations of treatment improvement based on a 50%+ reduction and meeting an absolute severity threshold did not suggest differences between the treatment arms by either criterion metric. Overall, more patients exhibited a response in terms of mania symptoms than in depressive symptoms. However, due to the lack of differences between groups, it is unclear whether this was associated with the administered treatment. Furthermore, treatment improvements frequently occurred in only one dimension. This suggests that the original two-scale criteria may have been too stringent to detect modest treatment effects. For both criteria, few patients demonstrated a treatment response (Tables 3.2 and 3.3): two by Criterion 1, and four using Criterion 2. Moreover, the results indicate a lack of agreement between the criteria as the patients classified as responders differed between the two metrics.

3.6 Issues with the Original Analysis Plan

This trial highlighted many of the challenges inherent in both the implementation of data and from analysis studies of bipolar disorder. While some of the issues were specific to this particular trial, others were the result of common features of the disease. In this section I discuss these features and how they impact statistical analyses in greater detail.

One challenge in this study is data sparsity. While the HRDS and YMRS are well-studied instruments and provide reliable metrics for severity of depression and mania symptoms, they both require a structured interview by a clinician. For longitudinal studies, the logistics of physician appointments make frequent data sampling impractical. Depending on the study's objectives, the data may be too sparse to capture important features of mood dynamics. This is particularly crucial when studying rapid cyclers who may experience frequent and acute shifts in mood polarity. With limited data granularity, these local fluctuations in mood episode severity may not be captured, and indeed whole episodes may be missed, resulting in a misrepresentation of the treatment response. Data sparsity makes it difficult to detect a true underlying signal from an already complex and irregular mood pattern and negatively affects precision and statistical power.

The data sparsity problem in this study was further compounded by irregularities in data sampling, resulting from missed visits, imperfect scheduling, and inconsistencies in the length of follow-up during the pre-randomization and post treatment-stabilization periods. Many standard analytical approaches are unable to accommodate such sampling irregularities. While the treatment efficacy measures originally proposed in this study do not directly rely on sampling regularity, the irregularities may still have an impact if they are associated with the patient's mood. For example, participants experiencing a depressive episode may not keep scheduled clinical appointments compared to participants in the euthymic state, resulting in a downward bias in HDRS scores. Similarly, patient follow-up times may be longer if the treatment is successful and have more scheduled visits, causing a bias toward a favorable treatment effect. A thorough statistical analysis must consider these potential bias and tailor analytical decisions according to the granularity of the data.

The varying patient-to-patient study period times in this study (as indicated by the high standard deviations in Table 3.1) reflect the challenge of obtaining a sufficient amount of data. Because bipolar disorder is inherently complex, it is important to obtain multiple data elements to properly characterize key features of the illness, preferably over time. Additionally, due to differences in the course of the disease, it is important for an analysis to consider the high degree of patient variability and weight the findings accordingly. The originally proposed analysis does not weigh the findings according to the amount of data available, and by using only a single summary score per time period, gives up much of the power of the longitudinal measures.

In addition to the challenges presented by the study measures, data collection processes, and the inherent nature of the disorder, this study highlights the difficulties associated with defining a treatment response. The original protocol defined a treatment response as a 50%+ reduction in both the HDRS and YMRS scores along with maximum score thresholds during the post treatment-stabilization period. This definition does not account for the possibility that a patient may improve in only one symptom or mood rating scale, nor does it makes concessions for patients who experience mood levels that improve along one criterion and not the other. For example, patients may exhibit mood severity levels below the specified HDRS and YMRS thresholds, but that may not reflect a 50%+ reduction. Similarly, patients who experience dramatic reductions in mood severity may be overlooked if their mood levels are still above the thresholds. The lack of apparent treatment response in the randomized trial using these strict criteria suggests that clinically-relevant treatment effects may be more subtle and that analyses with less restrictive definitions should be considered.

Related to the treatment effect definition are the statistical issues associated with using a maximum- or mean-based approach. Longer periods provide a greater likelihood of observing higher mood levels, making a treatment response based on maximum values susceptible to outliers. For example, a patient experiencing one early episode followed by a prolonged period of minimal scores may be counted as a non-responder despite the apparent symptom severity decrease. Though a mean-based approach would address cases such as these, the lengths of patient follow-up times vary and averaging over a different number of visits leads to differences in estimate precision. The criteria of the original analysis pland do not consider differential follow-up times among patients and may result in ineffective characterizations of treatment response.

Another shortcoming of the proposed analysis plan is its failure to consider the longitudinal structure of the data. As noted above, this leads to a reduction in power, but just as importantly, cyclicity is a key feature of the disorder. By examining the data over time, mood dynamics can be captured and key features of the illness can be characterized. This is especially crucial in small sample studies where most of the power comes from repeated measurements. The proposed plan does not account for longitudinal disease attributes, such as trends in mood severity or changes in episode frequency and duration. Ignoring the longitudinal data structure also ignores correlations that may exist between the scale scores and their properties over time. Relationships between mania and depression are not investigated in the proposed framework as the scales are analyzed independently. In particular, this masks periods when subjects are in the highly undesirable mixed state. The original analysis plan summarizes treatment efficacy to a single value for each study period and, given the challenges presented earlier, may not be an accurate characterization of a patient's longitudinal mood severity patterns.

3.7 Proposed Approaches

To better analyze the data from this and similar studies, the challenges outlined above require more sophisticated statistical methods. In the following chapters, two new approaches are presented to address these issues. The first approach reformulates mood data into clinically relevant states to form a Markov chain in an effort to both overcome data sparsity and irregularities and reflect the episodic nature of the illness. It then uses a bootstrap procedure for inference. The second approach aims to reconstruct the underlying process giving rise to mood episodes by proposing a flexible parametric model to describe depressive and manic symptom severity scores using an iterative optimization algorithm to determine estimates of the amplitude, duration, and frequency of mood episodes. These approaches represent refinements which attempt to maximize the information present in the data while generating results that address clinically relevant questions.

CHAPTER 4

Approach 1: Markov Chain with Bootstrap

4.1 Background

The three-arm randomized trial results highlight some of the statistical challenges present in analyzing bipolar disorder data. Some of the challenges arise from the sampling process and measures. Data were collected at doctor's appointments, leading to sparse and irregularly spaced observations. This makes it difficult to capture key local features of the mood data such as highly acute episodes. Another issue is low data volume. Mood cycles in bipolar disorder are erratic and irregular by nature and a higher data volume is required to fully capture the disease's dynamics, specifically episode start and end, peak severity, and rate. Limited data make it difficult to piece together the underlying features that are observed over time.

The power to detect treatment effects was limited not only by the sparse and irregular structure of the data but also by the original definition of efficacy and the corresponding analytical approach. Specifically, treatment efficacy was assessed by threshold-based improvements along both the HDRS and YMRS arcs. In order to be classified as a treatment responder, patients had to experience a 50% decrease (as measured by peak or average scores) in both depressive (HDRS) and manic (YMRS) symptom severity scores during the post treatment-stabilization period relative to the pre-randomization period. Additionally, the HDRS and YMRS scores during the post treatment-stabilization period had to be no greater than 10 and 5, respectively. While improvements on both scales are desirable, the proposed thresholds resulted in only a few treatment responders (two patients total based on Criterion 1 and four patients total based on Criterion 2). This suggests that the criteria may have been too stringent in the sense that few people were full responders. Partial responses are not detected by these methods but may still be of high clinical importance. Additionally, these measures may not detect clinically relevant patterns in mood episodes. For example, it is possible for peak severities to be identical, though episode durations and frequency may have dramatically decreased. By not considering temporal patterns and trends in the mood scale scores, key characteristics of bipolar disorder are not captured including episode duration, episode frequency, and episode severity. For this purpose, it might be better to define treatment efficacy more broadly. This might include examining improvements along each scale separately, considering decreases in symptom severity that are less than 50%, counting very large mood decreases that remain above the outlined post treatment-stabilization period thresholds, or focusing on the amplitude, duration, and spacing shape of mood levels within an episode. Due to the complex multidimensional nature of bipolar disorder, treatment efficacy definitions may need to consider both nuanced levels of improvement, and multiple metrics.

Another issue in the analytical approach relates to how the data are used to extract key pieces of information. Bipolar disorder is a lifelong condition and symptoms are episodic, requiring studies to collect data over time to capture all facets of the disease. It is not obvious how to process the longitudinal information to best summarize characteristics of the illness. In the original analysis, data for each mood scale were reduced to a single HDRS and YMRS score for each of the pre-randomization period and post treatment-stabilization periods. For determining treatment response using Criterion 1, the maximum mood scale score during each period was used; for Criterion 2, the average score was used. While easy to interpret, this approach completely ignores the rich longitudinal structure of the data.

Another challenge in processing the data is developing a method of combining the information contained in the two mood scale scores. The original analysis plan evaluated HDRS and YMRS scores separately rather than attempting to describe their joint mechanisms. While this may sufficiently describe depressive and manic symptom severity, it fails to capture periods of euthymia and mixed episodes which may result in misleading conclusions. For example, depressive and manic symptom severity may remain constant after treatment, but the joint analysis of HDRS and YMRS scores may reveal that fewer mixed episodes are occurring.

Compounding these challenges is the study's sample size. A small sample size reduces statistical power and diminishes the ability to detect a treatment effect. In the three-arm randomized trial, the PL, T_3, and T_4 groups have 9, 10, and 13 patients, respectively. Because the sample size is small, large differences, on the order of 50% in treatment response rates, would have been necessary to obtain statistically significant findings if one only considers the global measure, and neglects the repeated observations of actual mood scores. Indeed, even if all patients in the PL group had been classified as non-responders, at least 5 of the 10 patients in the T_3 group and 6 of the 13 patients in the T_4 group would have needed to be classified as treatment responders to obtain a statistically significant difference in response rate using $\alpha = 0.05$. This is an even bigger problem when comparing the two active thyroid hormone treatment groups. With these sample size limitations it is even more important to use methods that take advantage of all available data.

These issues indicate that many challenges are unaddressed by the originally proposed analysis plan. While some challenges – such as data sparsity, data irregularity, and the underlying erratic nature of bipolar disorder – are unavoidable, other obstacles may be addressed by changes to the analytical approach. Information present in the data may be better analyzed by adopting a perspective that provides a flexible treatment efficacy definition, intelligently processes and combines the mood scale score measures, and takes advantage of the longitudinal structure of the data. Presented next is an approach that incorporates these ideas by creating an analytical framework centered around the mood episodes and the transitions among them.

4.2 Proposed Approach

4.2.1 Overview

Rather than framing the analysis around the pre-post changes of the HDRS and YMRS scores, the approach proposed here focuses on the underlying construct that the scores are trying to measure: the mood states themselves. Bipolar disorder is clinically described in terms of depressive, manic, mixed, and euthymic periods, with subcategories such as hypomania and subthreshold depression also sometimes considered. During these periods, patients are perceived to be in a mood state with separate states specified for each period type. For example, patients in a major depressive episode are said to be in a depressed mood or in a state of depression. These mood states form the foundation of the proposed approach. By using the mood states as the building blocks of the analysis, the clinical perspective of episode types is maintained, which has the advantage of making the results more directly clinically relevant, interpretable, and applicative.

It should be noted that there are formal clinical criteria to identify depressive, manic, mixed, and euthymic periods (as described in the previous chapter), but unfortunately these ratings were not available in the three-arm trial. Therefore, to construct mood state equivalents, the quantitative symptom measures are processed to extract the underlying qualitative mood states. In the three-arm randomized trial, mood is captured by two measures: the HDRS measures depressive symptom severity while the YMRS captures manic symptom severity. Using these two measures, we create four "mood states" – euthymic, mixed, depressed, and manic – as follows. Euthymia is characterized by normal mood levels. Low symptom scores on both the HDRS and YMRS measures indicate an absence of depression and mania, respectively, and therefore imply a euthymic mood state. Depressive episodes are naturally characterized by elevated HDRS scores, and are differentiated from mixed episodes by the absence of manic symptoms. Therefore, high HDRS and low YMRS scores imply a depressed mood state. Similarly, low HDRS and high YMRS scores imply a manic mood state. Finally, a mixed episode is characterized by the presence of both depressive and manic symptoms. Therefore, elevated mood symptom scores on both the HDRS and YMRS indicate a mixed mood state. This approach naturally combines the outcomes from two separate rating scales into a single measure. This process as applied to the three-arm randomized trial is summarized in Figure 4.1.

It is important to note that while I refer to these states as euthymic, mixed, depressed, and manic, they are categories derived from symptom levels and not clinically defined criteria. Therefore, caution must be exercised when interpreting the results and applying them in a strict, clinical setting.

The obvious question for the above approach is what threshold values to use to differentiate low scores from high scores. Absolute thresholds outlining the interpretation of mood rating scale scores have been recommended by the American Psychiatric Association[51] and the National Institute for Health and Clinical Excellence[52], although consensus in the literature is unclear[49,54]. A review suggests that there is a general agreement that HDRS scores less than 8 indicate a non-depressed state[53]. Absolute thresholds for mania are more uncertain, with some studies interpreting YMRS scores less than 8 as a non-manic state[56,57], while others are more conservative and use a threshold of less than 6[58,59]. It is also possible to use thresholds based on other severity levels depending on sensitivity desired. For example, if the primary concern was severe depression, a higher HDRS cut-off can be employed. Similarly, lower values can be used to favor the detection of less severe forms, such as hypomania or subthreshold depression. Moreover, there is tremendous variation from patient to patient in the mood scores of our study, raising the question of whether an absolute threshold value is even appropriate. For example, two individuals may both exhibit HDRS scores less than 8 when in euthymia, but one patient may naturally register a score of 6 while the other may register a score of 2. Furthermore, patients in this study are treatment-refractory and may be subject to both greater severity and more erraticness of mood swings. In this case, the importance of detecting incremental mood changes (symptom reduction) may supersede the goal of episode remission (symptom elimination). In light of these considerations, patient-specific thresholds may be preferable to absolute thresholds. This has the advantage of focusing the analysis on change relative to the patient's unique mood episode history and accounts for the high patient-to-patient variability. However, it comes at the cost of being unable to directly match the results to clinical definitions of the four mood states in relation to absolute HDRS and YMRS thresholds. Therefore, careful interpretation is necessary before extending the findings to strict clinically-defined settings.

To maintain the longitudinal structure of the data, mood states are inferred from all available data across time. This results in the determination of a mood state at each observation point. However, due to data sparsity and irregularity, the amount of time between each observation may vary and it is unclear how long a patient may be in a particular mood state. To overcome these issues and to impose data regularity, I first linearly interpolated between the available depression and mania mood scores for each subject. After this interpolation, mood states were determined at fixed intervals using the interpolated data. This data processing resulted in a continuous chain of mood states for each patient, summarizing the longitudinal depression and mania scores into a mood state sequence. Full details of the procedure are given in Section 4.3.

With the symptom scale scores reconstructed as mood state sequences, a state-based analytical framework can be adopted to describe the underlying mood dynamics. Data in this structure lend themselves to a Markov chain approach[68]. Details of Markov chain theory are presented later, but the basic idea is as follows. Under the Markov chain paradigm, subjects are assumed to be moving among a fixed set of states over time. Using a counting process, state-shifting patterns are described as probabilities of transitioning from one state to another. These probabilities are then used to characterize the underlying dynamics of the data. Such state-based approaches have been used previously to examine movements among health states in schizophrenia and unipolar depression[69] and have inspired the approach discussed here.

One challenge that is not addressed by the Markov chain approach is the study's small sample size and the barrier it presents in statistical inference. While the Markov chain describes the mechanisms underlying the mood-transitioning dynamics, it does not automatically provide a means of statistically identifying mood differences. Therefore, it is up to the investigator to determine the treatment efficacy measure and how to appropriately detect group differences. The small sample size coupled with the bipolar disorder's inherent complexities makes it difficult to identify and validate an appropriate theoretical sampling distribution for hypothesis testing. To overcome this barrier, I chose to employ a bootstrap resampling technique[70] to approximate the underlying sampling distribution. Details of bootstrap theory are presented later, but a brief outline is as follows. In the bootstrap method, same-size replicates of the dataset (called bootstrap samples) are created by randomly sampling from the data with replacement. These datasets are analyzed and summary statistics of interest – in this case, the treatment efficacy measures – are obtained for each dataset (called bootstrap estimates). Collectively, these estimates form the underlying sampling distribution. In the bootstrap method, inference is based on the available data and does not rely on an assumed, and possibly incorrect, distributions. Additionally, the bootstrap easily accommodates complex and multidimensional treatment efficacy measures since the analysis performed on the original dataset is identical to that used on the bootstrap samples. This is especially beneficial in bipolar disorder studies because improvements may manifest in many ways and across multiple mood states.

In summary, the proposed approach adopts a state-based perspective and processes the HDRS and YMRS scores using patient-specific thresholds to extract the underlying mood state sequences over time. This preserves the longitudinal structure of the data and the underlying mood dynamics. The sequences and mood-shifting behavior are summarized using Markov chain principles. To detect differences in mood, a bootstrap approach is used to obtain the underlying sampling distribution and provides a means of determining statistical significance. Discussed next are the theoretical foundations and implementation details behind Markov chains and the bootstrap, followed by their application to the three-arm randomized trial.

4.2.2 Markov Chains

A Markov chain is a mechanistic model of a stochastic process characterized by three pieces of information: (1) the set of possible states, (2) the transition probabilities, and (3) the Markov chain time unit. Each of these components are described in detail below, followed by model assumptions and the key measures we will derive from them to examine treatment effects.

In a Markov chain, data are assumed to be generated by a mechanism that is encompassed by a finite set of discrete states. The data may consist of direct observations of the states themselves, or they may be complex or indirect manifestations of the states. A natural choice for the state set in bipolar disorder is the mood episodes: euthymia, depressed, manic, and mixed. Symptom data collected in bipolar disorder studies are manifestations of these mood periods and can be used to classify patients in states across time. In the case of the three-arm randomized trial, the mood states manifest through the HDRS and YMRS scores. I reversed-engineered the symptom score trajectories into mood state sequences. Full details of the procedure are given in Section 4.3.

The main interest of the Markov chain model is the dynamic behavior among the states. While the states may describe the structure of the underlying data mechanism, the dynamic behavior among the states describes how that mechanism operates. The dynamics of the Markov chain are summarized by transition probabilities. These describe the likelihood of shifting from one state to another – or remaining in the same state – conditional on the current state. The formulation of these probabilities is based on the memoryless property of Markov chains, which is described in detail later. In the case of bipolar disorder, these values describe the likelihood of switching from one mood state to another. This naturally translates into important characteristics of mood. For example, the probability of transitioning from the euthymic state to the depressed state describes the onset probability of a depressive episode and inherently the expected amount of time that elapses between episodes. Similarly, the probability of transitioning from the manic state to the euthymic state describes the probability of remission from a manic episode. Similar interpretations can be made for all other transition probabilities, including the probability of remaining in a specific mood state. Because the state set is finite, these probabilities can be summarized as a matrix whose entry at row i and column j corresponds to the probability of transitioning from state i to state j in a given time window.

Implicit to these probabilities is the time frame in which the transitions occur. The likelihood of transitioning from one state to the next depends on the length of the observation period. Therefore, specification of the Markov chain requires the transition time frame to be specified. This time frame is referred to here as the chain time unit. Choosing the chain time unit requires careful consideration. For very stable systems, transitions between states may be less likely to occur and are not efficiently described by short time frames. Conversely, highly active systems may experience multiple transitions in a short time span. Such systems are inadequately described by long time frames. Selection of an appropriate chain time unit depends on the scientific nature of the underlying mechanisms, the granularity of the data available, and the research question of interest. As an example, weather patterns in tropical climates may be highly volatile and a chain time unit of one hour may be theoretically appropriate. However, weather data may only be available on a day-to-day basis, preventing an hour-by-hour analysis. Unlike weather reporters, meteorologists may be interested in weekly patterns, suggesting that a chain time unit of one week is best. Similar considerations must be made when selecting a chain time unit in the analysis of bipolar disorder data. In the three-arm randomized trial, factors to consider include the study population (treatment-refractory rapid cycling bipolar disorder patients), the degree of data sparsity and irregularity, and the overall data volume. Choosing the chain time unit for rapid cycling patients is especially tricky since cycling rates can vary dramatically from person to person. Once the chain time unit is selected, the data can be processed to construct state sequences and transition probabilities can be calculated by counting the number of transition types.

The Markov chain approach makes two important assumptions. First, it assumes that the specified state set is exhaustive. That is, all possible states of interest are encompassed by the state set. With an exhaustive state set, all possible transition probabilities are specified and the process can theoretically continue for an indefinite period of time. Second, it assumes that the mechanism is memoryless[68], namely that, the transition behavior at any point depends only on the current state and not on prior states. We will assume such a "first-order" chain structure in subsequent analysis. The memoryless property assures that the mechanism is sufficiently described by a single transition matrix and mechanisms based on prior states need not be specified.

These assumptions carry important implications in applying a Markov chain model to bipolar disorder data. The construction of an exhaustive state set is straightforward as the disease can be sufficiently described by the four mood states, but alternative configurations for the state set are possible. For example, each of the non-euthymic states may be further divided into severity levels, such as subthreshold depression or hypomania, and each level may be treated as a separate state. The memoryless property assumes that state transitions depend only on the current state and not on prior states. This may not hold, especially when states are subdivided by severity levels, as certain patterns may be informative of transition behavior. For example, a moderately depressed state that is preceded by a mildly depressed state may be informative of an escalating episode, thereby making a severely depressed state more probable. Alternatively, the data may be characterized into two states: euthymic and non-euthymic. While this provides a simpler structure, the specifics of the non-euthymic states are lost. These examples suggest that careful attention must be applied when specifying the mood state set.

A similar issue arises when considering episode duration. The memoryless property indicates that the transition probabilities do not depend on the mood state sequence preceding the current state. This implies that mood-shifting dynamics are not dependent on the duration of the current mood state. As an example, this assumption suggests that a person in the manic state has an equal probability of transitioning to the euthymic state regardless of how long he or she has been in the manic state. That is, a person who has been in the manic state for 10 weeks has an equal probability of shifting to the euthymic state as a person who has been in the manic state for only 1 week. The appropriateness of these assumptions must be considered when performing a Markov chain analysis.

If these assumptions hold, then the Markov chain model can be used to describe the data's state dynamics in a variety of ways. In addition to the transition probabilities, one can obtain the average time it takes to transition from state i to state j (referred to as the mean first passage time) and the average time it takes to return to state i (referred to as the mean recurrence time)[71]. Of particular interest in the analysis discussed here is the stationary distribution. The stationary distribution describes the long-run total proportion of time subjects will spend in each state if the transition process described by the Markov chain model continues indefinitely[68]. Such distributions exists for a class of Markov chains knowns as ergodic chains[71]. For ergodic chains, the possibility of state i preceding state j in a state sequence exists for all pairwise combinations of states. That is, it is possible to eventually arrive at state j after departing from state i. This is immediately apparent for a transition matrix with all non-zero entries, but may be less obvious for transition matrices with multiple zero entries. If a Markov chain is ergodic, then the stationary distribution is obtained by solving a system of equations based on the transition probabilities. It can be approximated by repeatedly multiplying the transition matrix by itself until the desired precision is reached. In the context of our treatment trial, the stationary distribution describes the proportion of time spent in each mood state if the observed mood-shifting dynamics continue indefinitely, that is, no change in treatment or life circumstances. This provides a measure for comparing the transition processes of mood among groups of interest, and for seeing the effects of treatment.

4.2.3 Bootstrap

Next we present details of the bootstrap method. In the approach presented here, the bootstrap technique is used to construct multiple bootstrap samples. Each sample is analyzed as a Markov chain and measures of mood dynamics are obtained to form the underlying sampling distribution.

The bootstrap[72] is a non-parametric simulation method developed to assess the accuracy of a parameter estimate by empirically approximating its underlying sampling distribution. For many standard statistical approaches, the accuracy of an estimate is evaluated on the basis of distributional assumptions. For example, in classical mean estimation, the sampling distribution of the mean is assumed to be normally distributed and forms the foundation for hypothesis testing. This normality assumption is based on the Central Limit theorem which states that the sampling distribution of the mean asymptotically converges to a normal distribution for sufficiently large samples. While distributional assumptions such as this may be statistically appropriate in many scenarios, in others the assumptions may be questionable. In these cases it is difficult to evaluate the accuracy of the parameter estimate. Simple examples that demonstrate this are median estimation and the estimation of ratio-based estimates[73]. For more complex estimators, the distribution may not even be analytically derivable. In such cases, choosing to impose distributional assumptions may misrepresent the true accuracy of the estimate and can potentially confound the study's findings. In situations such as these, the bootstrap provides a means of constructing a parameter's estimated sampling distribution to fairly assess the estimate's accuracy. In our study we will use the bootstrap to evaluate the accuracy of parameters describing the mood-shifting mechanisms, such as those obtained under the Markov chain model. Due to the complexity of the disease, the distributions of such parameters are unknown. In the three-arm randomized trial, the bootstrap is especially useful because the appropriateness of an assumed distribution is difficult to evaluate in a small sample size setting.

Although the bootstrap can be applied to very complex scenarios, the method itself is relatively simple. The bootstrap involves three steps: (1) creating a bootstrap sample, (2) computing a bootstrap estimate, and (3) repeating steps 1 and 2 multiple times to form the parameter estimate distribution. The creation of a bootstrap sample involves sampling from the data with replacement. Conceptually, the bootstrap treats the current sample as a representative surrogate of the population and redraws samples from it. For a sample of size n, a new sample is created by drawing a single observation, then replacing it back into the original sample, then repeating this process until a full sample of size n is created. This new sample is referred to as a bootstrap sample. Because sampling is done with replacement, it is possible for an observation to appear multiple times within the bootstrap sample, while other observations may be omitted. The parameter of interest is then estimated for the bootstrap sample and is referred to as a bootstrap estimate. A virtue of the bootstrap lies in the fact that only the point estimate needs to be calculated. For many analyses, the point estimate is easy to calculate whereas the distribution of the estimate is difficult to obtain. The process of creating a bootstrap sample and obtaining a bootstrap estimate is repeated a large number of times (typically in the order of hundreds or thousands) to form a distribution. This distribution can be used to estimate the standard error of the original estimate, determine confidence intervals, or test hypotheses. This version of the bootstrap technique is known as the non-parametric bootstrap because the algorithm depends only on the data themselves and not on pre-specified population parameters or distributions.

Another version of the bootstrap, aptly known as the parametric bootstrap technique, aims to obtain the sampling distribution if pre-specified population parameters were known. Rather than resampling from the data to create bootstrap samples, the parametric bootstrap generates bootstrap samples based on population parameters. This often includes distributional parameters such as means, variances, and correlations, as well as sampling restrictions such as sample size or follow-up time. Once a bootstrap sample is created, the parametric bootstrap algorithm follows the same steps as the non-parametric version: parameters of interests are calculated for each bootstrap sample and those estimates are collectively analyzed to describe the estimate's distribution. As a simple example, consider a sample of n numbers that are assumed to be drawn from a normal distribution. The mean $\hat{\mu}$ and variance $\hat{\sigma}^2$ can be estimated from this sample, and parametric bootstrap samples can be drawn from a normal distribution parameterized by these estimates. The parametric bootstrap draws observations similar in distribution (though not necessarily identical in value) to that of the original sample. This makes use of the distributional assumptions imposed which, if correct, result in greater accuracy.

Though its simplicity and ease of implementation are appealing, care must be exercised when using the bootstrap and its underlying assumptions must be considered[74]. First, the method assumes that the originally drawn sample properly represents the population of interest. The method relies on the idea that the original sample is drawn from the population without bias and that it appropriately encompasses the breadth of information present in the population. Because the bootstrap draws from the original sample and creates new samples that are intended to be representative of the population, it is important that the parameter estimates, underlying relationships, and distributions present in the original sample appropriately and sufficiently reflect the characteristics of the population. Therefore, careful attention must be placed on ensuring that the original sampling procedure is free from bias. This assumption presents challenges for small samples – as is the case our study – since the data may not adequately represent the breadth of information or may be sensitive to spurious results. In such cases, the investigator must either assume that the information of interest is indeed contained within the small sample or accept the limitations attached to small samples. In the case of the parametric bootstrap, it is assumed that the underlying data-generating structure is sufficiently described by the parameters used to generate bootstrap samples and all specified values are correct. Second, application of the bootstrap must carefully consider the nature of the data, the parameter of interest, and how it is constructed. This includes the complexities of the sampling procedure, structural dependencies within the data, distributions with unique characteristics, and the choice of what unit to resample (e.g., subjects, single observations, single variables, etc.). This suggests that application of the bootstrap must respect the confines of the underlying data mechanisms and the limitations associated with the study's methodology. Examples of these scenarios and the associated shortcomings are discussed in further detail in other texts, as well as adaptations to the bootstrap to overcome some of these issues[73].

4.3 Application

Next we present the methodological details regarding the application of the Markov chain and bootstrap techniques to treatment studies in general and the three-arm randomized trial in particular. Modeling the three-arm randomized trial data as a Markov chain first requires the state set to be defined. Using the HDRS and YMRS scores, data are classified into the following four states: (1) *diminished depressive and manic symptom severity levels*, (2) *elevated depressive symptom severity levels only*, (3) *elevated manic symptom severity levels only*, and (4) *elevated depressive and manic symptom severity levels.* These states are intended to qualitatively correspond to the four standard mood states of bipolar disorder: euthymic, depressive, manic, and mixed, respectively. We will loosely use those terms in what follows.

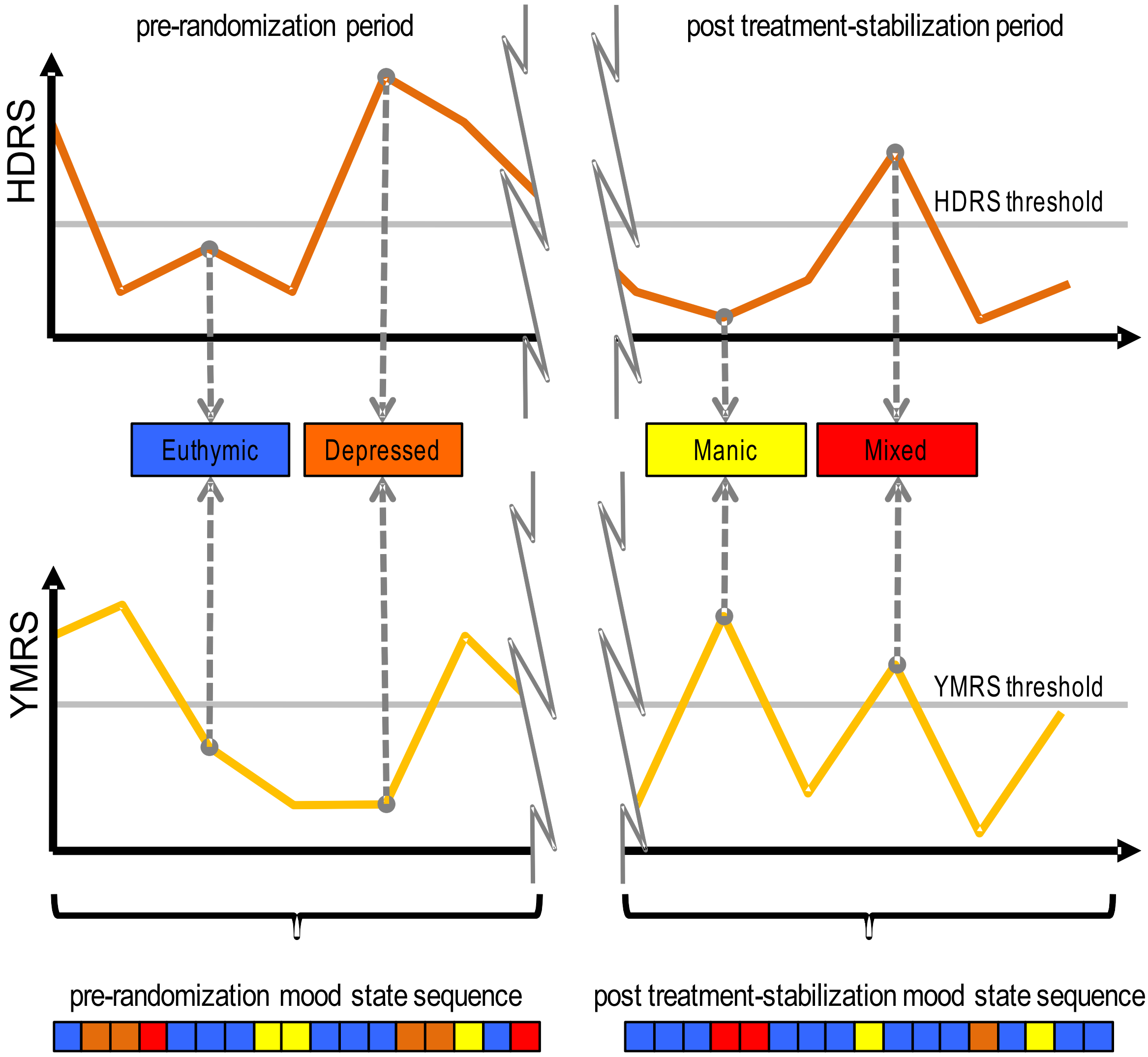
Because the sample describes a treatment-refractory population and exhibits a high degree of between subject heterogeneity, patient-specific thresholds for determining mood states are used rather than absolute thresholds. Specifically, the results presented in this chapter use each patient's median HDRS and YMRS scores during the pre-randomization period as the cut-off. This provides an easily interpretable cut-point for describing the mood severity levels of patients before receiving treatment. Based on these thresholds, data are classified into the four states. Though this method of mood classification does not match clinically defined standards, for the sake of simplicity these four states are referred to as euthymic, depressed, manic, and mixed for the duration of this chapter. The classification process is summarized in Figure 4.1.

Although the median is used as the threshold in this analysis, it is important to note that any threshold can be used in this analytical framework. The threshold can be as simple as the mean pre-randomization score, or more complex, such as the midpoint of the mediods based on a clustering algorithm. This offers flexibility in appropriately characterizing features based on research interest or the nature of the disease. For example, using the mean score suggests that research interests lie in changes in a patient's average severity, while using a clustering-based threshold suggests that symptom severities are reflective of an underlying categorical structure. For this particular study, investigations into these other thresholds yielded qualitatively similar results as the median-based threshold.

To address the sparsity and irregularity present in the data, the HDRS and YMRS scores of each patient are linearly interpolated. Based on this interpolation, patients are classified into mood states at one-week intervals. This results in a sequence of mood states with a chain time unit of one week. Two state sequences are formed for each patient. The first sequence corresponds to the pre-randomization period, while the second sequence corresponds to the post treatment-stabilization period. These sequences serve as the underlying signal describing each patient's unique mood-shifting dynamics with and without thyroid hormones.

The selection of a Markov chain time unit of one week is based on the study's protocol and the granularity of the data. The study was originally designed to collect data on a weekly basis. Although weekly data are unavailable due to limitations described earlier, data are available approximately every two weeks. It is assumed that weekly mood states can reasonably be inferred with bi-weekly data because mood will change relatively smoothly and

Figure 4.1: Mood State Classification


pre-randomization period
post treatment-stabilization period
HDRS
HDRS threshold
Euthymic
Depressed
Manic
Mixed
YMRS
YMRS threshold
pre-randomization mood state sequence
post treatment-stabilization mood state sequence

Thresholds are based on the median of the pre-randomization period data and are patient-specific. Data are classified into mood states relative to this threshold. Dots represent observation points and how they are classified according to their placement relative to the threshold.

continuously. Furthermore, while it is acknowledged that other interpolation methods may be used, the choice of linear interpolation in this analysis is selected based on its simplicity. The basic analytical framework is not restricted to linear interpolation and other methods may be employed, although investigations into these methods suggest that the degree of data sparsity and irregularity present in the three-arm randomized trial present its own set of challenges in the application of more complex shapes. Additionally, with data points spaced two weeks apart on average, it is reasonable based on our experience with these data that linear interpolation does not grossly overstate data granularity and that the potential bias introduced by this interpolation method for inferring mood states at one week intervals is not significant.

To capture the mood-shifting dynamics of the patients within a treatment group, state sequences are aggregated to form a transition matrix in order to examine efficacy. This matrix is constructed by tallying the number of times patients transition from one state to another, as well as the frequency with which they remain in the same state from week to week. These counts are used to form conditional probabilities describing the weekly transition behavior. Separate transition matrices are formed for each study period and treatment group. With three study arms in the trial, this results in three transition matrices describing the mood dynamics during the pre-randomization period and three transition matrices describing these mood dynamics during the post treatment-stabilization period. This information can also be globally summarized by the stationary distribution of each group for each of the two study periods. The stationary distribution describes the proportion of time spent in each mood state if the mood patterns described by the transition matrix continue indefinitely.

To test whether the memoryless property of first-order Markov chains is violated, a χ^2-based goodness-of-fit test is employed to detect second-order dependency[75]. First-order chains assume that the transition probabilities into the next state depend only on the current state. Second-order chains allow dependencies on the current state and the state immediately prior. This is tested by comparing the observed frequencies of triplet sequences (i.e., a sequence that is three chain units in length) to the expected frequencies based on the first-order transition matrix. The test statistic is calculated as follows:

$$\chi^2_{df=s(s-1)^2} = \sum_{ijk} \frac{(n_{ijk} - n_{ij}P_{jk})^2}{n_{ij}P_{jk}} \tag{4.1}$$

In equation (4.1), n_{ijk} is the observed frequency count of a triplet sequence that goes from state i to state j to state k. Similarly, n_{ij} is the observed frequency count of sequence pairs that go from state i to state j. P_{jk} is the entry in the transition matrix that describes the probability of transitioning from state j into state k. Asymptotically, this test statistic has a χ^2 distribution with $s(s-1)^2$ degrees of freedom where s is the number of states in the state set. Tests are conducted separately for each treatment group and by study period. a statistically significant p-value suggests that there is evidence of second-order dependencies and the memoryless property for first-order chains is violated.

Two types of measures of treatment efficacy are calculated based on the stationary distributions. The first type is the within-group change from the pre-randomization period to the post treatment-stabilization period. This is calculated as the difference in the time spent per the stationary distribution for a given treatment group and state and is measured in percentage-points. That is, for each state, the proportion of time spent in that state according to the pre-randomization mood pattern is subtracted from the proportion of time spent in that state according to the post treatment-stabilization mood pattern. This results in a difference of proportions. A within-group difference of zero indicates no treatment effect within the group. A difference greater than zero indicates that more time is spent in the given mood state during the post treatment-stabilization period. A difference less than zero indicates that less time spent in the given mood state during the post treatment-stabilization period. For example, suppose that the stationary distribution results indicate that the T_4 group spends 30% of the time in the depressed state based on the pre-randomization data, whereas the post treatment-stabilization data indicates 10%. This results in a within-group difference of -20%, meaning that T_4 participants experience a 20 percentage-point decrease in the time spent in the depressed state. For the euthymic state, an increase in within-group difference indicates a favorable treatment effect. For non-euthymic states, decreases in within-group differences indicate a favorable treatment effect.

The second type of efficacy measure is the between-group difference in treatment effects. This is calculated by subtracting the within-group change for one study arm from the within-group change of another study arm for a given state and is again measured in percentage-points. This measure represents a group by time interaction designed to get at differential treatment effects. A value of zero indicates no difference in treatment effects between the two groups. A value greater than zero indicates that the change observed in the first group is positively greater than the change observed in the second group. A value less than zero indicates that the change observed in the first group is negatively greater than the change observed in the second group. For consistency, the thyroid treatment group is always selected as the first group in these comparisons. When comparing the two thyroid treatment groups, T_4 is selected as the first group. Building on the previous example, suppose that a within-group difference of -5% is observed in the PL group for the depressed state. Comparing the T_4 group (with a within-group difference of -20%) to the PL group results in a between-group difference of -15%. This indicates that the change observed in the T_4 group is negatively greater than the change observed in the PL group. This implies that more T_4 patients have a greater improvement (decrease in time spent) in the depressed state than PL patients. For comparisons between the treatment groups and the PL group, a value less than zero (a negatively greater difference) is favored for the depressed, manic, and mixed states because it indicates that the treatment group is shifting *out* of these states to a greater degree or at a greater rate than PL patients. Similarly, a value greater than zero (a positively greater difference) is favored for the euthymic state as it indicates that the treatment group is shifting *into* a euthymic state at a higher rate than PL patients and also implies a beneficial treatment effect.

While the examples described above compare the mood states individually, it must be noted that these comparisons are not truly independent. Changes in one state is suggestive of changes in other states. These relationships may be lost when analyzing the stationary distribution separately by state. For example, significant increases in the euthymic state suggest significant decreases in the non-euthymic states. However, this corresponding decrease may be spread throughout the other states, which may result in smaller, less-detectable differences in state-by-state comparisons. Therefore, these dependencies must be considered when interpreting results.

We use a non-parametric bootstrap approach to evaluate statistical significance. A total of 10,000 bootstrap samples are created. Treatment efficacy measures are calculated for each bootstrap sample and distributions for each measure are constructed using the bootstrap estimates. Statistical significance is determined as follows. For distributions that are largely to the right of zero, the proportion of bootstrap estimates less than zero is calculated. For distributions largely to the left of zero, the proportion of bootstrap estimates greater than zero is calculated. These proportions are then multiplied by two to denote a two-sided test, and the resulting value reflects the p-value of the estimated effect. For within-group changes, a statistically significant result implies a treatment effect for the given mood state. For between-group differences, a statistically significant result implies a differential treatment effect between the two groups for the given mood state. Because the thyroid treatment group is always selected as the first group when measuring between-group differences, bootstrap interpretations are similar for both within-group changes and between-group differences. For the euthymic state, within-group changes and between-group differences greater than zero imply positive treatment effects and are depicted by a distribution that is largely to the right of zero. Similarly, for the non-euthymic states, values less than zero for both efficacy measures imply beneficial treatment effects and are depicted by a distribution that is largely to the left of zero. Additionally, the parametric bootstrap is employed using the observed transition matrices of each study period for each treatment group to generate mood state sequence data. Each bootstrap sample generated by the parametric bootstrap reflect the treatment group sample sizes and follow-up times of the original data. An agreement between the non-parametric and parametric bootstrap results suggest that the underlying sampling distribution of the original data is reflective of the distribution described by the observe transition matrices.

4.4 Results

The goodness-of-fit test described by equation (4.1) for the mood state sequences resulting from the process described by Figure 4.1 did not show violations of the memoryless property of first-order Markov chains ($p > .05$ for all comparisons). Table 4.1 summarizes the amount of time spent in each mood state according to treatment group by study period. The observed time is measured in weeks and is calculated by tallying the mood states along the mood state sequences. Although there is some variability, the proportions of time spent in each state during the pre-randomization period are roughly equal for all three study groups. The largest difference observed is in the T_4 group. According to the frequency counts, T_4 patients spend slightly more time in non-euthymic states (81.1%) compared to the PL and T_3 groups (72.0% and 72.1%, respectively). Differences appear when examining the time spent in the mood states during the post treatment-stabilization period. Both the T_3 and T_4 groups spend the majority of the time in the euthymic state (45.5% and 50.4%, respectively), while the PL group is largely unchanged (28.0% during the pre-randomization period versus 26.6% during the post treatment-stabilization period). These observed increases in the time spent in the euthymic state suggest that T_3 and T_4 may carry a favorable treatment effect.

The observed times in Table 4.1 summarize the actual amount of time spent by subjects in each mood state, but this does not provide information about transition rates or the evolution of the dynamic mood process. The observed proportions reflect the starting distribution which may not be equivalent to the one that arises from transitioning dynamics. To incorporate the mood-shifting mechanisms in the evaluation of the time spent in each mood state, we compute the transition matrix and from it the long-run stationary distribution. For the pre-randomization period, the long-run behavior describes the eventual mood patterns with no treatment intervention. For the post treatment-stabilization period, the long-run

Table 4.1: Time Spent in Mood States

		pre-randomization			post tx-stabilization		
		PL	T_3	T_4	PL	T_3	T_4
observed time (weeks)	Eut	33	39	30	34	71	128
	Dep	25	30	51	18	29	33
	Man	31	41	48	48	35	77
	Mix	29	30	30	28	20	16
	total	118	140	159	128	155	254
observed time (%)	Eut	28.0	27.9	18.9	26.6	45.8	50.4
	Dep	21.2	21.4	32.1	14.1	18.7	13.0
	Man	26.3	29.3	30.2	37.5	22.6	30.3
	Mix	24.6	21.4	18.9	21.9	12.9	6.3
stationary distribution (%)	Eut	27.7	29.8	14.7	21.2	45.8	47.8
	Dep	21.9	19.3	29.1	14.6	16.6	11.0
	Man	23.7	29.3	35.3	28.1	23.7	33.6
	Mix	26.8	21.6	20.9	36.0	13.9	7.6

The observed time is calculated by tallying the number of weeks subjects spent in each mood states from their derived mood state sequences. The stationary distributions are calculated using the transition matrices resulting from the mood state sequences.

behavior describes the eventual mood patterns under continued treatment. The stationary distributions are displayed in Table 4.1. Note that there are indeed some differences between the observed times and the stationary long-run distribution results, particularly the time spent in the mixed state during the post treatment-stabilization period for the PL group (21.9% based on the observed time versus 36.0% based on the stationary distribution). As would be hoped, the results of the stationary distribution analysis indicates increases in the time spent in the euthymic state within the thyroid hormone treatment groups, suggesting that T_3 and T_4 may carry a favorable treatment effect.

Displayed in Table 4.2 are the transition matrices for each group by study period. Transition matrix comparisons within the PL group suggest that subjects are even more likely to persist in manic and mixed states during the post treatment-stabilization period than pre-randomization. The probability of remaining in the manic state in the following week increases from 67.7% to 82.6% according to the pre-randomization and post treatment-stabilization transition matrices. For the mixed state, an increase from 70.8% to 88.0% is observed. Within the T_3 group, changes in mood patterns are mostly beneficial. The probability of remaining in the euthymic state increases from 64.9% to 76.5%, while the chance of transitioning from the depressed state to the euthymic state increases from 17.2% to 30.8%. However, an increase in mixed state persistence is also observed, from 51.9% to 61.1%. Favorable differences are observed within the T_4 group, particularly for the euthymic state. Persistence of the euthymic state increases (+39.7%), while shifts to the depressed and manic state decrease (-21.9% and -16.0%, respectively). Additionally, shifts from the mixed state to the euthymic state increase from 3.3% to 15.4%. An increased manic state persistence is observed (+11.5%), although this may be offset by a decrease in shifts to the mixed state (-13.0%). Overall, these findings suggest that the PL group may be experiencing prolonged manic and mixed episodes, while the treatment impact of the T_3 group may be related to increased periods of euthymia and shifts from the depressed state to the euthymic state. For the T_4 group, the treatment impact appears to be related to a highly persistent euthymic state and a tendency to shift out of the mixed state and into the euthymic state. However, it must be acknowledged that some transitions are rare and the observed estimates presented in Table 4.2 may be unstable.

Treatment efficacy measures based on stationary distribution differences are summarized in Table 4.3. Sampling distributions of these measures resulting from the non-parametric bootstrap method are displayed in Figures 4.2 and 4.3. Results provide no evidence of a placebo effect (i.e. a within-PL group change) for any of the mood states. Within-group changes for the PL group range from -6.5 to 9.3 percentage-point and none of the differences are not statistically significant. This suggests that the mood-shifting dynam-

Table 4.2: Mood State Transition Matrices

		pre-randomization				post tx-stabilization			
		Eut	Dep	Man	Mix	Eut	Dep	Man	Mix
PL	Eut	67.7	16.1	16.1	0.0	68.8	9.4	18.8	3.1
	Dep	21.7	52.2	4.3	21.7	18.8	62.5	6.3	12.5
	Man	12.9	6.5	67.7	12.9	8.7	2.2	82.6	6.5
	Mix	4.2	16.7	8.3	70.8	4.0	8.0	0.0	88.0
T_3	Eut	64.9	10.8	16.2	8.1	76.5	10.3	13.2	0.0
	Dep	17.2	58.6	3.4	20.7	30.8	57.7	0.0	11.5
	Man	18.9	2.7	64.9	13.5	20.6	0.0	64.7	14.7
	Mix	7.4	18.5	22.2	51.9	5.6	16.7	16.7	61.1
T_4	Eut	44.8	27.6	24.1	3.4	84.6	5.7	8.1	1.6
	Dep	19.6	65.2	6.5	8.7	18.8	65.6	3.1	12.5
	Man	4.9	7.3	70.7	17.1	12.3	1.4	82.2	4.1
	Mix	3.3	16.7	23.3	56.7	15.4	7.7	23.1	53.8

These values represent the probability of transitioning from the state indicated by the row to the state indicated by the column. Separate transition matrices are provided for each study group according to the pre-randomization and post treatment-stabilization periods. Values are presented as percentages.

ics within the PL group during the pre-randomization period are not significantly different from those during the post treatment-stabilization period.

Within-group results are similar for the T_3 group. The within-group changes for non-euthymic states range from -2.7 to -7.7 percentage-point differences. While these decreases are favorable for non-euthymic states, they are too small to establish statistical significance. The within-group difference is most dramatic for the euthymic state with an observed increase of 16.0 percentage-points. Although this increase suggests that more time is spent in the euthymic state after T_3 treatment administration, the non-parametric bootstrap results fail to reflect statistical significance (p=.1834). However, with the majority of the non-parametric bootstrap distribution (90.8%) to the right of zero for the euthymic state (Figure 4.2), T_3 may prove to be a worthwhile treatment in a larger study setting.

The most dramatic within-group changes are observed within the T_4 group. Statistically significant effects are observed in the proportion of time spent in the euthymic, depressed, and mixed states (p=0.0022, 0.0022, and 0.0312, respectively). For the euthymic state, there is an estimated 33.1 percentage-point increase based on the original sample, with more time spent in the euthymic state after T_4 treatment administration. For the depressed state, there is an estimated 18.1 percentage-point pre-post decrease indicating that significantly less time is spent in depression. For the mixed state, there is a 13.3 percentage-point decrease post treatment, indicating that significantly less time is spent in the mixed state as well. Results for the manic state are inconclusive: the estimated decrease is only 1.6 percentage-points and the non-parametric bootstrap distribution does not indicate statistical significance.

Between-group differences and the corresponding sampling distributions resulting from the non-parametric bootstrap method are displayed in Table 4.3 and Figure 4.3. Results comparing the T_3 group to the PL group yield statistically inconclusive results. Nevertheless, the non-parametric bootstrap distributions suggest that T_3 may have a favorable effect over PL, as demonstrated by a euthymic state distribution that is largely to the right of

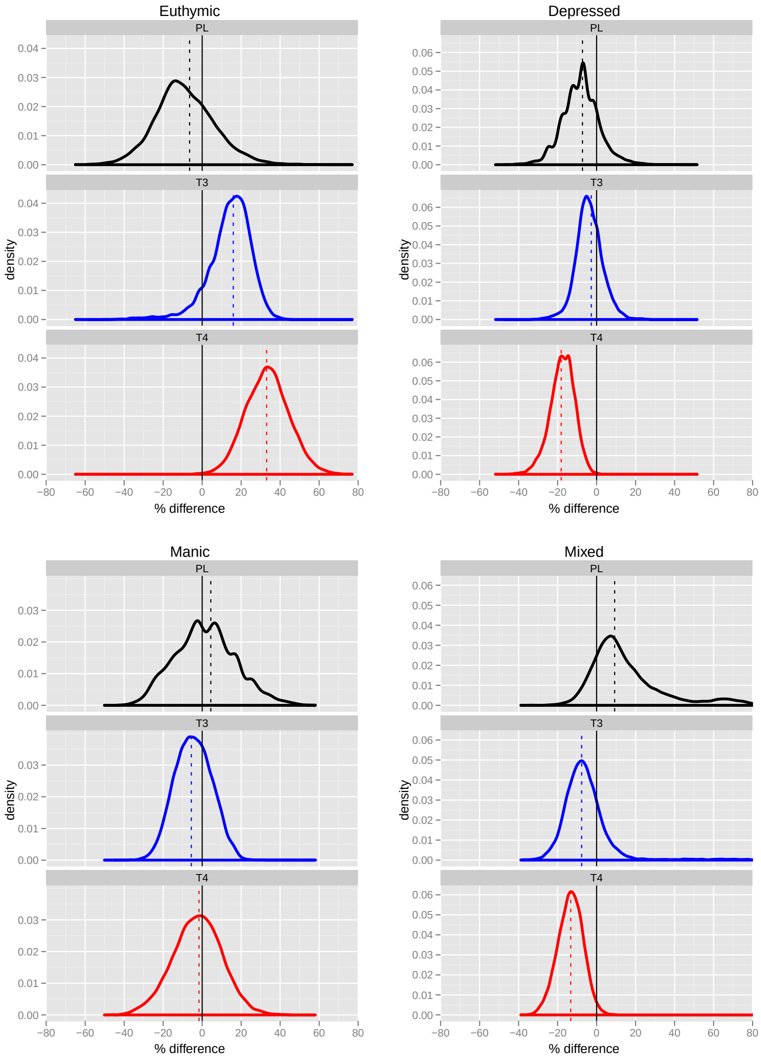
Table 4.3: Non-Parametric Bootstrap Treatment Efficacy Measures

	within-group differences					
	PL	p-value	T_3	p-value	T_4	p-value
Eut	-6.5	0.5246	16.0	0.1834	33.1	0.0022
Dep	-7.2	0.3358	-2.7	0.5348	-18.1	0.0022
Man	4.4	0.9532	-5.6	0.6458	-1.6	0.8590
Mix	9.3	0.3248	-7.7	0.3934	-13.3	0.0312

	between-group differences					
	T_3 vs PL	p-value	T_4 vs PL	p-value	T_4 vs T_3	p-value
Eut	22.4	0.2194	39.5	0.0328	17.1	0.1958
Dep	4.5	0.6930	-10.9	0.3840	-15.4	0.1216
Man	-10.0	0.7500	-6.1	0.8622	3.9	0.8988
Mix	-16.9	0.2160	-22.5	0.0454	-5.6	0.5402

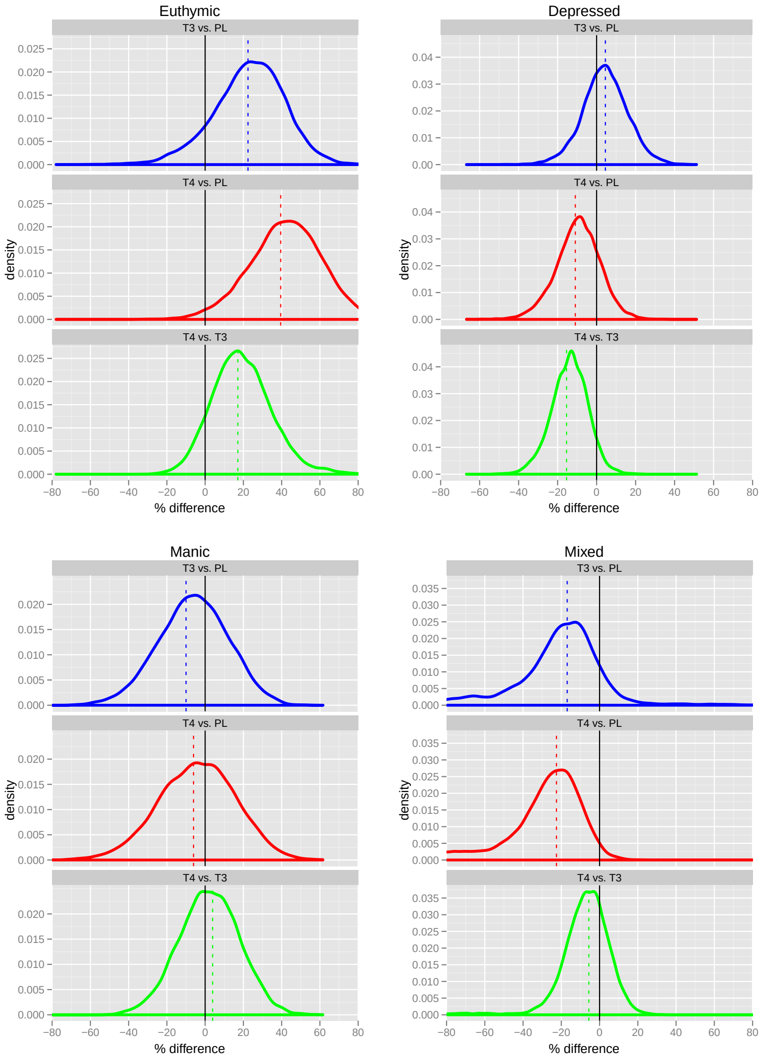
Within- and between-group differences are based on the stationary distribution results calculated from the state sequences of the original sample and are measured as percentage-point differences. P-values are based on 10,000 non-parametric bootstrap samples.

Figure 4.2: Bootstrap Distributions of Within-Group Differences



The non-parametric bootstrap distributions are based on the stationary distribution results of 10,000 bootstrap samples. The dashed lines represent the stationary distribution results of the original sample. For the euthymic state, a distribution to the right of zero suggests a favorable treatment effect. For non-euthymic states, a distribution to the left of zero suggests a favorable treatment effect. A significant result is obtained if the zero-line falls in the extreme tail of the bootstrap distribution.

Figure 4.3: Bootstrap Distributions of Between-Group Differences



The non-parametric bootstrap distributions are based on the stationary distribution results of 10,000 bootstrap samples. The dashed lines represent the stationary distribution results of the original sample. The competing treatment is the treatment that is listed second. For the euthymic state, a distribution to the right of zero suggests a favorable treatment effect that is greater than the competing treatment. For non-euthymic states, a distribution to the left of zero suggests a favorable treatment effect that is greater than the competing treatment. A significant result is obtained if the zero-line falls in the extreme tail of the bootstrap distribution.

zero (indicating more time spent in the euthymic state for the T_3 group compared to the PL group) and a mixed state distribution that is largely to the left of zero (indicating less time spent in mixed state for the T_3 group compared to the PL group). This suggests that further study is needed in a larger sample size setting to elucidate the impact of T_3 treatment.

Between-group comparisons of the T_4 group to the PL group reveal statistically significant differences. This is indicated by positively greater differences in the euthymic state (p=0.0328) and negatively greater differences in the mixed state (p=0.0454). This suggests that the changes observed in T_4 patients are more favorable than the changes observed in PL patients for the euthymic and mixed states. Comparisons involving the other mood states are suggestive of a positive treatment effect, but are not statistically significant.

Comparisons between T_4 and T_3 do not yield statistical significance for any mood state. However, the depressed state distribution is suggestive of an effect that favors T_4 over T_3. A similarly favorable distribution pattern is also observed in the euthymic state distribution. This suggests that T_4 may potentially carry an effect over T_3 for these mood states, although further study is needed.

Displayed in Table 4.4 are the results of the parametric bootstrap. These results reflect the p-values of bootstrap samples generated from the transition matrices of the original sample (Table 4.2). In general, these results confirm the findings based on Table 4.3 and support the efficacy patterns of T_3 and T_4. However, unlike the non-parametric bootstrap results, the treatment effect of T_4 over PL for the mixed state (a 22.5 percentage-point decrease) is no longer statistically significant (p=.2196 versus p=.0454). This suggests that the underlying sampling distribution may not fully match the distribution described by the original sample's transition matrices. The non-parametric bootstrap estimates the empirical distribution free of parametric assumptions, and is therefore the favored result in this setting, but the parametric bootstrap findings suggest that there may be interesting relationships that are not completely apparent in the transition matrices themselves, particularly in dynamics involving the mixed state and T_4 treatment, and requires further study.

Table 4.4: Parametric Bootstrap Treatment Efficacy Measures

	within-group differences					
	PL	p-value	T_3	p-value	T_4	p-value
Eut	-6.5	0.5600	16.0	0.1574	33.1	0.0002
Dep	-7.2	0.4014	-2.7	0.7274	-18.1	0.0260
Man	4.4	0.7802	-5.6	0.5642	-1.6	0.8764
Mix	9.3	0.6052	-7.7	0.3408	-13.3	0.0290

	between-group differences					
	T_3 vs PL	p-value	T_4 vs PL	p-value	T_4 vs T_3	p-value
Eut	22.4	0.1622	39.5	0.0092	17.1	0.2296
Dep	4.5	0.7174	-10.9	0.3802	-15.4	0.1894
Man	-10.0	0.5974	-6.1	0.7480	3.9	0.7942
Mix	-16.9	0.3818	-22.5	0.2196	-5.6	0.5812

Within- and between-group differences are based on the estimated transition matrices for the original sample and are measured as percentage-point differences. P-values are based on 10,000 parametric bootstrap samples.

In summary, within-group effects indicate statistically significant improvements in the T_4 group in the time spent in the euthymic, depressed, and mixed states. Potential within-group effects are also observed in the T_3 group for the euthymic and mixed states, although these findings are not statistically significant. Between-group effects favor T_4 over PL for the euthymic state and possibly for the mixed state as well, although due to the conflicting results between the non-parametric and parametric bootstrap, the latter effect requires further investigation. Comparisons between the T_3 and PL group are also suggestive of a favorable thyroid treatment effect in the time spent in the euthymic and mixed states, although these findings are statistically inconclusive and a larger study is necessary to increase statistical power. Comparisons of T_4 to T_3 suggest that an effect favoring T_4 may be present in the time spent in the depressed state, although this finding also fails to reach statistical significance.

It must be reiterated that, although comparisons are performed separately for each state, the effects based on the stationary distribution are not truly independent. Increases observed in one state are suggestive of decreases in another (potentially multiple). Therefore, while a large difference may be observed for individual state, corresponding minor or moderate differences may exist across other states.

4.5 Conclusions

The Markov chain results suggest that T_4 treatment positively affects the mood dynamics among treatment-refractory rapid cycling bipolar disorder patients. However, the treatment effect varies depending on the mood state. The results suggest that T_4 treatment may be most influential in increasing the time spent in the euthymic state. Of the four mood states, this is the most favorable, suggesting that adjunctive T_4 treatment may carry a highly desirable impact. The results also suggest a corresponding decrease in the time spent in the mixed and depressed states. Results suggest that the impact of T_4 on the manic state may be minimal, implicating that T_4 treatment maybe be less effective for patients whose disease is mainly characterized by periods of mania. However, interpreting these results as a whole indicate an overall shift into the euthymic state.

While statistically significant differences are observed within the T_4 group, differences are less pronounced when comparing the effects between treatment groups. This may be a result of the increased variance when comparing two highly variable measures. This makes the detection of differences between treatment groups more difficult and limits the statistical power to the detection of dramatically large differences. This is further exacerbated by the high variance and small sample size of the PL group. Nevertheless, patterns suggest potential effects than lend itself to further study. This includes the additional impact of T_4 over PL on the depressed state and the impact of T_3 over PL on the euthymic and mixed states. Patterns also suggest a possible edge that T_4 may have over T_3 in the time spent in the euthymic and depressed states.

The approach presented here offers a novel means of analyzing bipolar disorder data. By adopting an episode-based perspective, data can be translated into mood state sequences and mood-shifting dynamics can be modeled as a Markov chain. To detect treatment differences, the bootstrap can be employed and offers a means of statistically testing hypotheses that are free of distributional assumptions. However, this approach only describes the amount of time spent in the mood states. Though some timing information may be inferred from the transition matrices, specific characteristics of the mood episodes themselves are not directly captured by this approach. These include the degree of symptom severity, episode duration, episode frequency, and changes over time. Information containing these characteristics are lost when the data are translated into mood states and collapsed into a transition matrix. In the next chapter, I present an approach that aims to specifically capture these elements by developing a flexible model parameterized by these episode characteristics and fitting it to the HDRS and YMRS data using an iterative alogorithm known as particle swarm optimization.

CHAPTER 5

Approach 2: Particle Swarm Optimization

5.1 Background

In the Markov chain approach, data from bipolar disorder patients are translated into a sequence of mood states. This provides insight about the mood-shifting patterns of the patients and the proportion of time spent in each state. While the Markov chain perspective centers the analysis on the mood states, certain aspects of the underlying mood episodes themselves are lost in this approach. These include episode duration, severity, frequency, and changes in dynamics occurring over time.

The following approach focuses the analysis on the more detailed characterization of the mood symptoms within each episode. The aim of this approach is to create a model that directly speaks to clinically important treatment targets of the disorder – namely the duration, severity, and frequency of mood episodes. In particular, this approach aims to identify periods when a patient is in an episode and to describe attributes of those episodes. Mood symptom data are described using a flexible function that is specifically parameterized to identify these features and the course of the illness for an individual patient. These parameters are then used to assess treatment effects and evaluate group differences. By characterizing the longitudinal patterns in symptom severities, the complete volume of data is leveraged to detect treatment effects and time-dependent nuances, such as trends in episode duration or changes in episode frequency.

This model-based approach has three key steps. The first step is to parameterize longitudinal symptom severity scale scores as a function of key mood episode characteristics. This *mood function* must take a form that is flexible enough to fit general symptom severity patterns of the disease (such as increases and decreases in severity), while also accommodating a range of specific episode manifestations, such as severe acute episodes and mild chronic episodes. Additionally, the function must maintain clinical relevance and interpretability. The second step is to fit the function to the data and obtain parameter estimates describing episode characteristics specific to a patient. The method proposed here uses a stochastic technique known as particle swarm optimization (PSO)[76], a metaheuristic algorithm based on using swarm intelligence to inform the search and discovery of a globally optimal solution. The third step involves using the individual mood function parameter estimates to characterize features present across a group of patients and to test for differences in those features between groups of patients. The method proposed here uses a series of generalized linear models to detect pre-post differences between treatment groups.

In the next section, I present the theoretical foundations of this approach, specifically mood function specification and the PSO algorithm. This is followed by the application of this approach to the three-arm randomized trial data, with adaptations to address study-specific challenges. The chapter concludes with interpretations of the results and important precautions.

5.2 Proposed Approach

The particle swarm approach centers around the estimation of mood duration, severity, and frequency over time by modeling the longitudinal trajectory of mood symptom severity scores as a time-dependent function. This function – the *mood function* – must appropriately describe the salient mood trends and features of bipolar disorder, yet be flexible enough to capture patient-specific variances. Described next is the construction of the mood function and how it is specifically parameterized to capture mood episode features. This is followed by details of the PSO algorithm, including its components and required specifications.

5.2.1 Mood Function Specification

The key to parameterizing the mood symptom severity scores is to focus on the structural form of an individual episode. The basic characteristics of interest for a particular episode are its severity (statistically amplitude) and duration. We assume that an episode k, has a functional form that is uniquely described by its duration parameter σ_k and severity parameter α_k, where σ_k describes the episode duration in units of time and α_k describes the severity of an episode as measured in the corresponding mood scale units. Because the data are longitudinal and episodes occur at specific moments in time, the episode's location must also be specified. In the following formulation, an episode's location is parameterized as μ_k. The episode's influence on a patient's mood scale score at time t is therefore captured by the episode function $e_k(\sigma_k, \alpha_k, \mu_k, t)$. In the three-arm randomized trial, mood symptom severities for depressive and manic episodes are measured by the HDRS and YMRS, respectively, while time is measured in days. The set of parameters describing episode k is abbreviated as ψ_k. The duration, severity, and location parameters describing a set of episodes is denoted as σ, α, and μ, respectively, and is abbreviated as ψ.

In addition to the episodes themselves, symptom severity scores are influenced by a patient's baseline mood severity, $b(t)$ during non-episode periods (i.e., periods of euthymia) over time. Additionally, symptom ratings are subject to random noise, such as typical mood fluctuations due to local events and measurement error. It is assumed that this noise fluctuates around zero and is parameterized as $\epsilon\left(\tau(t)\right)$ where $\tau(t)$ measures the variance of the noise over time. For now, it is assumed for simplicity that the noise variation is constant regardless of whether a patient is in an episode or non-episode period, though this could be modified to allow for greater noise during mood episodes since possible scores encompass a larger range. These components can be combined as follows to form the following mood function:

$$m(\sigma, \alpha, \mu, t) = b(t) + \sum_k e_k(\sigma_k, \alpha_k, \mu_k, t) + \epsilon\left(\tau(t)\right) \tag{5.1}$$

In (5.1), k indexes episodes and α_k, σ_k, and μ_k are the shape and location parameters of the corresponding episode. This equation captures the basic mood severity influences in bipolar disorder. During a period of euthymia, a patient registers an inherent mood level captured by $b(t)$. When a patient is in an episode, mood symptom severity levels are impacted at time t by amounts described by $e_k(\sigma_k, \alpha_k, \mu_k, t)$. Additionally, patients are subject to random fluctuations and measurement error reflected by $\epsilon\left(\tau(t)\right)$.

Although this formulation of the mood function is rather general, it is not without its implicit assumptions. First, it characterizes the effects of each component as additive, with an episode's effect upon a patient's mood symptom severity simply added to the baseline mood severity measure. Similarly, it suggests that an episode's effect is in addition to remnant effects of adjacent episodes. For example, in theory a patient's first episode may begin on day 0 and end on day 15, while a second episode may begin on day 10 and end on day 20. This means that there is an overlap of these two episodes from day 10 to day 15. During this time interval, the effects of both episodes are added on top of each other and the baseline mood severity. This allows for the flexibility in potentially capturing partial episode remission and relapse, as well as nuances of within-episode fluctuation. However, in settings where episodes are indeed distinct, restrictions to the mood function are necessary. Second, this formulation assumes that episodes are uniquely identified by episode location μ_k. That is, the episode locations of any two episodes are assumed to be non-identical. This means that at most only one episode can be located at a specified time, although due to additivity the effects of multiple episodes may be at play at a single time point. Lastly, it assumes the functional form $b(t)$ is distinct relative to the functional form of the episode function $e_k(\sigma_k, \alpha_k, \mu_k, t)$. This maintains model identifiability by keeping episode effects distinct from baseline mood severity patterns. Depending on the research setting and longitudinal mood features of interest, restrictions may need to be imposed to create an appropriate mood function.

5.2.2 Particle Swarm Optimization

Particle swarm optimization (PSO) is a stochastic technique that aims to identify the global maximum or minimum of a function. It is generally applied in cases where the function is mathematically complex or contains multiple local optima. Intricate functions such as these pose problems for traditional deterministic approaches which rely on the mathematical form of the function to enable minimum/maximum identification through differentiation techniques and which may have difficulty distinguishing between local optima and global optima. While brute force techniques can be employed to search the entire parameter space, they can be inefficient, especially in large multidimensional spaces. Rather than systematically searching for the solution, stochastic optimization techniques iteratively search the parameter space by a random process.

Particle swarm optimization is a member of the class of stochastic optimization techniques known as neighborhood search algorithms, or local search algorithms[77]. Neighborhood search algorithms can be generally described as follows. Suppose we have a complex function $f(\theta)$ where θ is the characterizing parameter, and we are interested in finding the optimal estimate of that paragon, $\hat{\theta}$. The optimality of a candidate solution $\tilde{\theta_i}$ is described by an objective function $g(\theta)$ such that $g(\hat{\theta}) = \min\left[g(\theta)\right]$. Neighborhood search algorithms aim to find $\hat{\theta}$ by iteratively searching the parameter space of θ. Like many iterative search approaches, the procedure begins with an initial solution, $\tilde{\theta_0}$. Next, a candidate solution, $\tilde{\theta_0}^*$, is randomly drawn from the region of the parameter space neighboring $\tilde{\theta_0}$. The optimality of the current solution, $g(\tilde{\theta_0})$, and the candidate solution, $g(\tilde{\theta_0}^*)$, are compared to determine which provides a better result. Based on this evaluation process, a new candidate solution $\tilde{\theta_1}$ is generated. This process is repeated and candidate solutions are updated until a set of criteria are met. Because the search is stochastic, it cannot be guaranteed that the process will indeed identify the true global optimum. However, such algorithms provide a means of efficiently searching the parameter space and identifying near-global (and often sufficient) solutions in mathematically complex settings.

The PSO algorithm takes the basic iterative principles of neighborhood search algorithms and adds two refinements: (1) search space dynamics and (2) swarm-based intelligence. The foundational component in PSO is the *particle.* A particle represents a candidate set of parameter values, $\tilde{\theta}_i$, for the function being optimized. In this context, θ corresponds to the entire parameter vector of the mood function, which includes the baseline severity and the durations, amplitudes, and locations of *all* episodes. Much like other neighborhood search algorithms, this particle iteratively searches the solution space and seeks the global optimum. A particle is thought of as our entity that moves about the search space with a speed and direction referred to as its *velocity.* PSO differs from the basic neighborhood search algorithm in that it initiates a set of particles – referred to as the *particle swarm* – which all search the space simultaneously. This carries the advantage of creating a profile of solutions across the search space in a single iteration. This information – the *swarm-based intelligence* – is leveraged to determine more promising regions of the search space and guide the dynamics of each particle to these areas. This is accomplished by keeping a record of two key types of values: (1) a particle's best solution across all previous iterations, and (2) the global best solution seen so far by any particle. These values serve as attractors and influence the particle velocities, directing the swarm to its next iteration and eventually to the global optimum. Details behind these mechanisms and its required specifications are described next.

5.2.2.1 Search Space

Before the PSO algorithm can be implemented, initial specifications relating to the search space are in order. First, the objective function to be optimized, $g(\theta)$, must be specified and θ must be restricted to a fixed dimensional space. Second, boundaries of the search space, θ_{bound}, must be specified to prevent swarm velocities from driving particles to regions with nonsensical candidate solutions. These boundaries are usually determined by phenomena underlying the data generating mechanisms and serve as a means of introducing prior knowledge into the search specifications. The last required specification is the swarm size s. Complex spaces may require a larger swarm size to adequately cover the area and identify nuanced optimal regions. However, this comes at the expense of additional computational power and total algorithm run time. Therefore, adequate solution accuracy and practical limitations must be appropriately balanced.

5.2.2.2 Particle Velocity and Position Update

The next set of required specifications relate to the procedure for updating particle velocities and candidate solutions. Letting x_i^j represent particle i's position x_i at iteration j, the location of the particle at the next iteration is determined by:

$$x_i^{j+1} = x_i^j + v_i^{j+1} \tag{5.2}$$

Here, v_i^{j+1} is referred to as the update velocity and guides where the particle will look next. The update velocity consists of three components: (1) a particle's current velocity, (2) its personal influence, and (3) the swarm's influence. The personal influence represents the impact that the particle's best solution has on the particle. Similarly, the swarm's influence represents the effect of the global best solution. Velocity updating is summarized by the following equation:

$$v_i^{j+1} = wv_i^j + c_1 r_1^j (B_i^j - x_i^j) + c_2 r_2^j (B^j - x_i^j) \tag{5.3}$$

In equation (5.3), the current velocity of particle i at iteration t is represented by v_i^j and its position in the search space is x_i^j. The particle's best and global best solution at iteration t are represented by B_i^j and B^j, respectively.

The personal influence – described by $c_1 r_1^j (B_i^j - x_i^j)$ – is a function of a local exploration constant c_1, a local stochastic element r_1^j, and the distance between the particle's best solution and its current position $(B_i^j - x_i^j)$. The local exploration constant is a positive number that controls the degree of influence of the particle's best solution. It is sometimes referred to as an acceleration coefficient because it determines how quickly a particle accelerates toward the best solution. The local stochastic element is a random variable and adds an element of randomness to the search. To illustrate the impact of these values, suppose $c_1 = 1$ and $r_1^j = 1$. This would result in a personal influence of $(B_i^j - x_i^j)$. Since this describes the distance between the particle and it's best solution, this would direct the particle to head directly toward B_i^j. If the product $c_1 r_1^j = 2$, the personal influence would be $2(B_i^j - x_i^j)$ and the particle would overshoot it's personal best solution by two-fold. Similarly, a product of $c_1 r_1^j = 0.5$ would cause the particle to undershoot it by 50%. In the original formulation, the local stochastic element was drawn from a uniform distribution ranging from 0 to 1 and the local exploration constant was set to 2. Taken together, these velocity settings suggest that the particle will overshoot the particle's best solution about half the time on average.

The swarm's influence – described by $c_2 r_2^j (B^j - x_i^j)$ – is a function of the global exploration constant c_2, a global stochastic element r_2^j, and the distance between the global best solution and the particle's current position $(B^j - x_i^j)$. It has a structure analogous to the personal influence and carries a similar interpretation: c_2 controls the degree of influence of the global best solution and r_2^j adds a random element to the search.

The parameter w is referred to as the inertia weight and was proposed as a modification to the original PSO algorithm[78]. It is a constant that controls the influence of a particle's current velocity upon its updated velocity. A high inertia weight suggests that a particle's current velocity carries high influence, so the particle is less swayed by the personal and swarm influence. A low inertia weight suggests that a particle is eagerly drawn toward the personal and/or global best solutions. As an example, suppose $w = 0$. This would suggest that the particle is completely directed by the personal influence and the swarm's influence. Values greater than zero diminishes the impact of the personal and swarm's influence, while very high values may negate it completely. The original research on inertia weight suggested values in the range of 0.9 to 1.2 based on simulation studies.

Standard versions of the PSO algorithm and its velocity parameter specifications have been proposed[79]. The most recent is the Standardized Particle Swarm Optimization

2011 (SPSO-2011) which is the version used in this analysis. Its parameters are as follows: $c_1 = c_2 = 5 + log(2) = 1.193$, $w = \frac{1}{2log(2)} = .721$, and $s = 40$. While these are presented as standard values reflective of theoretical research available at the time of its development, new findings may recommend features and adjustments that lead to an updated standard version.

Note that a particle's position will always be updated at each iteration. This differs from other local search algorithms which incorporate acception or rejection of candidate estimates based their optimality measures. In the PSO algorithm, a particle will always be updated even if the resulting new position is less optimal. However, because the particle's best and global best solutions are incorporated in the update process, even though particles are allowed to roam to less optimal regions they have an inclination to return to previously-encountered optimal solutions unless a more optimal solution is found. Additionally, if a particle's best solution, global solution, and current position are identical, then the update velocity will be very low and the particle will have a tendency to stay at that position. As the process evolves, particles will eventually swarm at the same position in the search space.

In summary, the dynamics of the PSO algorithm are controlled by the local and global exploration constants which are used to tune the balance between intensive local searching versus favored global searching. Low values translate to a weak influence, allowing the particle to venture far from the current best solutions. High values translate to a strong influence, restricting the particle's exploration range to nearby regions. The overall speed of the algorithm is tuned by the inertia weight. Lower inertia weights translate to easier acceleration toward the best solutions, although this may result in premature convergence. Higher inertia weights translate to slower convergence, but allow for a more thorough search before being influenced by the best solutions. Particles explore the search space by these mechanisms and are eventually drawn to an optimum. Having many widely dispersed particles at the start helps ensure adequate coverage.

5.2.2.3 Stopping Criteria

The last set of required specifications relates to stopping the search. The easiest criteria to implement are based on the number of iterations. The simplest would be to restrict the total number of iterations for the search. This assumes that the specified number of iterations is sufficient enough to identify the global optimum. Another option is to specify the maximum number of stagnate iterations. This assumes that the search has arrived at least at a local optimum, with only less optimal solutions in its neighboring regions and therefore stagnates. Depending on the research setting, convergence may be determined by an absolute tolerance level. This stops the search after the calculated objective function value has reached an acceptable level of optimality specified by a threshold. Choices for stopping criteria depend on the objective function used, the interpretation of its calculated values, and what is deemed to be an acceptable solution. This allows the algorithm to be applied across a variety of research settings, but the investigator is faced with defining and determining a solution's sufficiency.

5.3 Application

In this section, I outline specifications and modifications made to apply these methods to treatment trials for bipolar disorder. Described first is the functional form of the mood function, its parameterization, and its associated definitions and interpretations. Next I describe details of a grid-based approach I employ to detect time intervals that are likely to contain mood episodes. This creates search space restrictions and boosts the efficiency of the PSO algorithm. Next I outline parameter specifications of the PSO algorithm. Finally, I present statistical models used to detect treatment differences based on the results of the PSO algorithm.

5.3.1 Adaptations to the Mood Function

For the implementation of this approach, I make some simplifying restrictions to the mood function described by equation (5.1). First, I restrict the baseline mood severity function $b(t)$ to be constant over time. This explicitly assumes that there are no fluctuations in the baseline mood severity, and also implicitly assumes that the baseline mood severity remains unchanged after treatment initiation. This suggests that treatment effects are only manifest in the episode-specific parameters σ and α. Second, I restrict the random noise variance $\tau(t)$ to also be constant over time. Next, I specify the functional form of the episode function as a Gaussian curve such that for a given episode:

$$e_k(\sigma_k, \alpha_k, \mu_k, t) = \alpha_k \exp\left(-\frac{(t-\mu_k)^2}{2\sigma_k^2}\right) \tag{5.4}$$

Here, α_k describes the maximum the mood symptom severity score for episode k and can be interpreted as the amplitude of the episode, while μ_k, the location parameter, corresponds to the time when the influence of episode k is at its greatest. While other functional forms can be used, Gaussian curves have mathematical properties well suited to describe increasing and decreasing processes. It exhibits a rise and fall pattern – a qualitative feature of mood episodes – but does so in a smooth, exponential fashion. Modeling episodes as Gaussian curves assumes that they are symmetric and unimodal. While there is no marked beginning and end to an episode under this parameterization scheme, the degree of an episode's duration is captured by σ_k. This is similar to normal distributions and how standard deviations are use to describe an interval surrounding the curve's peak. With these adjustments, the mood function described in equation (5.1) reduces to:

$$m(\sigma_k, \alpha_k, \mu_k, t) = b + \sum_k \alpha_k \exp\left(-\frac{(t-\mu_k)^2}{2\sigma_k^2}\right) + \epsilon(\tau) \tag{5.5}$$

For purposes of interpretation, duration is measured by the size of the interval $(\mu_k \pm 2\sigma_k)$ and is referred to as the *duration window*. Similarly, σ can also be used to describe the peak region of an episode. In this analysis, I define this interval as $(\mu_k \pm .75\sigma_k)$ and is referred to as the *peak window*. The choice of .75 was loosely based on the idea that it approximately covers the middle three-eighths of the duration window and reflects a reasonably restrictive region when implementing penalties (described later in Section 5.3.3). Note that under this formulation, the number of parameters increase as k increases and may vary across patients. Due to this varying dimensional space, the PSO algorithm with be used to fit each patient separately.

5.3.2 Episode Region Detection

One challenge in applying the PSO algorithm to episode-based data is the specification of the search space. According to the mood function described in equation (5.1), the number of episodes determines the size of the dimensional space. This requires that the number of episodes be determined before applying the PSO algorithm. Additionally, the episodic structure of the mood function implies a grouping to the parameter estimates. That is, episode k is described by the parameters σ_k, α_k, and μ_k. To properly characterize the mood function, these parameters must be considered as a set corresponding to episode k and must be distinguished from the parameters σ_{k+1}, α_{k+1}, and μ_{k+1}, which form the parameter set of episode $k + 1$. Permuting these parameters without respecting the grouping will result in a completely different mood function. Furthermore, the mood function implies that a curve may be described by multiple parameter orderings. This presents issues of statistical identifiability. For example, suppose ψ_1, ψ_2, and ψ_3 correspond to a patient's episode parameters (i.e., σ_k, α_k, and μ_k) of episodes 1, 2, and 3, respectively. Based on the formulation of the mood function, a parameter vector of (ψ_1, ψ_2, ψ_3) corresponds to a mood curve that is identical to (ψ_3, ψ_2, ψ_1) or (ψ_2, ψ_1, ψ_3). This label-switching problem presents ambiguities in parameter estimation and requires a system that keeps track of which parameters correspond to specific episodes.

To address these challenges, I used a grid-based approach to identify regions of probable episode locations. This is done by proposing a set of episodes spanning a range of durations and amplitudes and fitting them across multiple time points that span the data.

The fit of each episode is evaluated by a fitness function and the collection of these fitness values are plotted across time. This takes advantage of the one-dimensional nature of time and how episodes are distinct across time. If the set of episodes do not fit well across a time interval, then that region will exhibit poor fitness values and will be interpreted as a non-episode region. However, time intervals with well-fitting episodes will exhibit favorable fitness values and will be interpreted as probable episode regions. By identifying these regions, the time dimension is divided into mutually exclusive intervals with each one corresponding to a unique episode. This allows episode parameters to be grouped according to these regions and presents a means of uniquely identifying episodes. Additionally, it reduces the parameter space and restricts the PSO search to episode-probable locations.

For the proposed set of episodes, I selected a total of 100 different episode shapes by varying α and σ across ten different values for each parameter, resulting in 100 possible combinations. For α, the episode set ranged from 1 to 35 for HDRS data and 1 to 25 for YMRS data and encompassed the entire range observed in the trial data. For σ, the episode set ranged from 1.75 to 21 for both the HDRS and YMRS data. Recall that an episode's duration is equivalent to 4σ. Therefore, this range corresponds to episodes ranging from 7 to 84 days (i.e., 1 to 12 weeks). These values were selected based on the intended sampling rate of the study (1 week) and the minimum frequency required for rapid cycling (at least 4 episodes per year). The selected values for both α and σ were equally spaced along their corresponding intervals. For the episode location parameter μ, points spanned the entire range of the data at intervals of 1.75 days, resulting in 4 time points per week.

The fitness function used to evaluate episode fit consists of three components: (1) data fit with no episode present, (2) data fit with the episode present, and (3) fit restrictions. For episode k, fit is evaluated according to data within the interval $\mu_k \pm 2\sigma_k$, the peak window. The data points spanning this region are denoted by the vector Y_k and consists of time points $(t_{1k}, t_{2k}, ..., t_{dk})$ and its mood scale scores $(y_{1k}, y_{2k}, ..., y_{dk})$. Denoting the average mood scale score within the interval as $y_{.k}$, we take a sum of squares approach and evaluate as the data fit in the absence of an episode:

$$f_0(Y_k) = \sum_{i=1}^{d} (y_{ik} - y_{.k})^2 \tag{5.6}$$

This value represents the total variance within the interval. To evaluate the fit with the episode present, a similar approach is used. Using the mood function specified in equation (5.5), this fit is specified by:

$$f_1(Y_k, \sigma_k, \alpha_k, \mu_k) = \sum_{i=1}^{d} (y_{ik} - m(\sigma_k, \alpha_k, \mu_k, t_{ik}))^2 \tag{5.7}$$

This value is the episode's sum of squares error and represents the residual variance that is unexplained by the episode. Note that the function $m(\sigma_k, \alpha_k, \mu_k, t_{ik})$ requires the baseline severity b to be specified. The parameter b is chosen such that the function $f_1(Y_k, \sigma_k, \alpha_k, \mu_k)$ is minimized. Taking the results of equations (5.6) and (5.7), the episode fitness value is calculated as:

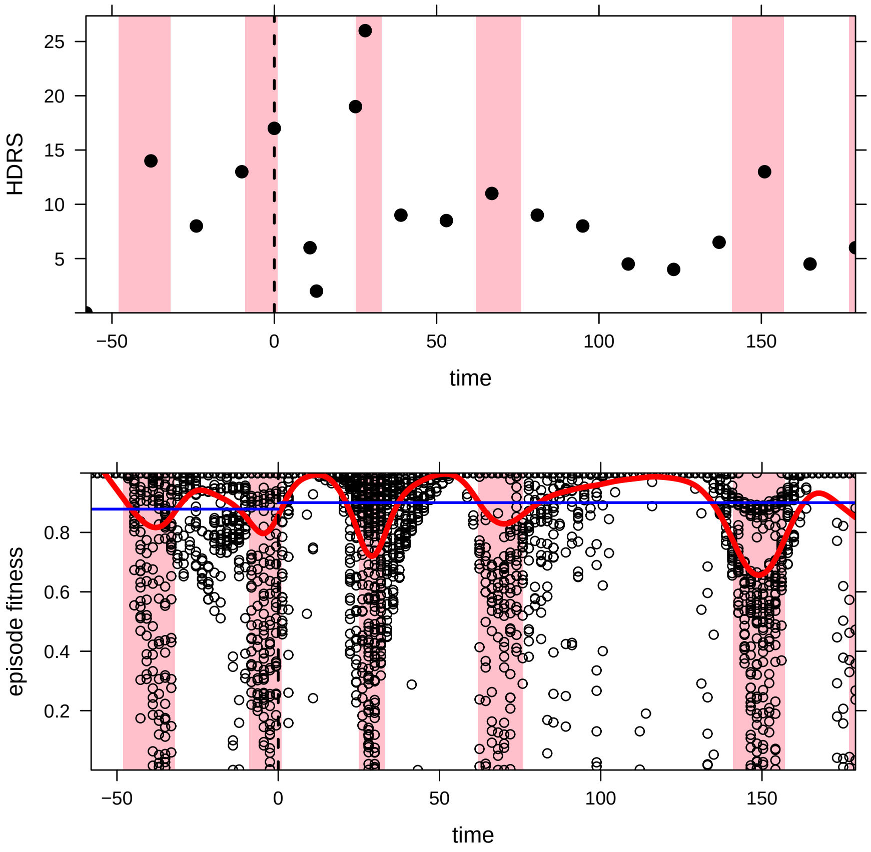
$$F_k(Y_k, \sigma_k, \alpha_k, \mu_k) = \frac{f_1(Y_k, \sigma_k, \alpha_k, \mu_k)}{f_0(Y_k)} \tag{5.8}$$

This value is interpreted as the proportion of variance unexplained by episode k and is expected to have a range between 0 (an episode that fits the data perfectly) and 1 (an episode that fits equally as well as a constant value). This suggests that better-fitting episodes correspond to lower values, while higher values equate to episodes that do not fit the data well. However, for very poorly-fitting episodes, it is possible to obtain a value greater than 1. This indicates an interval whose data variance is better explained by a constant value rather than the candidate episode's bell-shaped curve. For these instances, the maximum value of 1 is imputed. Additionally, the selected baseline severity parameter b may lie outside the range of possible values of the mood severity scale. Moreover, it is also possible that Y_k may not contain any data within the episode's peak window (defined as $\mu_k \pm .75\sigma_k$), suggesting that the calculated fit is based on points along the episode's edges only and is not reflective of the episode's bell-shaped curve. To address these issues, episodes meeting any of the latter two criteria are removed from the episode detection analysis.

After episode fitness values are calculated across the entire spectra of the data, they are plotted across time. Data are fit by a cubic smoothing spline with the degree of smoothing determined by the leave-one-out cross-validation method[80]. While other smoothing techniques can be employed in this approach, the cubic smoothing splines method is chosen because its formulation consists of piecewise cubic curves with a smooth second derivative. This facilitates the identification of curve minima and inflection points. These form the basis of episode region detection. Locations exhibiting favorable fitness values are identified by the curve's minima and episode-probable intervals are marked by the inflection points containing the minima. It is assumed that each region contains a single episode. To filter out spurious regions that may result from random noise, the mean episode fitness value across the entire curve is calculated as a threshold and the identified minima must be less than this value. This suggests that the region is exhibiting episode fitness values that are better than average. In cases where episode profiles may change across time, separate thresholds may be set. Similarly, values other than the average can be selected, allowing the threshold to be properly tuned to the research setting. In applying this method to the three-arm randomized trial data, thresholds are calculated separately according to the pre-randomization and post treatment-stabilization periods. An example of this episode detection process is shown in Figure 5.1.

Identifying these regions provides three main benefits. First, it identifies the number of episodes to be fit, fixing the parameter space of the search and allowing the PSO algorithm to be applied. Second, it provides a labeling mechanism to identify which parameters correspond to a given episode as an episode can now be identified according to the interval that encompasses location parameter μ_k. This addresses the label-switching ambiguity discussed earlier. Third, it restricts the dimensional space and ensures that the search efforts are focused on promising regions. Discussed next are the details of applying the PSO algorithm to the three-arm randomized trial data and how the episode detection results are used.

Figure 5.1: Episode Region Detection Plot



The first panel is a plot of a patient's HDRS data versus time. The second panel reflects the results of the episode region detection. Episode fitness values are plotted across time and data are smoothed by cubic smoothing splines. Lower values correspond to better fitness values. The horizontal lines are the thresholds of each study period to determine favorable fitness values and the highlighted intervals are the resulting regions. The dashed line at zero indicates the point of randomization.

5.3.3 Objective Function

Before the PSO algorithm can be applied, the objective function must be defined. The algorithm is flexible enough to accommodate a variety of objective functions and allows the investigator to evaluate optimality over multiple components. However, because optimality is calculated for multiple particles at each iteration, complex objective functions often translate to longer run times. Therefore, it is necessary to select an objective function that is simple enough to minimize computational complexity, yet is sophisticated enough to incorporate key features indicative of optimality.

The challenge in using the PSO approach for the three-arm randomized trial centers around data sparsity and the associated difficulty in distinguishing a true signal from background noise. The objective function I developed to fit the data addresses these concerns by incorporating penalties. The function is developed as follows. Following the approach used in episode region detection and equation (5.7), overall fit is measured by the sum of squared differences between the data values and the candidate solution's predicted values. The fit value is interpreted as the unexplained variance with lower values indicating a better overall fit. To address issues introduced by data sparsity, each fitted episode is inspected to check whether the following criteria are met. First, the episode's amplitude α_k is compared to the noise level resulting from the fit. This noise level, described by τ in equation (5.5), is estimated by the root mean squared error of the fit. Second, the number of data points encompassed by an episode's estimated peak window region $\mu_k \pm .75\sigma_k$ is counted. Episodes with amplitudes less than twice the estimated noise level or containing no points within their peak window regions are flagged as poorly-fitting episodes. The overall fit calculated earlier is inflated by the proportion of episodes not meeting the criteria. For example, if a candidate solution composed of 5 episodes results in an overall fit of 100, yet has 1 episode failing to meet the criteria, the overall fit value is inflated by 20% and results in a value of 120. At best, all episodes meet the criteria and no inflation occurs. At worst, all episodes violate the criteria and the initial overall fit value is doubled. This adjusted overall fit is the final objective function value.

5.3.4 Particle Swarm Optimization Parameters

To fit the randomized trial data using the PSO algorithm, the parameter space boundaries, velocity parameter values, and stopping criteria must be defined. The algorithm is applied to each patient individually as the number of episodes – and consequently the parameter space – differ across patients. The number of episodes to fit and the boundaries of μ are determined by the episode region detection results. The boundaries for α are dependent on the range of the patient's mood scale scores. The minimum amplitude is set to 1 and the maximum amplitude is the difference between the highest and lowest scale scores observed. Duration estimates are allowed to range from 1 to 12 weeks and correspond to σ values of 1.75 and 21 days, respectively. Similar to the rationale behind the episode set in the region detection method, these values were selected based on the intended sampling rate of the study (1 week) and the minimum frequency required for rapid cycling (at least 4 episodes per year). The baseline severity b is allowed to range from 0 to the highest scale score observed. For the PSO velocity parameters, the values recommended by SPSO-2011 are used: $c_1 = c_2 = 1.193$, $w = .721$, and $s = 40$. The algorithm is stopped after 500 iterations or after no improvement in the optimal solution is observed for 100 iterations.

5.3.5 Filtering

Although the objective function penalizes episodes that are unreliable due to data sparsity issues, there is no guarantee that the resulting fit is free of such episodes. This may arise when a detected episode region is deemed favorable when considered as a single interval, but is poor when considering the fit of all other episodes simultaneously. In its search, the algorithm must fit an episode in this region and the optimal fit may specify a penalized episode. To safeguard against this, a filter is imposed on the resulting PSO fit. If the resulting fit contains penalized episodes, their corresponding episode regions are removed and the data are refitted with this updated search space. This provides a mechanism of refining the episode regions in instances where the episode region detection method resulted in false-positives.

5.3.6 Modeling Episode Characteristics

To test for differences in episode characteristics across treatment groups over time, a linear mixed effects model is fit to the subject level episode parameters resulting from the PSO algorithm. This model is used to fit episode duration and amplitude data because multiple episodes per subject occur within a single study time period, necessitating a within-subject repeated measures framework. Data are assumed to be normally distributed conditional on treatment assignment, study time period, and the treatment-time interaction effect. Data within a patient are assumed to be correlated and patient effects are modeled as having a random intercept. Data between patients are assumed to be independent. The model is specified in matrix notation as follows:

$$y_i = X_i\beta + Z_ib_i + \epsilon_i \tag{5.9}$$

The data for patient i is captured by the set of episode characteristics y_i, the fixed-effect design matrix X_i, and the random-effect design matrix Z_i. The fixed-effect parameter vector, β, consists of an intercept term, treatment effects, and treatment by time interaction effects. The random-effect parameter b_i is patient-specific and consists of an intercept term. The primary parameter of interest is the treatment by time interaction effect which measures differential changes associated with treatment initiation.

Episode frequency is modeled using negative binomial regression. This model is used to model count variables exhibiting a high degree of variance. Episode frequency is modeled as follows:

$$log(y_{i1}) = log(y_{i0}) - log(\phi_0) + log(\phi_1) + \beta_0 + \beta_1 x_i \tag{5.10}$$

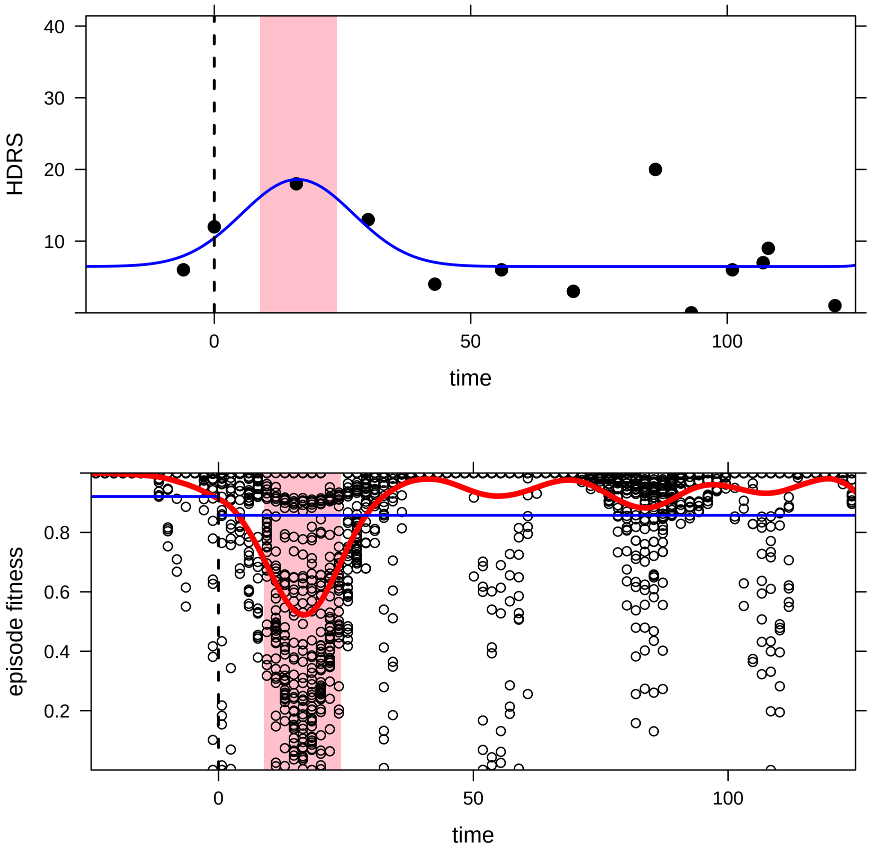
In this model, y_{i0} and y_{i1} correspond to the number of episodes for patient i during the pre-randomization and post treatment-stabilization period, respectively. Treatment group assignment is represented by x_i. The model parameters β_0 and β_1 correspond to the intercept and treatment effects, respectively. When exponentiated, these parameters are interpreted as rate ratios. The primary parameter of interest is β_1 which captures the association between treatment assignment and episode frequency changes. Because of varying follow-up times, offsets for the pre-randomization and post treatment-stabilization periods are included, represented by $log(\phi_0)$ and $log(\phi_1)$ respectively.

5.4 Results

The quality of the PSO algorithm's fit to the data is summarized by the root mean squared error (RMSE). RMSE is calculated for each patient and mood scale. These values are averaged across patients by treatment group. Summary statistics of RMSE are displayed in Table 5.1. The results suggest a fair degree of unexplained variance in the HDRS data. The greatest discrepancies are observed in the T_4 group with RMSE = 3.03. Fits for the other treatment groups are more favorable with RMSE = 1.23 and 1.90 for the PL and T_3 groups, respectively. Errors are less pronounced in the YMRS data with RMSE values of 0.82, 0.92, and 0.78 for the PL, T_3, and T_4 groups, respectively, and exhibit relatively less variance. Visual inspection of the patients with higher RMSE values indicate that poor fit may be due to undetected highly-acute episodes and an attenuated baseline severity. An example of this is shown in Figure 5.2. This shows the possibility of over-smoothing the episode fitness values or overly-restrictive thresholds, leading to the undetected episodes. This suggest that the failed detection of acute severe episodes may bias both amplitude and duration estimates. Therefore, this method requires finer tuning to detect sudden shifts in sparse data settings.

The average episode duration is calculated for each patient by mood scale and study

Figure 5.2: PSO Fit Example



The first panel is a plot of a patient's HDRS data versus time with the predicted fit by the PSO algorithm in blue. The second panel reflects the results of the episode region detection method. Although there is an elevated HDRS measure at approximately time = 80 days, an episode region is not specified for this interval. This suggests that the detection method may be not be robust to acute episodes due to possible over-smoothing when data are sparse.

Table 5.1: PSO Fit by Group

	HDRS		YMRS	
group	mean	SD	mean	SD
PL	1.23	0.88	0.82	0.38
T_3	1.90	1.33	0.92	0.36
T_4	3.03	1.23	0.78	0.37

The RMSE comparing the PSO fit and the observed values is calculated for each patient. Listed here are the mean and SD across patients for each treatment group for HDRS and YMRS data.

Table 5.2: Episode Duration in Days by Group and Study Period

	HDRS				YMRS			
	pre		post		pre		post	
group	mean	SD	mean	SD	mean	SD	mean	SD
PL	38.17	19.93	42.26	20.70	31.54	14.63	27.73	11.67
T_3	31.10	15.48	37.96	27.22	30.79	20.88	24.41	10.24
T_4	42.60	25.89	25.31	13.13	33.40	21.87	30.69	20.36

The average episode duration in days is calculated for each patient by study period. Listed here are the mean and standard deviation (SD) across patients for each treatment group by mood scale.

period. Summary statistics are calculated for these values across patients by treatment group and are presented in Table 5.2. Recall that an episode's duration is interpreted as 4σ. In general, episodes estimated along the depression axis (HDRS) are longer than those estimated along the mania axis (YMRS), although estimates on both scales are highly variable. These summary statistics suggest that the T_4 group may be exhibiting a decrease in average depressive episode duration (42.6 days versus 25.3 days). Changes in average mania episode duration are comparable across treatment groups, although the largest difference is observed in the T_3 group (30.8 versus 24.4 days).

Table 5.3: Episode Amplitude by Group and Study Period

	HDRS				YMRS			
	pre		post		pre		post	
group	mean	SD	mean	SD	mean	SD	mean	SD
PL	15.00	6.40	14.23	4.09	5.86	2.85	4.31	2.21
T_3	14.19	4.16	10.91	6.57	8.43	3.02	7.18	3.11
T_4	17.59	4.20	14.32	6.12	6.91	5.75	7.02	4.58

The average episode amplitude is calculated for each patient by study period. Listed here are the mean and standard deviation (SD) across patients for each treatment group by mood scale.

Like episode duration, the average episode amplitude is calculated for each patient by mood scale and study period, and summary statistics are calculated across patients by treatment group. These statistics are shown in Table 5.3. Results indicate that the typical peak severity for depressive episodes is 10 to 20 units. For manic episodes, the typical peak severities were estimated to be around 4 to 9 units. Within-group differences suggest that the T_3 and T_4 groups may be experiencing a decrease in episode amplitude of depression symptoms, although the observed differences are not large relative to the observed standard deviations. The estimated mania episode amplitudes exhibit substantial variance. While the estimated means may suggest minor changes within the PL and T_3 groups, high variability casts doubt on these effects.

Due to variable follow-up times, episode frequencies are annualized to describe the 12-month episode rate. Summary statistics of these observed rates by treatment group, mood scale, and study period are presented in Table 5.4. Results indicate a high degree of between-subject variability, but the average episode rates are roughly comparable across treatment groups by study period. The estimates do suggest some possible within group differences: decreases in the average episode rate are observed within PL group along the HDRS (11.1 versus 7.1 per year) and within the T_3 group along the YMRS (8.9 versus 3.3 per year), but much like the other episode characteristics, high variances mean these differences

Table 5.4: 12-Month Episode Rate by Group and Study Period

	HDRS				YMRS			
	pre		post		pre		post	
group	mean	SD	mean	SD	mean	SD	mean	SD
PL	11.09	5.33	7.06	3.89	6.89	3.33	5.61	3.79
T_3	8.41	3.07	5.92	4.69	8.90	4.01	3.29	2.79
T_4	6.69	3.93	5.09	4.29	5.57	3.69	6.86	5.98

The annualized episode rate is calculated for each patient by study period. Listed here are the mean and SD across patients for each treatment group by mood scale.

are not statistically significant.

Overall, summary statistics of episode duration, amplitude, and rate reflect a high degree of variability. Additionally, these statistics are based on values that have been averaged within a patient and glosses over patient-level variability. Although the results may be suggestive of some trends, large variances suggest more refined approaches to detect differences while addressing data uncertainties, such as the models presented next.

Results of the linear mixed-effect models of episode duration are presented in Table 5.5. For the HDRS data, the model estimates an average episode duration of 36.3 days for the PL group for the pre-randomization period. A lower duration is estimated for the T_3 group (31.1 days), and a higher duration for the T_4 group (43.3 days), although these differences are not statistically different than the pre-randomization period duration of the PL group (p=.5610 and .4182, respectively). For the PL group, the model estimates an increase in duration of 6.6 days, although this pre-post treatment result is not statistically significant (p=.3500). The model estimates a smaller increase in duration for the T_3 group (2.6 days), although not statistically different from the PL group. A statistically significant difference is observed for the T_4 group with an estimated decrease in duration of 17.6 days (p=.0172). This suggests that T_4 is associated with a significant decrease in depressive episode duration

Table 5.5: Parameter Estimates: Episode Duration Model

	HDRS			YMRS		
effect	β	SE	p-value	β	SE	p-value
Intercept	36.29	6.13	<.0001	32.10	6.35	<.0001
T_3	-5.15	8.62	.5610	-0.22	8.38	.9792
T_4	7.00	8.42	.4182	1.70	8.63	.8448
time	6.58	7.19	.3500	-3.55	7.73	.6478
$T_3 \times time$	-3.95	10.29	.6675	-1.22	10.87	.9108
$T_4 \times time$	-24.15	10.08	.0172	3.22	10.23	.7540

The fixed effects parameter estimates of the episode duration linear mixed effects regression models. Separate models are fit to the HDRS and YMRS data. Estimates and standard errors (SE) are in days.

relative to the PL group.

Results of the episode duration model for YMRS are inconclusive. The estimated pre-randomization period episode durations for the T_3 and T_4 groups are 31.9 and 33.8 days, respectively, and are not statistically different from that of the PL group (32.1 days; p=.9792 and .8448, respectively). A decrease in duration is estimated for the PL group (3.6 days), but is not statistically significant. Decreases of 4.8 days and 0.3 days are estimated for the T_3 and T_4 groups, respectively, which also are not significantly different from that of the PL group (p=.9108 and .7540, respectively). Overall, these results suggest that adjunctive T_4 treatment is associated with shorter depressive episode durations as measured by the HDRS, while the impact of T_3 or T_4 on episode durations measured by the YMRS are inconclusive.

Results of the linear mixed-effect models of episode amplitude are presented in Table 5.6. Recall that the amplitude is defined as the increase in mood scale score relative to the estimated baseline score. For the HDRS data, the estimated amplitude for the PL group during the pre-randomization period is 14.7 points on the HDRS scale. The estimated amplitudes for the T_3 and T_4 groups are 14.1 and 16.9 points, respectively, and did not significantly differ from the PL group (p=.8122 and .3910, respectively). A decrease of 1.1 points

Table 5.6: Parameter Estimates: Episode Amplitude Model

	HDRS			YMRS		
effect	β	SE	p-value	β	SE	p-value
$Intercept$	14.73	1.64	<.0001	6.05	1.36	<.0001
T_3	-0.61	2.30	.8122	2.30	1.79	.2093
T_4	2.18	2.24	.3910	0.90	1.85	.6298
$time$	-1.10	1.92	.4566	-1.44	1.70	.4002
$T_3 \times time$	-1.43	2.76	.6879	0.99	2.39	.6788
$T_4 \times time$	-1.55	2.66	.6065	1.41	2.25	.5314

The fixed effects parameter estimates of the episode amplitude linear mixed effects regression models. Separate models are fit to the HDRS and YMRS data. Estimates and standard errors (SE) are in units of the respective mood scales.

is estimated for the PL group, although this result is not statistically significant. Although greater decreases are estimated for the T_3 and T_4 groups (2.5 and 2.7 points, respectively), these effects are also not large enough to claim superiority over the PL group (p=.6879 and .6065, respectively).

Results of the episode amplitude model for YMRS are also inconclusive. The estimated pre-randomization period episode amplitudes for the T_3 and T_4 groups are 8.4 and 7.0 points on the YMRS scale, respectively, and do not differ from the amplitude estimate of 6.1 points for the PL group (p=.2093 and .6298, respectively). The estimated change in amplitude is greatest for the PL group (-1.4 points) followed by the T_3 (.5 points) and T_4 (.01 points) groups, not close to statistically significant effects. Overall, these results suggest that are no treatment effects in peak severities in mania episodes.

Results of the negative binomial regression model for episode rate are presented in Table 5.7. Episode frequencies are mathematically modeled as the logarithm of the episode rate and the effects of the covariates are described as the logarithm of the rate ratio. To facilitate interpretation, the exponentiated parameter estimates are also given in Table 5.7.

Table 5.7: Parameter Estimates: Episode Rate Model

	HDRS				YMRS			
effect	β	SE	e^{β}	p-value	β	SE	e^{β}	p-value
Intercept	-.412	.182	—	.0232	-.366	.193	—	.0580
T_3	.029	.257	1.030	.9090	-.021	.279	.979	.9390
T_4	-.178	.250	.837	.4772	-.414	.260	.661	.1120

The parameter estimates of the episode rate negative binomial regression models. Separate models are fit to the HDRS and YMRS data. The parameter estimate β is exponentiated to reflect the estimated rate ratio comparing the thyroid treatment arms with the placebo arm.

For the HDRS data, the model does not detect any significant post-treatment rate differences in the T_3 or T_4 groups relative to the PL group. The model estimates an episode rate for the T_3 group that is only 3.0% greater than the PL group and is not statistically significant. A larger effect is estimated for the T_4 group with an episode rate that is 16.3% lower than the PL group, but this difference does not achieve statistical significance. For the YMRS data, results suggest a possible effect in the T_4 group. While the model estimates a 2% decrease in episode rate for the T_3 group relative to the PL group (p=.9390), the estimated decrease for the T_4 group is 33.9% (p=.1120). Though this effect does not achieve statistical significance, it is suggestive of a possible effect that may be detected in a larger sample setting.

5.5 Conclusions

The results described above indicate a high degree of variability present in the three-arm randomized trial data. This may be a consequence of the small sample size and data sparsity. Although many estimates are in the expected direction, the large variances preclude statistical significance. Modeling episode characteristics in a repeated measures framework addresses some of this variance and reveals some group differences in the three-arm randomized trial data. The episode duration model of the HDRS data suggests that T_4 is associated with shorter episode durations relative to PL (21.4 less days; p=.0188). Furthermore, the episode rate model of the YMRS data point toward a possible association between T_4 and a decreased episode rate relative to PL (33.9% decrease; p=.1120). However, these findings are contingent upon the appropriateness of the episode region detection and PSO algorithm elements. Visual inspection of the PSO fit suggests that highly acute episodes and fluctuations in baseline severity may lead to undetected episodes and bias the subsequent results. This may be problematic in certain research settings and must be fully addressed by tuning the components of the episode detection method and the PSO algorithm accordingly.

This approach presents a novel method of estimating the mood scale score trajectories of bipolar disorder patients. It incorporates a flexible function to describe longitudinal mood dynamics that is characterized by episode features. Through the grid-based episode detection approach, data can be explored in ways that address study-specific challenges such as sparsity and key regions of clinical interest can be identified. By using the PSO algorithm, data described by a complex function can be fit in an efficient manner and parameter estimates can be obtained. These extracted parameters can then be used as inputs to a variety of models. Because of the generality of the approach described here, components can be tailored according to specific study settings. However, the performance of this approach is driven by the decisions of the investigator. This includes the functional form of the mood function, objective function formulation, and the models used to process the estimates selected by the PSO algorithm. The appropriateness of this approach hinge on whether the assumptions truly represent the underlying data processes. In the next chapter, I evaluate the strengths and weaknesses of this and the Markov Chain with Bootstrap approach through simulation studies.

CHAPTER 6

Simulation Studies and Application Recommendations

While the approaches presented here have addressed some of the analytical problems that arise when working with bipolar disorder data, the statistical performance properties of these methods are unknown. In this chapter, I evaluate the strengths and weaknesses of these approaches through simulation studies to identify the settings in which they should most appropriately be used. I first present the results of the Markov chain with bootstrap approach, testing scenarios with varying data sampling frequencies, noise, sample size, and chain time units. Performance is evaluated based on the method's ability to correctly classify states and properly detect within- and between-group differences. A similar evaluation is done for the particle swarm optimization approach, looking at the overall fit to the data, correct episode detection, parameter estimate precision, and the proper detection of treatment differences. In both simulation studies, pre-treatment and post-treatment data are generated for a placebo (PL) and a treatment (TX) group. The agreement between the two methods is analyzed through two sets of cross-method simulation studies. In the first set, data are generated using a Markov chain mechanism and are analyzed using the particle swarm optimization approach. In the second set, data are generated based on a function of episode parameters and are analyzed using the Markov chain with bootstrap approach. This chapter concludes with recommendations for the application of these two methods.

6.1 Markov Chain with Bootstrap

The Markov chain with bootstrap approach requires specific parameters be set before its application to bipolar disorder data: (1) the threshold values (or method) for determining the mood states, (2) the data interpolation or imputation method, (3) the number of bootstrap replicates for inference, and (4) the chain time unit for analysis. Choices of these parameters for the simulation study are outlined below. I tested the method's sensitivity in settings with different data sampling frequencies, noise levels, and sample sizes. Described next are details of the data simulation process, followed by measures for evaluating performance, and finally the results.

6.1.1 Simulation Process

6.1.1.1 Data Generating Specifications

In the Markov chain framework, data are generated according to a transition matrix. The transition matrices used in the simulation study are listed in Table 6.1. The structure of these matrices restricts mood transitions between non-euthymic states. This results in episodes that are distinctively flanked by euthymic periods. Additionally, the transition probabilities of all non-euthymic mood states are equal and therefore exhibit identical behavior. While these structural restrictions may not reflect the real world patterns,especially for rapid cyclers, they create regularity in the data that is useful in testing performance properties which might otherwise be confounded with complexities of the disease process.

Pre- and post-treatment data for the PL group were generated by Matrix 1 (see Table 6.1), corresponding to no pre-post changes. For scenarios involving treatment effects for the TX group, pre-treatment data were generated by Matrix 1 (representing no baseline differences between TX and PL), while post-treatment data were generated by Matrix 2.

Table 6.1: Markov Chain with Bootstrap: Simulation Transition Matrices

	Matrix 1				Matrix 2			
	Eut	Dep	Man	Mix	Eut	Dep	Man	Mix
Eut	41.7	33.3	8.3	16.7	83.3	8.3	4.2	4.2
Dep	25.0	75.0	0	0	25.0	75.0	0	0
Man	25.0	0	75.0	0	25.0	0	75.0	0
Mix	25.0	0	0	75.0	25.0	0	0	75.0
stationary distribution	30.0	40.0	10.0	20.0	60.0	20.0	10.0	10.0

These values represent the probability of transitioning from the state indicated by the row to the state indicated by the column. Values are presented as percentages.

In scenarios testing the null case, both the pre-treatment and post-treatment data of the PL and TX groups were generated by Matrix 1. The stationary distributions of Matrix 1 and 2 result in a within-TX difference of +30, -20, 0, and -10 percentage-points for the euthymic, depressed, manic, and mixed states, respectively. These values also correspond to the expected between-group differences. These values were chosen for simulation because they incorporate a variety of effect sizes that reflect a favorable treatment outcome. In all scenarios, both the pre- and post-treatment follow-up times were set to 24 weeks each.

Mood scale scores were generated to match the simulated mood sequences using the function described in Equation (5.5). For HDRS data, the baseline severity was set to $b = 6$ and the amplitudes during a depressive or mixed episode were set to $\alpha_k = 15$. This corresponds to a baseline severity that is within a normal range (HDRS=6) and an episode peak severity that is within the severely depressed range (HDRS=6+15=21). For YMRS data, the baseline severity was set to $b = 6$ and the amplitudes were set to $\alpha_k = 7$ during a manic or mixed episode, corresponding to a baseline severity within a normal range (YMRS=6) and a peak severity that is within a mildly manic range (YMRS=6+7=13). Episode durations were automatically determined by generated mood sequence. The episode duration parameter σ was specified by the convention from Chapter 5 (see Section 5.3.1) that episode duration is equivalent to 4σ. For example, if the mood sequence indicated that an episode lasted for 28 days, then the matching σ_k was set to 7. These specifications were applied to both patient groups for the two study periods.

In these simulations, fixed thresholds were used to categorize subjects' mood states. For HDRS data, the threshold for being in a depressed state was 8 and was based on the cut-off commonly used in the literature[54] to describe a severity level corresponding to minor depression. A threshold of 7 was used for YMRS data corresponding to hypomania; this was also the cut-off value proposed in the three-arm randomized trial. Fixed thresholds were chosen to maintain a similar structure across all scenarios. However, the approach can accommodate other thresholding methods and this is an important area of future study.

6.1.1.2 Simulation Scenario Specifications

Simulation scenarios varied according to data sampling frequency, chain time unit, and noise. Data sampling rates were either weekly or once every two weeks. These were chosen based on the intended sampling frequency of the three-arm randomized trial (one week) and its observed frequency (approximately every two weeks). For similar reasons, chain time units of (a) seven days and (b) fourteen days were tested. Linear interpolation was used to impute data for scenarios with a seven-day chain time unit and a bi-weekly sampling rate.

To evaluate the effects of noise, five different levels were investigated. All noise introduced were normally distributed with mean zero, but varied according to standard deviation: 0, 0.25, 0.5, 1, and 2. For a given simulation scenario, the same degree of noise was added to both the HDRS and YMRS data. The choice of these noise levels was based on the differences between the baseline severities of the scales and the thresholds used, with higher noise levels expected to cause mood state misclassification.

Two sample sizes were selected for simulation study: $n = 10$ and $n = 50$. These reflect the sample sizes for each group. In all scenarios, the sample sizes of the PL and TX groups were equal. The smaller sample size choice was based on the number of patients in the three-arm randomized trial.

For each simulation scenario, 200 replicates were created. Inference for each replicate was based on 400 bootstrap samples. These values were selected based on computation time and limitations. Further in-depth study may require more replicates.

6.1.2 Measures

Three measures are used to evaluate performance: (1) correct state classification, (2) statistical power, and (3) Type I error. Correct state classification compares the generated mood state sequence to the one inferred by the Markov chain with bootstrap approach. The denominator for this measure is based on the generated sequence and the numerator is the number of matches found by the inferred sequence. For example, if the generated sequence contains 20 euthymic mood states and the inferred sequence matches 15 of them, then the calculated value is 75% for the euthymic state. Correct state classification is calculated by treatment group, study period, and mood state. Reported values are averaged across replicates.

Statistical power and Type I error calculations are based on the within- and between-group test results for the Markov chain with bootstrap approach applied to the replicate simulation data sets. Statistical significance is based on a p-value of $p < .05$. For each simulation scenario, the proportion of replicates resulting in statistical significance is calculated. For comparisons where a difference is expected, the proportion corresponds to statistical power. These comparisons are the within-group differences in the TX group and the between-group differences (excluding manic state comparisons). For comparisons where no difference is expected, the proportion corresponds to Type I error. These include the within-group differences in the PL group, the within-group difference of the manic state in the TX group, the between-group difference of the manic state, and all comparisons of the null case scenarios.

6.1.3 Results

6.1.3.1 Correct State Classification

The correct state classification results are summarized in Table 6.2. Only the results of the TX group are shown because that is the only group for which the pre- and post-treatment mood-shifting dynamics differed by study period and thus provides a better summary of the effects of sampling frequency, chain time unit, and noise under varying data generating mechanisms. Moreover, because classification is done at the level of the individual, it is not dependent on sample size. Therefore, only the simulation scenarios with $n = 50$ per group are displayed. Results indicate sensitivity to noise. Across all states, classification performance decreases as noise increases. Overall, the euthymic state exhibits the greatest sensitivity to noise. Closer inspection reveals that this pattern is a result of multiple factors.

One factor influencing euthymic state misclassification is threshold selection. For these simulated data, the selected HDRS and YMRS thresholds (8 and 7, respectively) are closer to the baseline severity value of 6 than the episode peak values (21 and 13, respectively). With thresholds closer to baseline severity values, the introduced noise is more likely to cause euthymic HDRS and YMRS data points to cross threshold values compared to non-euthymic data points, resulting in higher misclassification for the euthymic state.

Another factor driving euthymic state misclassification is data interpolation. In these scenarios, data are linearly interpolated. For a missing data point that lies in between a euthymic and non-euthymic point, the interpolated value will be the mean of these two values. However, because the threshold is closer to baseline severity value, the interpolated value is more likely to be classified as a non-euthymic data point. This classification bias leads to greater misclassification rates for euthymic data points. This is demonstrated by poorer euthymic state classification for scenarios involving interpolation relative to non-interpolation scenarios. Furthermore, because shifting out of the euthymic state is more probable, more episodes (and consequently more misclassification due to interpolation bias) are expected during the pre-treatment period relative to the post-treatment period. This pattern is observed in Table 6.2 and demonstrates an effect of episode frequency.

6.1.3.2 Power

The statistical power results for scenarios with a sample size of $n = 50$ per group are summarized in Table 6.3. Shown in the table are the TX within-group comparisons and the between-group comparisons. The underlying differences are +30, -20, 0, and -10 percentage-points for euthymic, depressed, manic, and mixed states, respectively, for both within- and between-group comparisons. Results indicate that, for a fixed noise level, a bi-weekly sampling rate with a 14-day chain time unit results in the lowest power. This suggests that the statistical power of this approach is impaired by low data volume and poor data granularity. Scenarios with a bi-weekly sampling rate and a 7-day chain time unit have a comparatively higher power, suggesting that interpolating points may be an effective method of recouping statistical power. Mood state comparisons indicate that euthymic state differences are well detected for both within- and between-group comparisons, although this result is expected because of the large underlying difference. Depressed state differences are also well detected, but performance suffers at high noise levels. This suggests that noise may be a factor in detecting moderate differences. For the mixed state, performance is reasonable for the within-group comparison, but not detectable for the between-group comparison. This suggests that smaller between-group differences may be difficult to detect.

The statistical power results for scenarios with a sample size of $n = 10$ per group are displayed in Table 6.4. The overall pattern is the same as with the larger sample size

Table 6.2: Markov Chain with Bootstrap: Correct State Classification

			pre-treatment				post-treatment			
sampling freq	chain time	noise	Eut	Dep	Man	Mix	Eut	Dep	Man	Mix
weekly	7 days	0	92.9	100	98.9	98.9	99.4	100	99.2	99.2
bi-weekly	14 days	0	91.3	100	94.0	94.0	99.1	100	96.3	96.5
bi-weekly	7 days	0	68.1	96.4	93.1	93.1	91.6	97.2	94.4	94.7
weekly	7 days	.25	85.6	98.0	93.4	92.2	95.7	98.1	93.7	93.2
bi-weekly	14 days	.25	80.6	95.4	89.0	86.3	92.2	96.0	89.8	89.3
bi-weekly	7 days	.25	63.3	94.7	91.3	90.0	89.2	95.4	91.5	91.4
weekly	7 days	.5	77.5	93.3	90.7	88.2	91.0	93.5	90.6	89.9
bi-weekly	14 days	.5	73.0	89.4	87.0	82.5	87.5	90.7	87.3	86.4
bi-weekly	7 days	.5	58.4	91.4	89.0	87.1	85.8	92.6	89.6	89.0
weekly	7 days	1	61.7	77.5	85.4	82.9	75.2	77.9	84.6	84.9
bi-weekly	14 days	1	60.2	73.9	83.4	78.5	72.4	75.9	81.7	82.4
bi-weekly	7 days	1	48.0	79.8	85.5	83.5	73.4	80.6	86.1	85.5
weekly	7 days	2	45.4	61.2	69.2	76.4	53.3	61.6	69.1	78.3
bi-weekly	14 days	2	45.2	58.8	67.9	73.4	52.6	60.8	67.7	76.8
bi-weekly	7 days	2	36.4	63.9	71.7	78.0	54.2	64.8	72.7	80.0

These values represent the correct state classification results of the TX group with $n = 50$ averaged across simulation replicates. Noise values correspond to the standard deviation of the normally distributed noise introduced. Values represent the proportion of correct classifications where the denominator value is based on the generated sequence. Values reported are given as percentages.

simulations, but there is a marked decrease in statistical power. Only comparisons in the euthymic state have acceptable levels of power, suggesting that only larger differences are adequately detectable in smaller sample size studies. Additionally, the lower data volume of scenarios with a bi-weekly sampling rate and a 14-day chain time unit results in poor statistical power. This suggests that this method may be inappropriate in small sample studies with infrequent sampling rates.

6.1.3.3 Type I Error

Type I error was assessed by testing null case scenarios in which Matrix 1 in Table 6.1 was used to generate pre- and post-treatment data for both the PL and TX group. Results are displayed in Table 6.5. Noise was not introduced in the null case scenarios to distill Type I error properties from the effects of noise, therefore representing a best case situation. Additionally, because the PL and TX data are generated by identical mechanisms, within-group comparisons of the two groups are comparable. Therefore, only the within-treatment comparisons for the PL group are presented. Results reflect a large degree of Type I error across all scenarios with the greatest sensitivity to small sample size. There do not appear to be clear patterns in Type I error as a function of sampling frequency, chain time, comparison type (within or between), or interpolation. Overall, these results indicate that statistical significance is being overstated and is particularly problematic in small sample size settings.

One potential source of inflated Type I errors is the lack of independence among the mood states. The Markov chain with bootstrap approach tests each mood state individually. However, the stationary distribution has an inherently dependent structure: increases in one mood state must equate to decreases in other states. Investigations indicate that a Bonferroni correction is sufficient to mitigate this inflation in most cases. The Markov chain with bootstrap approach involves four mood state comparisons, although the stationary distribution is a function of three independent components. Therefore, a correction factor of 3 is applied, resulting in a corrected significance level threshold of $p < \frac{.05}{3}$. Results with

Table 6.3: Markov Chain with Bootstrap: Power Analysis (n=50 per group)

			within-group			between-group		
sampling freq	chain time	noise	Eut	Dep	Mix	Eut	Dep	Mix
weekly	7 days	0	100	100	75.5	100	89.5	41.0
bi-weekly	14 days	0	100	87.0	45.0	99.5	54.5	24.0
bi-weekly	7 days	0	100	99.5	75.5	100	88.5	47.0
weekly	7 days	.25	100	99.5	80.0	100	88.0	47.5
bi-weekly	14 days	.25	100	86.5	58.0	100	60.5	22.0
bi-weekly	7 days	.25	100	100	82.5	100	86.5	44.0
weekly	7 days	.5	100	99.5	84.5	100	88.5	43.5
bi-weekly	14 days	.5	100	86.0	66.5	99.5	61.5	27.0
bi-weekly	7 days	.5	100	100	87.0	100	88.0	42.5
weekly	7 days	1	100	100	96.0	100	89.5	67.5
bi-weekly	14 days	1	100	90.5	79.0	100	60.0	42.0
bi-weekly	7 days	1	100	99.5	97.0	100	87.0	61.0
weekly	7 days	2	100	97.5	99.0	100	67.5	85.0
bi-weekly	14 days	2	100	70.0	82.5	95.5	40.5	47.0
bi-weekly	7 days	2	100	95.0	97.5	100	71.0	80.0

These values represent the percentage of replicates that result in statistical significance for the within-group comparisons of the TX group and the between-group comparisons. Statistical significance is defined by p < .05. These percentages correspond to statistical power for the euthymic, depressed, and mixed comparisons with effect sizes of +30, -20, and -10 percentage-points, respectively, for both within- and between-group comparisons. Differences in the manic state are not expected and are not shown.

Table 6.4: Markov Chain with Bootstrap: Power Analysis (n=10)

			within-group			between-group		
sampling freq	chain time	noise	Eut	Dep	Mix	Eut	Dep	Mix
weekly	7 days	0	89.5	58.0	30.0	79.5	29.0	16.0
bi-weekly	14 days	0	45.0	26.0	13.5	34.5	11.5	10.5
bi-weekly	7 days	0	83.5	64.0	27.5	75.0	32.0	16.5
weekly	7 days	.25	92.5	56.5	30.5	82.5	33.5	18.5
bi-weekly	14 days	.25	61.0	23.5	28.5	41.0	13.0	13.5
bi-weekly	7 days	.25	87.0	58.5	34.5	79.0	32.5	16.0
weekly	7 days	.5	98.0	60.0	35.5	90.5	41.0	19.0
bi-weekly	14 days	.5	61.5	31.5	21.5	49.0	15.5	12.5
bi-weekly	7 days	.5	92.5	57.5	31.5	85.5	34.5	17.0
weekly	7 days	1	97.5	63.5	52.5	85.0	35.0	29.5
bi-weekly	14 days	1	71.5	40.5	31.5	51.5	17.5	15.0
bi-weekly	7 days	1	95.0	57.5	46.5	84.0	30.5	20.0
weekly	7 days	2	88.5	40.5	61.0	71.5	21.5	39.5
bi-weekly	14 days	2	55.5	26.5	33.0	43.0	10.0	18.0
bi-weekly	7 days	2	86.5	45.5	57.0	69.5	21.0	29.0

These values represent the percentage of replicates that result in statistical significance for the within-group comparisons of the TX group and the between-group comparisons. Statistical significance is defined by $p <$ .05*. These percentages correspond to statistical power for the euthymic, depressed, and mixed comparisons with effect sizes of +30, -20, and -10 percentage-points, respectively, for both within- and between-group comparisons. Differences in the manic state are not expected and are not shown.*

Table 6.5: Markov Chain with Bootstrap: Type I Error: Null Case

			within-group				between-group			
sampling freq	chain time	n	Eut	Dep	Man	Mix	Eut	Dep	Man	Mix
weekly	7 days	50	9.5	5.0	6.0	3.5	5.5	7.5	5.0	5.0
bi-weekly	14 days	50	8.0	7.5	5.5	7.0	5.0	7.0	6.5	4.5
bi-weekly	7 days	50	8.0	6.0	5.0	6.0	5.0	6.0	6.0	5.0
weekly	7 days	10	5.0	9.5	15.0	8.0	6.5	9.0	12.0	9.0
bi-weekly	14 days	10	9.5	10.0	12.5	13.0	6.5	7.5	9.5	8.5
bi-weekly	7 days	10	9.5	7.5	9.0	12.0	8.0	9.5	5.5	12.5

These values represent the percentage of replicates that result in statistical significance in the null case (i.e., within- and between-group differences of zero). Signifiance is defined by $p < .05$*. For the within-group comparisons, only the PL group is shown. Noise was not introduced in any of the scenarios.*

the correction are displayed in Table 6.6. With the Bonferroni correction, Type I errors for all comparisons in the larger sample setting are less than 5%. Errors in the smaller sample setting are approximately 5%, with larger errors remaining for some within-group comparisons.

6.1.4 Conclusions

Simulation results suggest that correct mood state classification is contingent upon noise levels and threshold selection. If the euthymic state exhibits values that are closer to the threshold compared to non-euthymic states (or vice versa), then misclassification bias may be introduced. Related to this issue is the misclassification of interpolated data points. Depending on the interpolation method and the selected threshold, values imputed in between euthymic and non-euthymic data points may favor one state over another. Therefore, in addition to noise levels and threshold selection, the underlying episode frequency must be considered.

Table 6.6: Markov Chain with Bootstrap: Type I Error: Null Case (corrected)

			within-group				between-group			
sampling freq	chain time	n	Eut	Dep	Man	Mix	Eut	Dep	Man	Mix
weekly	7 days	50	4.0	1.5	2.5	1.5	1.5	2.0	4.0	2.5
bi-weekly	14 days	50	4.0	3.5	2.5	4.5	2.5	4.0	3.0	2.0
bi-weekly	7 days	50	3.0	3.0	2.5	3.0	3.0	3.0	3.5	2.5
weekly	7 days	10	1.0	6.5	10.0	4.5	4.0	4.5	6.0	3.5
bi-weekly	14 days	10	5.0	6.5	8.5	8.0	2.0	3.0	4.5	5.0
bi-weekly	7 days	10	5.0	4.5	5.5	7.0	2.5	5.5	2.0	5.0

These values represent the percentage of replicates that result in statistical significance in the null case (i.e., within- and between-group differences of zero). Signifiance is Bonferroni corrected and is defined by p $< \frac{.05}{3}$. *For the within-group comparisons, only the PL group is shown. Noise was not introduced in any of the scenarios.*

As expected, statistical power is related to sample size, data volume, and effect size. Analyses suggest that the method performs well in larger sample settings with frequently sampled measures. In cases where data are infrequently sampled, values may be imputed to recoup some power, although potential misclassification bias resulting from imputation and noise must be considered to ensure proper analysis. In small sample settings, the method performs well in detecting large differences, although statistical power is limited when data are infrequently sampled and values are not imputed.

Type I error analysis yields values higher than 5%, with substantially greater errors in the smaller sample size setting. These results indicate that adjustments are necessary to obtain nominal acceptance levels. The higher Type I error rates may be driven by the dependent structure of the stationary distribution. Applying Bonferroni corrections yields acceptable Type I errors for most comparisons, although further corrections may be necessary for some cases in a smaller sample size setting.

6.2 Particle Swarm Optimization

The particle swarm optimization approach can be summarized by four steps: (1) mood function formulation, (2) episode region detection, (3) particle swarm optimization, and (4) modeling to detect group differences. Each of these steps requires specifications that must consider data characteristics such as sample size, data sparsity, and noise. In the simulation studies presented here, I focus on a small sample setting and investigate the effects of data sparsity and noise. Because episode region detection is a critical aspect to this approach, I also investigate the impact of episode location and the impact of overlapping episodes. Details of the data simulation process and performance metrics used are described next, followed by simulation testing results.

6.2.1 Simulation Process

6.2.1.1 Mood Function Parameters

For the purpose of simulation study, attention is focused on the HDRS. Data for PL and TX patients are generated according to the mood function described by Equation (5.5). Data for $n = 10$ patients per group are simulated and a total of 200 replicates are used for each simulation scenario. Follow-up times are set to 24 weeks for each of the study periods. Within each patient, all simulated episode durations and amplitudes are identical by study period. Mood function parameters are specified for each patient as follows. For both the PL and TX groups, baseline severity is set to $b = 6$, corresponding to an HDRS score within the normal range. For the PL group, episode amplitudes (α_k) are drawn from a truncated normal distribution with a mean of 15, standard deviation of 2, and truncation at 4 and 26. A truncated distribution is used to prevent the simulation of nonsensical values and to ensure proposed values are within the particle swarm search space. The mean amplitude value of 15 corresponds to an HDRS score of 21 at episode peak, which falls within the severely depressed range. Episode durations ($4\sigma_k$) are also drawn from a truncated normal distribution with a mean of 28, standard deviation of 3.5, and truncation at 7 and 49. This corresponds to a mean duration of 4 weeks and ranges from 1 to 7 weeks. For the TX group, the same distributions as the PL group are used for episodes during the pre-treatment period, while the post-treatment period distributions differ only in the means. Specifically for the TX group, the mean post-treatment episode amplitude is 7 (corresponding to a mean peak HDRS score of 13) and the mean episode duration is 21 days.

Two different episode location (μ_k) settings are studied. The first scenario – referred to here as the fixed-episode setting – forced the episode locations for all patients to occur at -126, -84, -42, 42, 84, and 126 days, where negative days correspond to the pre-treatment period. In the second setting – the overlap-episode setting – the number of episodes and their locations are allowed to vary. The number of episodes is determined by a truncated Poisson distribution with a rate parameter of 8 episodes per year and truncation at 0 and 10. A truncated distribution is used to ensure that the number of episodes simulated are within reason and can be identified for the time horizon and granularity of the data. For the post-treatment period of the TX group, a rate parameter of 4 episodes per year is used. Locations are determined by a Dirichlet process to allow variation in placement of individual episodes (often resulting in episode overlap), even though on average they are equally spaced.

To assess Type I error, null case scenarios are studied. In these scenarios, mood function parameters in the TX group are identical to those of the PL group.

6.2.1.2 Episode Region Detection Parameters

The collection of episodes used in the episode region detection process is specified as follows. For episode amplitude, the specified range has a minimum of 4 and a maximum equivalent to the patient's largest HDRS score. Ten equally-spaced amplitudes spanning this range are selected. For episode duration, the range is specified with a minimum equivalent to twice the sampling frequency and a maximum of 49 days. Ten equally-spaced durations spanning this range are selected. This yields a total of 100 different episodes in the episode set. Episode fitness values for this set is calculated at a weekly interval from -168 days (i.e., 168 days before treatment randomization) to 168 days. Episode fitness is calculated using the formula described by equation 5.8.

6.2.1.3 PSO Parameters, Sampling Rate, and Noise

The velocity updating parameters are identical to those specified by SPSO-2011 (see Section 5.3.4). Boundaries for episode amplitude and duration are identical to the ranges used in the episode region detection process. To reduce computation run time, some simplifications are made. First, the stopping criteria are reduced to a maximum of 100 total iterations and 25 stagnate iterations. Second, no penalties are incorporated into the objective function. Three different sampling rates are studied: daily, weekly, and bi-weekly (i.e., every 2 weeks). Noise levels are normally distributed with mean zero and vary in standard deviation values: 0, 1, and 2. Noise is not introduced in null case scenarios.

6.2.2 Measures

To assess overall fit, the root mean squared error (RMSE) is used. This is calculated by comparing the true underlying HDRS scores with those predicted by the PSO fit. The RMSE is calculated for each patient and is averaged across patients by treatment group for each replicate. Descriptive statistics are used to summarize the overall fit across replicates.

To assess the accuracy of the episode region detection method, three rates are used: (1) true positive detection, (2) false positive detection, and (3) false negative detection. True positive detection counts the number of true episodes and calculates the proportion of these episodes that are uniquely identified by an episode region. For an episode to be considered as uniquely identified, it must be the only episode in the specified region. False positive detection counts the number of episode regions and calculates the proportion of regions that do not encompass a true episode. This describes instances where the detection method incorrectly suspects there to be an episode. False negative detection considers the set of true episodes and calculates the proportion of true episodes that are not uniquely identified by a detected episode region. This describes instances where the method fails to identify an episode or specifies a single region that contains more than one episode. In cases where multiple episodes are encompassed by a region, all but one of those episodes are counted toward the false negative rate.

To evaluate the fit of episode durations and amplitudes, the RMSEs comparing the predicted parameter estimates to the true parameters is calculated. These values are calculated for each patient and are averaged across patients by treatment group and study period for each replicate. Descriptive statistics are used to summarize the fit across replicates.

To assess statistical power and Type I error, the same linear mixed-effect models employed in the three-arm randomized trial are used. Both episode duration and amplitude are modeled. Attention is focused on the treatment by time interaction. Using $p < .05$ as the significance threshold, parameter estimates are examined and statistical power is calculated as the proportion of replicates resulting in significance. This process is also applied to null case scenarios and the calculated value is interpreted as Type I error.

6.2.3 Results

6.2.3.1 Overall Fit

Overall fit results are presented in Table 6.7. In general, the overall fit does not exhibit gross errors. In the fixed-episode setting, the discrepancy between the true and predicted HDRS scores is less than 1 point on average and is robust to noise. In the overlap-episode setting,

Table 6.7: PSO: Overall Fit

		fixed-episode				overlap-episode			
		PL		TX		PL		TX	
sampling freq	noise	mean	SD	mean	SD	mean	SD	mean	SD
daily	0	0.00	0.00	0.01	0.03	1.17	1.35	0.95	1.27
weekly	0	0.00	0.00	0.01	0.02	0.84	1.04	0.60	0.87
bi-weekly	0	0.24	0.21	0.31	0.13	0.74	0.96	0.62	0.75
daily	1	0.24	0.04	0.26	0.10	1.26	1.32	1.13	1.34
weekly	1	0.63	0.11	0.66	0.13	1.23	0.94	1.13	0.87
bi-weekly	1	0.23	0.21	0.31	0.13	0.75	0.96	0.63	0.76
daily	2	0.47	0.08	0.60	0.30	1.45	1.37	1.48	1.55
weekly	2	1.32	0.21	1.37	0.26	1.75	0.89	1.75	0.92
bi-weekly	2	0.23	0.20	0.32	0.13	0.73	0.96	0.63	0.77

These values represent the root mean squared error differences between the true underlying HDRS scores and those predicted by the PSO fit averaged across patients by treatment group. Noise values correspond to the standard deviation of the normally distributed noise introduced. These values are interpreted as the average error of the fit.

the discrepancies are slightly higher, reaching values between 1 and 1.5. Moreover, standard deviations are greater relative to their fixed-episode counterparts in scenarios with a higher sampling rate. This suggests that, despite greater data granularity, the method is limited in capturing the underlying mood curve in the overlap-episode setting. Instead of informing the fit with greater data volume, the frequent sampling rate results in more data points that are poorly captured by the fitted curve, equating to higher RMSE and standard deviation values. These results suggest that the method overall is able to detect the underlying mood function with a fair amount of precision, but performance suffers when episodes overlap. The only anomaly in these findings is the scenario with a weekly sampling frequency and a noise level of 2. This scenario represents a unique case and its performance is examined in further detail later.

6.2.3.2 Episode Region Detection

Episode region detection results are presented in Table 6.8. In the fixed-episode setting, detection is near perfect. True positive detection rates are nearly 100%, while false positive and false negative detection rates are less than 1%. The overlap-episode setting reflects poor episode detection. Approximately 50% to 60% of the true episodes are uniquely identified by the specified episode regions across all scenarios. In comparison, false positive detection rates are lower, ranging from 2.9% to 9.6%, indicating that the method is conservative in episode specification. However, the false negative rate is high, indicating that some episodes are not being detected or regions are encompassing more than one true episode. Overall, these findings indicate that distinct episodes across time is an important feature for proper episode range detection. In cases where episodes are less distinct, the method is conservative and specifies fewer episodes. While these episodes overlap according to their mathematical specification by the mood function, it is unclear whether these overlapping episodes clinically represent multiple periods of elevated mood. These episodes may collectively represent a single episodic period and its characterization as a single episode may be a fair reflection of the clinical trajectory of the patient. Once again, the only exception to these results is the scenario with a weekly sampling frequency and a noise level of 2.

6.2.3.3 Episode Duration Estimation

Discrepancies in episode duration estimation are presented in Table 6.9. In general, episode duration estimates are affected by both noise and sampling frequency, with the latter having a much stronger effect. This is expected as data sparsity makes it difficult to pinpoint episode onset and resolution. In the fixed-episode setting, RMSE values roughly range from 0 to 10 days with noticeably higher values for bi-weekly sampling rate scenarios. Post-treatment estimates for the TX group indicate greater errors. Because the episode durations of the TX group during this period are shorter, these larger RMSE values suggest that there may

Table 6.8: PSO: Episode Range Detection

		fixed-episode			overlap-episode		
sampling freq	noise	TP	FP	FN	TP	FP	FN
daily	0	100	0.0	0.0	53.6	2.9	49.9
weekly	0	100	0.0	0.0	57.6	2.8	41.5
bi-weekly	0	100	0.0	0.0	57.9	7.4	43.5
daily	1	99.6	0.0	0.4	52.8	4.6	51.1
weekly	1	99.5	1.4	0.5	57.4	9.6	40.2
bi-weekly	1	100	0.0	0.0	56.7	7.5	44.8
daily	2	98.1	0.6	2.0	52.0	7.3	53.1
weekly	2	97.5	8.0	2.4	55.4	22.3	38.9
bi-weekly	2	100	0.0	0.0	56.6	7.9	44.9

These values represent the true positive (TP), false positive (FP), and false negative (FN) detection rates averaged across simulation replicates. Noise values correspond to the standard deviation of the normally distributed noise introduced. Values listed are in percentages.

Table 6.9: PSO: Episode Duration Error

		fixed-episode				overlap-episode			
		PL		TX		PL		TX	
sampling freq	noise	pre	post	pre	post	pre	post	pre	post
daily	0	0.00	0.00	0.04	0.41	7.87	8.17	8.01	6.41
weekly	0	0.00	0.00	0.02	0.13	7.79	7.62	7.31	5.71
bi-weekly	0	3.03	3.02	2.87	9.71	7.74	7.35	7.83	7.81
daily	1	0.99	0.99	0.99	2.15	8.05	7.97	7.94	6.59
weekly	1	3.03	3.27	2.89	4.97	8.33	8.57	8.33	7.11
bi-weekly	1	3.00	2.89	2.84	9.56	7.71	7.61	7.92	8.01
daily	2	2.36	2.40	2.05	4.78	8.54	8.45	8.48	8.22
weekly	2	7.30	7.79	6.71	10.11	10.11	10.53	10.41	9.92
bi-weekly	2	2.90	2.87	2.97	9.54	7.67	7.57	7.90	8.22

These values represent the root mean squared error differences between the true underlying episode durations and those predicted by the PSO fit, averaged across patients by treatment group and study period, and averaged across simulation scenario replicates. Noise values correspond to the standard deviation of the normally distributed noise introduced. These values are interpreted as the average error of episode duration estimation measured in days.

be a lower limit to estimation accuracy based on the sampling frequency. This is expected as acute episodes are difficult to detect when data sparsity is high. Scenarios of the overlap-episode setting reflect greater RMSE values compared to their fixed-episode counterparts. The discrepancies within overlap-episode scenarios are roughly comparable with RMSE values approximately ranging from 6 to 9 days. Like the fixed-episode scenarios, slight patterns across overlap-episode scenarios are observed for noise, although the sensitivity to sampling frequency is less pronounced. This may indicate that the estimation errors associated with episode overlap may supersede the effects of noise and sampling frequency.

6.2.3.4 Episode Amplitude Estimation

Results presented in Table 6.10 indicate that amplitude estimation is sensitive to overlapping episodes. In the fixed-episode setting, RMSE values are generally no more than 1 point and performance during the pre-treatment and post-treatment periods are identical for each treatment group. In the overlap-episode setting, performance is worse, with RMSE values of approximately 7 points across most scenarios. Results also differ by treatment period for the TX group, reflecting RMSE values of approximately 2 points during the post-treatment period. Investigations suggest that this improved performance may be reflective of the decreased episode rate during the post-treatment period. Although episodes may overlap during this period, the decreased episode rate makes for fewer overlaps and episodes are more distinct. This mimics the fixed-episode setting and results in improved estimation.

6.2.3.5 Episode Duration: Power and Type I Error

The statistical power and Type I error results of the episode duration model are summarized in Table 6.11. Statistical power for detecting a treatment by time interaction is high in the fixed-episode setting, except in scenarios with a bi-weekly sampling rate. A similar trend is observed in the overlap-episode setting, although power is much lower overall. No clear patterns are not observed across noise levels for either the fixed- or overlap-episode settings. These results suggest that the statistical power to detect a difference in episode duration is primarily dependent upon a sampling rate that is frequent enough to capture the difference and non-overlapping episodes.

Type I error results for the null cases (rows 10 through 12 in Table 6.11) indicate acceptable coverage in the fixed-episode setting. For the treatment by time interaction, Type I error is approximately 5% across all sampling frequency scenarios, and the coverage for the treatment and time effects are 5% on average. Nominal significance in the overlap-episode

Table 6.10: PSO: Episode Amplitude Error

		fixed-episode				overlap-episode			
		PL		TX		PL		TX	
sampling freq	noise	pre	post	pre	post	pre	post	pre	post
daily	0	0.00	0.00	0.01	0.04	6.59	6.58	6.58	1.75
weekly	0	0.00	0.00	0.01	0.01	6.75	6.73	6.49	1.78
bi-weekly	0	0.39	0.39	0.46	0.40	7.00	6.95	7.37	2.18
daily	1	0.44	0.43	0.44	0.50	6.84	6.69	6.64	1.88
weekly	1	1.51	1.83	1.41	1.45	7.20	7.18	6.95	2.30
bi-weekly	1	0.38	0.38	0.46	0.40	7.31	7.24	7.37	2.20
daily	2	1.02	1.05	0.89	1.05	6.68	6.76	6.44	1.93
weekly	2	4.18	4.71	3.60	2.93	8.00	8.22	8.40	3.31
bi-weekly	2	0.37	0.38	0.46	0.41	7.35	7.15	7.44	2.25

These values represent the root mean squared error differences between the true underlying episode amplitudes and those predicted by the PSO fit, averaged across patients by treatment group and study period, and averaged across simulation scenario replicates. Noise values correspond to the standard deviation of the normally distributed noise introduced. These values are interpreted as the average error of episode amplitude estimation measured in HDRS scale points.

setting are inflated. For the treatment by time interaction, Type I error ranges from 7% to 9.5%, while the coverage for the main effects are 8% on average. There are no clear patterns across sampling frequency for either the fixed- or overlap-episode settings. These results suggest that Type I errors are not maintained when episodes overlap, but acceptable coverage may be attainable if episodes are distinct. Achieving nominal significance in analyses with data with episode overlap may involve modifications to the mood or objective function and is a topic of further study.

6.2.3.6 Episode Amplitude: Power and Type I Error

Presented in Table 6.12 are the statistical power and Type I error results for the episode amplitude model. Statistical power for the treatment by time interaction is high for both fixed-episode and overlap-episode settings. Across all scenarios of the fixed-episode setting, all replicates resulted in a statistically significant treatment by time interaction effect. For scenarios in the overlap-episode setting, statistical power is 93.8% on average. These results suggest that in spite of sampling frequency, noise, and potential episode overlap, the power to detect an eight-point difference in amplitude between groups – an effect size large enough to differentiate between mild and moderate or moderate and severe depression – is maintained. However, further study is required to examine the performance of other effect sizes.

Type I error results for the null cases (rows 10 through 12 in Table 6.12) generally reflect acceptable coverage in the fixed-episode setting. For the treatment by time interaction, Type I error across all sampling frequency scenarios is 5.3% on average, while the mean error for the treatment and time effects are both 4.8%. Nominal significance in the overlap-episode setting are slightly inflated. For the treatment by time interaction, Type I error is within 5% for only one scenario. While Type I error is mostly within limits for the treatment effect, coverage is high for the time effect. No clear patterns are observed across sampling frequency for either the fixed- or overlap-episode settings. These results suggest that statistical significance may be overstated when episodes overlap.

Table 6.11: PSO: Power and Type I Error: Episode Duration Model

		fixed-episode			overlap-episode		
sampling freq	noise	TX	time	$TX \times time$	TX	time	$TX \times time$
daily	0	5.0	5.5	87.5	5.0	8.5	62.5
weekly	0	7.0	5.0	90.0	5.5	6.5	52.5
bi-weekly	0	6.0	8.0	27.0	8.0	7.5	21.0
daily	1	6.5	5.0	87.0	8.0	7.0	63.5
weekly	1	7.0	7.5	86.0	4.5	8.0	60.0
bi-weekly	1	3.5	9.5	25.5	8.5	10.0	18.5
daily	2	8.5	8.0	90.0	8.0	9.0	63.5
weekly	2	8.0	4.0	67.5	5.0	8.0	48.2
bi-weekly	2	2.5	13.5	28.5	7.0	7.0	19.0
daily	0	3.0	7.0	5.5	4.5	10.0	9.5
weekly	0	5.0	6.0	4.5	9.0	7.0	7.0
bi-weekly	0	5.0	4.0	5.5	9.0	8.5	9.0

These values represent the proportion of replicates resulting in a statistically significant result for the given effect. The first nine scenarios include a seven-day decrease in episode duration. For these scenarios, the $TX \times time$ *column reflects statistical power. The last three rows are null case scenarios and do not include a difference in episode duration. For these scenarios, the* $TX \times time$ *column reflects Type I error. For all scenarios, the* TX *and time columns reflect Type I error. Noise values correspond to the standard deviation of the normally distributed noise introduced. Values are in percentages.*

Table 6.12: PSO: Power and Type I Error: Episode Amplitude Model

		fixed-episode			overlap-episode		
sampling freq	noise	TX	time	$TX \times time$	TX	time	$TX \times time$
daily	0	6.5	10.0	100	6.0	11.5	95.5
weekly	0	5.0	7.0	100	5.0	8.0	95.0
bi-weekly	0	8.0	5.5	100	8.0	12.0	94.5
daily	1	5.0	7.5	100	6.5	7.5	97.0
weekly	1	6.5	9.5	100	8.0	14.0	93.5
bi-weekly	1	8.5	7.0	100	8.0	10.0	95.0
daily	2	2.5	5.0	100	7.5	10.0	93.5
weekly	2	11.0	8.5	100	8.5	14.1	87.9
bi-weekly	2	4.5	7.0	100	8.0	11.5	92.0
daily	0	5.0	5.0	8.5	4.5	10.5	5.0
weekly	0	7.0	5.0	4.5	7.0	7.5	9.0
bi-weekly	0	2.5	4.5	3.0	3.5	11.5	9.0

These values represent the proportion of replicates resulting in a statistically significant result for the given effect. The first nine scenarios include an eight-point decrease in episode amplitude. For these scenarios, the $TX \times time$ *column reflects statistical power. The last three rows are null case scenarios and do not include a difference in episode amplitude. For these scenarios, the* $TX \times time$ *column reflects Type I error. For all scenarios, the* TX *and time columns reflect Type I error. Noise values correspond to the standard deviation of the normally distributed noise introduced. Values are in percentages.*

6.2.4 Conclusions

Simulation studies of the PSO approach reflect that it is robust to data sparsity and noise when detecting the underlying signal. However, episode detection suffers dramatically when episodes are not distinct over time. If two episodes are relatively close to each other, the episode region detection method may identify a wide interval that encompasses both episodes and choose to fit one episode in that space. Although this may indicate poor episode detection in a statistical sense if episodes are viewed as coming from a mixture model, it may actually reflect proper clinical interpretation. For example, if two episodes are close to each other, the HDRS scores in between the two peak severities may still exceed threshold levels. Even though the curve is in some sense generated by two episodes or events, the lack of episode resolution in between them is suggestive of one continuing episode. For example, two overlapping episodes may manifest as a bimodal curve whose intermediate nadir point remains above the specified threshold levels and be clinically interpreted as a single episode. This explains the poor performance in duration and amplitude estimation in the overlap-episode setting, while greater precision is observed in the fixed-episode setting. When detecting between-group differences, the method performs well for episode amplitude changes, although significance may be overstated in the overlap-episode setting. Performance in detecting episode duration differences is impacted by episode overlap, but is largely associated with the sampling frequency. Further study is required to investigate the performance for other effect sizes and sample sizes.

Throughout all investigations, the scenario with a weekly sampling rate and a noise level of 2 exhibits consistently poor performance and represents a very unique case. Examination of this scenario's replicates reveal that episode range detection is markedly poor. This is driven by two things: (1) a data sampling rate that matches the episode region detection's time grid, and (2) high noise levels. At high noise levels, an elevated HDRS score due to noise may cause this region to be interpreted as an episode-probable location. However, because the sampling rate matches the episode region detection's time grid exactly (i.e., every 7 days), there are no intermediate points to smooth over this region and have it be classified otherwise. This causes noise fluctuations to be detected as episodes. This error cascades into all other analyses and explains the overall poor performance of this scenario. This suggests that the episode region detection process should calculate episode fitness at an interval that is more frequent than the data sampling rate and should not propose episodes whose amplitude is less than the expected noise level.

6.3 Cross-Method Simulations

The bipolar disorder features that these two methods model are conceptually different. The Markov chain with bootstrap approach models the mood-shifting dynamics and the amount of time spent in each mood state. The PSO approach estimates a mood function parameterized by episode characteristics and models differences across time. Although parameterized differently these approaches are intrinsically related. For example, a decrease in the time spent in the manic state may be related to decreases in episode duration or frequency. Similarly, a decrease in episode severity may be related to less time spent above a mood state threshold and be translated as a decrease in time spent in that mood state. To study the relationships between these features and examine each method's ability to detect changes, cross-method simulations are used. In the first set of simulations, data are generated by a Markov chain mechanism and are analyzed using the PSO approach. In the second set of simulations, the roles are reversed: data are generated by a mood function and are analyzed using the Markov chain with bootstrap approach. Data simulation details and results are described next.

6.3.1 Simulation Process

Markov chain cross-method data are generated by the processes described in Section 6.1.1 with a few differences. In summary, these changes restrict treatment effects to a reduction in the time spent in the depressed state and an increase in the time spent in the euthymic state through transition matrix modifications. Additionally, differences in baseline severity and amplitude between the two scales are not implemented to simplify the analysis. The baseline severity for both the HDRS and YMRS is set to 6 points, the amplitude during the pre-treatment period is 15 points for both scales, and the amplitude during the post-treatment period is set to 7 points for the TX group (amplitude remains at 15 points for the PL group). The stationary distribution during the pre-treatment period for both groups is set to one-half of the time spent in the depressed state and one-sixth of the time spent in each of the remaining states. The PL group post-treatment stationary distribution is identical to the pre-treatment stationary distribution. For the TX group, the post-treatment stationary distribution is set to one-half of the time in the euthymic state and one-sixth of the time spent in each of the remaining states, resulting in a 33% decrease in the time spent in the depressed state and a 33% increase in the time spent in the euthymic state. No differences are simulated for the manic and mixed states. The transition matrices used to simulate these data are structurally similar to those described in Table 6.1, although for the post-treatment period of the TX group, the probability of transitioning from the depressed state to the euthymic state is increased from 25% to 45%, resulting in a decrease in depression duration. A total of 100 replicates are simulated and data are analyzed using the PSO approach with the algorithm parameters described in Section 6.2.1. The linear mixed-effect model structure from the three-arm randomized trial analysis is used to detect group differences in episode duration, amplitude, and frequency. Attention is focused on detecting between-group differences.

Particle swarm cross-method data are generated by processes similar to those described in Section 6.2.1. Both HDRS and YMRS data are simulated. Baseline severity and amplitude parameters are identical to those of the Markov chain cross-method data: a baseline severity of 6 points, a pre-treatment amplitude of 15 points, and a post-treatment amplitude of 15 and 7 points for the PL and TX group, respectively. Episode duration for both groups and rating scales are set to 21 days, although the episode frequencies differ. For the pre-treatment period of both groups, 6 depression and 2 manic episodes are simulated. For the post-treatment period of the PL group, these parameters remain unchanged. For the post-treatment period of the TX group, the number of depressed episodes is decreased to 2 episodes. To prevent episode overlap, the locations of simulated episodes are restricted to be at least 14 days apart. A total of 500 replicates are simulated and data are analyzed using the Markov chain with bootstrap approach with the algorithm parameters described in Section 6.1.1. A Bonferroni-correction is implemented and a p-value of $p < \frac{.05}{3}$ is used to denote statistical significance.

For both cross-method simulations, data are generated at a weekly sampling rate for $n = 50$ patients per group with a pre-treatment and post-treatment period of 252 days each. To simplify cross-method analysis, noise is not added to the data.

6.3.2 Results

In the Markov chain cross-method data, patients in the TX group spend 33% less time in the depressed state and have an increased probability of transitioning out of the depressed state. Simulation results indicate that changes in episode amplitude are readily detectable. Significant amplitude differences are observed across 100% and 99% of replicates along the HDRS and YMRS, respectively. Performance differed for episode duration and frequency. Statistically significant between-group episode duration differences along the HDRS are observed in 72% of the simulated replicates, while significant episode frequency differences along the HDRS are observed in 49% of the replicates. These findings suggest that underlying mood-shifting dynamics may be difficult to pinpoint by the PSO approach. Episode duration and frequency are analyzed separately, but stationary distribution changes may manifest along both of these characteristics.

In the PSO cross-method data, patients in the TX group have a decreased frequency of depressed episodes and decreased episode amplitudes in both mania and depression. Application of the Markov chain with bootstrap approach to these data suggests that these decreases manifest as stationary distribution differences. For all replicates, significant between-group differences of the euthymic, depressed, and mixed state are observed (significant differences in the manic state are observed in 2.4% of the replicates). While this suggests high agreement between the two methods, it also suggests that changes along one scale impact multiple states. The PSO cross-method data for the TX group are designed to reflect less time spent in the depressed state (and consequently more time in the euthymic state), but based on the Markov chain with bootstrap results, such changes may also impact mixed state results.

6.3.3 Conclusions

In general, these results suggest agreement between the two methods. Underlying differences in the mood-shifting mechanisms manifest as differences in the PSO models. Similarly, differences in episode characteristics manifest as differences in mood state stationary distributions. However, it is difficult to directly correlate features based on the results.

Significant differences in stationary distribution may imply differences in episode characteristics, but do not indicate which characteristics are driving the results. Simulations studied here only examined an underlying stationary distribution difference in the euthymic and depressed states. Even in this simplified example, the cross-method simulation results suggest that the difference may be driven by both episode duration and frequency. Because these features are modeled separately in the PSO approach, the collective effect is not captured. More complex settings, such as differences involving the mixed state or a combination of non-euthymic states, may further convolute the relationship between stationary distribution and episode characteristics. This may result in significant findings according to the Markov chain with bootstrap approach, but inconclusive findings when applying the particle swarm optimization approach.

The PSO cross-method data simulation results indicate that changes in episode characteristics coincide with stationary distribution differences, but it is unclear how these differences manifest. In the simulations studied here, episode differences are mainly introduced on the depression axis, but results indicate differences in the euthymic, depressed, and mixed states. This highlights the dependent structure of the stationary distribution, but complicates the relationship between changes in episode characteristics and mood-shifting dynamics.

Overall, the cross-method simulation results suggest that there is a relationship between mood-shifting dynamics and episode characteristics, but separate analyses of these features may not necessarily result in agreement. Mood-shifting dynamics may be complex and not adequately described by analyzing the HDRS and YMRS separately. Similarly, a mixture of episode features may be driving the changes in stationary distribution. While it is reassuring that the results in the cross-method simulations reflect agreement between the methods, further study is necessary to investigate more complex scenarios and effect sizes, as well as approaches that jointly model the data.

6.4 Application Recommendations

6.4.1 Markov Chain with Bootstrap

Simulation studies suggest that the Markov chain with bootstrap approach adequately detects stationary distribution mood state differences. Like many statistical methods, overall performance of this approach improves with larger sample sizes. In particular, bigger samples allow for the detection of smaller effect sizes; small sample studies many only be adequately powered to detect large differences. However, methodological choices can improve performance in both small and large sample settings.

One consideration to make is the data sampling rate. This method assumes that mood-shifting dynamics can sufficiently be characterized according to the specified sampling interval. Therefore, it is important to select a sampling rate that best matches this interval and to set this time period as the chain time unit. For example, if it is suspected that mood can adequately be characterized at weekly intervals, then it is important to sample data at this frequency and to select a chain time unit of seven days for the analysis. However, if data cannot be sampled at this rate, then imputation is recommended to recoup statistical power. The selection of an appropriate imputation method must consider the thresholds used in determining mood states, and the potential for biased mood state misclassification – particularly during episode onset and resolution – must be examined. If the imputation method proposes data point values that are more likely to be on one side of the threshold than the other, then results will be biased and another data interpolation method (or threshold) must be used.

Related to the issue of misclassification bias are threshold selection and data measurement noise levels. It is recommended that the selected threshold maximally discriminate between typical euthymic state and non-euthymic state values. This ensures that classification remains unbiased in light of noise fluctuations. If the threshold is more similar to typical non-euthymic state values, then variance due to noise may favor data points to be classified as a non-euthymic state (or vice versa). This recommendation, however, complicates the selection of a fixed threshold in settings with a high degree of patient variability. In these cases, patient-specific thresholds – such as the subject's pre-treatment median – are recommended to accommodate patient heterogeneity. Moreover, the interpolation method used to impute data must also consider the selected threshold and facilitate bias reduction.

One challenge to this approach is the overstatement of statistical significance, as evidenced by the inflated Type I errors in the simulation studies. It is suspected that this is a result of the dependent structure of the stationary distribution. This method compares mood states individually, but the mood states are not independent: increases in one mood state imply decreases in other mood states. To rectify this issue, it is recommended that a Bonferroni-adjusted significance value of $\alpha = \frac{.05}{3}$ be applied. The simulation studies suggest that this corrects the Type I error for most comparisons.

6.4.2 Particle Swarm Optimization

The particle swarm optimization approach can be summarized by four steps: (1) mood function formulation, (2) episode region detection, (3) particle swarm optimization, and (4) modeling to detect group differences.

The first step to the PSO approach is mood function formulation. Therefore, the application of this method hinges on proper episode characterization. The simulation results indicate that statistical performance improves dramatically if episodes are distinguishable across time and do not overlap. If episode overlap is a primary feature of interest, then the episode functional form and parameterization must reflect this. This highlights the importance of choosing an appropriate functional form of an episode, but also underscores the method's flexibility to accommodate various parameterizations. Moreover, it emphasizes clinical relevance in the model building process. While overlap may present complications in capturing nuanced fluctuations, the overall trend may be interpreted as a single episode from a clinical perspective.

Like the Markov chain with bootstrap approach, sampling frequency is an important consideration in applying the PSO approach. The sampling rate primarily impacts episode duration estimation and the detection of duration changes. If data are sparse, then the data may not adequately capture acute episodes or small changes in duration. If episode duration is of primary interest, it is recommended that data be sampled frequently to capture subtle shifts in duration.

Unlike episode duration, the method is robust in estimating amplitude and detecting amplitude changes. However, complex amplitude characterization is still possible and is primarily driven by episode overlap and noise. If two episodes overlap, this region may be characterized as a single episode and result in an averaged amplitude value. Therefore, it is important to properly specify an episode's functional form to accommodate for these instances. It also highlights the importance of characterizing episodes as distinguishable components across time. Additionally, amplitude estimation is impacted by noise levels. If noise levels are high, some regions may be misinterpreted as an episode. To overcome this, it is recommended to use a fine time grid in the episode detection process to potentially smooth over these areas.

There may be some structural limitations that are not completely addressed by functional form selection, sampling rate, or noise level accommodations. One example may be the minimum amplitude necessary to be considered an episode or the number of data points necessary to confidently describe a region as an episode. For these limitations, adjustments can be incorporated into the objective function to ensure estimates are within desired limits.

Finally, it is important to consider the secondary models that will analyze the resulting estimates. These models must be appropriate for the hypotheses of interest, the study's sample size, and the data's underlying distributions.

CHAPTER 7

Future Work

7.1 Refinements and Extensions to the Markov Chain with Bootstrap Approach

Approach 1 focuses on marginal distributions of the mood states. That is, the treatment efficacy measures focus on each mood state individually. An alternative approach may extend the analysis to include the joint distribution of the mood states. This carries the benefit of preserving correlations between the states and may provide deeper insight of the mechanisms governing mood dynamics. It may also uncover specific scenarios where one therapy outperforms another, providing a refined treatment prescription. One way to investigate the joint mood state stationary distribution is to analyze the mood states as a three-dimensional measure. Because the proportions must sum to one, only three of the four mood states are needed to fully describe the stationary distribution. Data can be plotted in a three-dimensional space and metrics can be developed to evaluate differences in the multivariate distributions. Another possibility is to create a test statistic that evaluates the overall differences between two stationary distributions. While details of changes within specific mood states are lost in this omnibus measure, it provides the advantage of potentially detecting the effect of multiple minor changes occurring simultaneously.

One limitation to Approach 1 is that it does not adjust for covariates. For example, age and gender information are not incorporated into the analysis and differential treatment effects may be undetected. While analyses can be stratified by these variables, doing so would further reduce an already small sample size. It may be possible to automatically and efficiently incorporate other variables into the analysis in a unified manner, though additional assumptions may be necessary. Future developments may focus on adapting the approach to a multivariate setting.

Another limitation to Approach 1 is the implicit weighting of the data. In its present formulation, data are weighed according to the amount of data available for each patient. Patients with more data have longer mood state sequences and therefore provide greater influence on the transition matrix relative to patients with little data. While this may be appropriate in settings where patient-to-patient variability is acceptable, it may be problematic in highly heterogeneous populations. Depending on the research question of interest, a different weighting scheme may be more appropriate, such as a weight based on disease severity, type, or prevalence. Future work may focus on adaptations to adjust for heterogeneity.

To partially address data sparsity issues, Approach 1 uses linear interpolation to approximate intermediate points. However, other imputation methods are available. This may include a smoothing algorithm, the PSO fit in Approach 2, or models that incorporates other measures. In particular, the ChronoRecord and its correlations with HDRS and YMRS scores may provide a means of informative imputation. Future developments may focus the feasibility and effectiveness of model-based imputation methods.

Approach 1 focuses on the stationary distribution in characterizing treatment group differences. However, other measures are available within the Markov chain framework. One example are the transition matrices. These matrices directly describe the mood-shifting dynamics. Based on these matrices, approaches can be developed that evaluate patterns of episode onset or remission, switches between specific mood states, and the persistence of mood states. Furthermore, other measures derived from the transition matrices can form the basis of the analysis. For example, the mean recurrence time for the euthymic state may be used as a measure of mean episode duration. These measures may provide additional insight on the dynamics of the illness.

Future work can also focus on applying Approach 1 to other longitudinal assessment scales such as the ChronoRecord. With more frequent measures, mood-shifting dynamics can be described for shorter intervals and capture acute features of the illness. This is especially useful in rapid cycling bipolar disorder as episode rates are high and are better characterized by data collected over frequent time intervals.

7.2 Refinements and Extensions to the Particle Swarm Optimization Approach

Approach 2 focuses on a model-based approach to characterizing mood episode characteristics from the HDRS and YMRS data. The foundational component of Approach 2 is the mood function. Future research can focus on other functional forms that better characterize symptom severity trajectories or capture other features of interest. This can include within-episode fluctuations, partial remission, time trends in baseline severity, or the shape of the episode itself.

Refinements can also be made to the episode region detection method. This can involve the episode fitness function and adapting it to properly penalize poorly-fitting episodes or incorporating adjustments to capture unique episodes. For example, sensitivity for acute episodes may be increased by creating a weighting scheme that favors candidate episodes with high amplitudes and short durations. Similarly, the smoothing method used to fit the episode fitness values can be tuned to increase detection sensitivity. Other refinements may focus on the thresholding method used to distinguish true episodes from random fluctuations. Instead of using a fixed value, thresholds may vary according to time intervals. For example, a moving window approach would define a threshold based on neighboring regions rather than the entire curve and may detect episodes that differ from previous ones. The episode region detection method may also be refined by extending it into a multidimensional space that includes episode amplitude and duration. This would assist in not only identifying time intervals, but amplitude and duration intervals associated with those regions. This can help with the efficiency of the PSO algorithm by restricting the search space even further.

Much like the episode fitness function, the objective function used in the PSO algorithm may also be refined. The function used in Approach 2 uses the sums of square error and penalizes according to characteristics of the fitted episodes. Adjustments may be incorporated that examine episode regions independently from non-episode intervals, allowing a weighting scheme to be employed that can favor the fit in the regions of most interest. Other options can incorporate the candidate estimates themselves. For example, settings where estimates are believed to follow a specified distribution can include a likelihood component into the objective function. Similarly, the objective function can also be adapted to favor specific types of episodes. This can involve adjustments that create favorable objective values for candidate estimates that involve such episodes, or penalties for those that include none.

In Approach 2, the estimates derived from the PSO fit are used as inputs in regression models. Similarly, Approach 2 can be extended by using the entire curve as an input in a functional data analysis setting. By using the entire curve trajectory, patterns over time may be identified and reveal important temporal features of the illness. Additionally, the functional data analysis approach naturally extends to a multivariate setting, allowing the simultaneous analysis of both HDRS and YMRS data. Future work in this area would involve proper curve registration as it is unclear how to appropriately align highly variable episodic data on a unified time dimension.

A natural extension to Approach 2 is incorporating model fit within the PSO algorithm. This would require fitting data across patients simultaneously and allow the fit of one patient to be influenced that another. Recent research has incorporated the PSO algorithm into a nonlinear mixed effects setting[81], though the use of PSO and other swarm-based global search algorithms in statistical modeling is sparse.

Bibliography

1. D. A. Regier, J. K. Myers, M. Kramer, L. N. Robins, D. G. Blazer, R. L. Hough, W. W. Eaton, and B. Z. Locke. The NIMH Epidemiologic Catchment Area program. Historical context, major objectives, and study population characteristics. *Arch. Gen. Psychiatry*, 41(10):934–941, Oct 1984.

2. D. A. Regier, M. E. Farmer, D. S. Rae, B. Z. Locke, S. J. Keith, L. L. Judd, and F. K. Goodwin. Comorbidity of mental disorders with alcohol and other drug abuse. Results from the Epidemiologic Catchment Area (ECA) Study. *JAMA*, 264(19):2511–2518, Nov 1990.

3. R. C. Kessler, P. Berglund, O. Demler, R. Jin, K. R. Merikangas, and E. E. Walters. Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the National Comorbidity Survey Replication. *Arch. Gen. Psychiatry*, 62(6):593–602, Jun 2005.

4. K. R. Merikangas, R. Jin, J. P. He, R. C. Kessler, S. Lee, N. A. Sampson, M. C. Viana, L. H. Andrade, C. Hu, E. G. Karam, M. Ladea, M. E. Medina-Mora, Y. Ono, J. Posada-Villa, R. Sagar, J. E. Wells, and Z. Zarkov. Prevalence and correlates of bipolar spectrum disorder in the world mental health survey initiative. *Arch. Gen. Psychiatry*, 68(3):241–251, Mar 2011.

5. P. S. Wang, M. Lane, M. Olfson, H. A. Pincus, K. B. Wells, and R. C. Kessler. Twelve-month use of mental health services in the United States: results from the National Comorbidity Survey Replication. *Arch. Gen. Psychiatry*, 62(6):629–640, Jun 2005.

6. S. C. Dilsaver. An estimate of the minimum economic burden of bipolar I and II disorders in the United States: 2009. *J Affect Disord*, 129(1-3):79–83, Mar 2011.

7. *Diagnostic and statistical manual of mental disorders.* American Psychiatric Association, Washington, DC, 5th edition, 2013.

8. L. B. Alloy, S. Urosevic, L. Y. Abramson, S. Jager-Hyman, R. Nusslock, W. G. Whitehouse, and M. Hogan. Progression along the bipolar spectrum: a longitudinal study of

predictors of conversion from bipolar spectrum conditions to bipolar I and II disorders. *J Abnorm Psychol*, 121(1):16–27, Feb 2012.

9. D. L. Dunner and R. R. Fieve. Clinical factors in lithium carbonate prophylaxis failure. *Arch. Gen. Psychiatry*, 30(2):229–233, Feb 1974.

10. C. D. Schneck, D. J. Miklowitz, J. R. Calabrese, M. H. Allen, M. R. Thomas, S. R. Wisniewski, S. Miyahara, M. D. Shelton, T. A. Ketter, J. F. Goldberg, C. L. Bowden, and G. S. Sachs. Phenomenology of rapid-cycling bipolar disorder: data from the first 500 participants in the Systematic Treatment Enhancement Program. *Am J Psychiatry*, 161(10):1902–1908, Oct 2004.

11. S. Lee, A. Tsang, R. C. Kessler, R. Jin, N. Sampson, L. Andrade, E. G. Karam, M. E. Mora, K. Merikangas, Y. Nakane, D. G. Popovici, J. Posada-Villa, R. Sagar, J. E. Wells, Z. Zarkov, and M. Petukhova. Rapid-cycling bipolar disorder: cross-national community study. *Br J Psychiatry*, 196(3):217–225, Mar 2010.

12. R. W. Kupka, D. A. Luckenbaugh, R. M. Post, G. S. Leverich, and W. A. Nolen. Rapid and non-rapid cycling bipolar disorder: a meta-analysis of clinical studies. *J Clin Psychiatry*, 64(12):1483–1494, Dec 2003.

13. D. J. Miklowitz and S. L. Johnson. The psychopathology and treatment of bipolar disorder. *Annu Rev Clin Psychol*, 2:199–235, 2006.

14. R. J. Schloesser, J. Huang, P. S. Klein, and H. K. Manji. Cellular plasticity cascades in the pathophysiology and treatment of bipolar disorder. *Neuropsychopharmacology*, 33(1):110–133, Jan 2008.

15. H. K. Manji, J. A. Quiroz, J. L. Payne, J. Singh, B. P. Lopes, J. S. Viegas, and C. A. Zarate. The underlying neurobiology of bipolar disorder. *World Psychiatry*, 2(3):136–146, Oct 2003.

16. F. K. Goodwin and K. R. Jamison. *Manic-Depressive Illness: Bipolar Disorders and Recurrent Depression.* Oxford University Press, New York, 2007.

17. G. Hennemann, R. Docter, E. C. Friesema, M. de Jong, E. P. Krenning, and T. J. Visser. Plasma membrane transport of thyroid hormones and its role in thyroid hormone metabolism and bioavailability. *Endocr. Rev.*, 22(4):451–476, Aug 2001.

18. J. Bernal. Action of thyroid hormone in brain. *J. Endocrinol. Invest.*, 25(3):268–288, Mar 2002.

19. P. M. Yen. Physiological and molecular basis of thyroid hormone action. *Physiol. Rev.*, 81(3):1097–1142, Jul 2001.

20. P. Kopp. Thyroid hormone synthesis. In L. E. Braverman and R. D. Utiger, editors, *Werner and Ingbar's The Thyroid: A Fundamental and Clinical Text*. Lippincott Williams and Wilkins, Philadelphia, PA, 9th edition, 2005.

21. M. Bauer, T. Goetz, T. Glenn, and P. C. Whybrow. The thyroid-brain interaction in thyroid disorders and mood disorders. *J. Neuroendocrinol.*, 20(10):1101–1114, Oct 2008.

22. S. Chakrabarti. Thyroid functions and bipolar affective disorder. *J Thyroid Res*, 2011:306367, 2011.

23. M. Bauer, A. Heinz, and P. C. Whybrow. Thyroid hormones, serotonin and mood: of synergy and significance in the adult brain. *Mol. Psychiatry*, 7(2):140–156, 2002.

24. J. H. Hsu and G. A. Brent. Thyroid hormone receptor gene knockouts. *Trends Endocrinol. Metab.*, 9(3):103–112, Apr 1998.

25. C. Venero, A. Guadano-Ferraz, A. I. Herrero, K. Nordstrom, J. Manzano, G. M. de Escobar, J. Bernal, and B. Vennstrom. Anxiety, memory impairment, and locomotor dysfunction caused by a mutant thyroid hormone receptor alpha1 can be ameliorated by T3 treatment. *Genes Dev.*, 19(18):2152–2163, Sep 2005.

26. P. C. Whybrow, A. J. Prange, and C. R. Treadway. Mental changes accompanying thyroid gland dysfunction. A reappraisal using objective psychological measurement. *Arch. Gen. Psychiatry*, 20(1):48–63, Jan 1969.

27. R. T. Joffe and M. Marriott. Thyroid hormone levels and recurrence of major depression. *Am J Psychiatry*, 157(10):1689–1691, Oct 2000.

28. M. A. Frye, K. D. Denicoff, A. L. Bryan, E. E. Smith-Jackson, S. O. Ali, D. Luckenbaugh, G. S. Leverich, and R. M. Post. Association between lower serum free T4 and greater mood instability and depression in lithium-maintained bipolar patients. *Am J Psychiatry*, 156(12):1909–1914, Dec 1999.

29. O. Abulseoud, N. Sane, A. Cozzolino, L. Kiriakos, V. Mehra, M. Gitlin, S. Masseling, P. Whybrow, L. L. Altshuler, J. Mintz, and M. A. Frye. Free T4 index and clinical outcome in patients with depression. *J Affect Disord*, 100(1-3):271–277, Jun 2007.

30. M. Bauer and P. C. Whybrow. Thyroid hormones and the central nervous system in affective illness: interactions that may have clinical significance. *Integr. Psychiatry*, 6:75–100, 1988.

31. J. D. Amsterdam, A. Winokur, I. Lucki, S. Caroff, P. Snyder, and K. Rickels. A neuroendocrine test battery in bipolar patients and healthy subjects. *Arch. Gen. Psychiatry*, 40(5):515–521, May 1983.

32. A. Winokur, J. D. Amsterdam, J. Oler, J. Mendels, P. J. Snyder, S. N. Caroff, and D. J. Brunswick. Multiple hormonal responses to protirelin (TRH) in depressed patients. *Arch. Gen. Psychiatry*, 40(5):525–531, May 1983.

33. M. Bauer, H. Baur, A. Berghofer, A. Strohle, R. Hellweg, B. Muller-Oerlinghausen, and A. Baumgartner. Effects of supraphysiological thyroxine administration in healthy controls and patients with depressive disorders. *J Affect Disord*, 68(2-3):285–294, Apr 2002.

34. R. M. A. Hirschfeld, C. L. Bowden, M. J. Gitlin, P. E. Keck, T. Suppes, M. E. Thase, K. D. Wagner, and R. H. Perlis. *Practice Guideline for the Treatment of Patients With Bipolar Disorder*. American Psychiatric Association, Washington, DC, 2000.

35. University of South Florida: Florida best practice medication adult guidelines. *http://medicaidmentalhealth.org/Guidelines/Adult.aspx*, Oct 2011.

36. T. A. Ketter and P. O. Wang. Management of rapid-cycling bipolar disorders. In T. A. Ketter, editor, *Handbook of Diagnosis and Treatment of Bipolar Disorders*. American Psychiatric Publishing, Arlington, VA, 2010.

37. J. R. Calabrese, M. D. Shelton, C. L. Bowden, D. J. Rapport, T. Suppes, E. R. Shirley, S. E. Kimmel, and S. J. Caban. Bipolar rapid cycling: focus on depression as its hallmark. *J Clin Psychiatry*, 62 Suppl 14:34–41, 2001.

38. P. C. Whybrow. Sex differences in thyroid axis function: Relevance to affective disorder and its treatment. *Depression*, 3(1–2):33–42, 2005.

39. H. C. Stancer and E. Persad. Treatment of intractable rapid-cycling manic-depressive disorder with levothyroxine. Clinical observations. *Arch. Gen. Psychiatry*, 39(3):311–312, Mar 1982.

40. M. S. Bauer and P. C. Whybrow. The effect of changing thyroid function on cyclic affective illness in a human subject. *Am J Psychiatry*, 143(5):633–636, May 1986.

41. M. S. Bauer and P. C. Whybrow. Rapid cycling bipolar affective disorder. II. Treatment of refractory rapid cycling with high-dose levothyroxine: a preliminary study. *Arch. Gen. Psychiatry*, 47(5):435–440, May 1990.

42. M. S. Bauer, M. Aldi, T. Bschor, A. Heinz, N. Rasgon, M. Frye, H. Grunze, R. Kupka, and P. C. Whybrow. Clinical applications of levothyroxine in refractory mood disorders. *Clinical Approaches in Bipolar Disorders*, (2):49–56, 2003.

43. R. Aronson, H. J. Offman, R. T. Joffe, and C. D. Naylor. Triiodothyronine augmentation in the treatment of refractory depression. A meta-analysis. *Arch. Gen. Psychiatry*, 53(9):842–848, Sep 1996.

44. M. Bauer, R. Hellweg, K. J. Graf, and A. Baumgartner. Treatment of refractory depression with high-dose thyroxine. *Neuropsychopharmacology*, 18(6):444–455, Jun 1998.

45. L. L. Altshuler, M. Bauer, M. A. Frye, M. J. Gitlin, J. Mintz, M. P. Szuba, K. L. Leight, and P. C. Whybrow. Does thyroid supplementation accelerate tricyclic antidepressant response? A review and meta-analysis of the literature. *Am J Psychiatry*, 158(10):1617–1622, Oct 2001.

46. B. C. Appelhof, J. P. Brouwer, R. van Dyck, E. Fliers, W. J. Hoogendijk, J. Huyser, A. H. Schene, J. G. Tijssen, and W. M. Wiersinga. Triiodothyronine addition to paroxetine in the treatment of major depressive disorder. *J. Clin. Endocrinol. Metab.*, 89(12):6271–6276, Dec 2004.

47. C. J. Miller, S. L. Johnson, and L. Eisner. Assessment Tools for Adult Bipolar Disorder. *Clin Psychol (New York)*, 16(2):188–201, Jun 2009.

48. P. Bech. The use of rating scales in affective disorders. *Euro Psychiatr Rev*, 1:14–18, 2008.

49. T. A. Furukawa. Assessment of mood: guides for clinicians. *J Psychosom Res*, 68(6):581–589, Jun 2010.

50. M. Hamilton. A rating scale for depression. *J. Neurol. Neurosurg. Psychiatr.*, 23:56–62, Feb 1960.

51. American Psychiatric Association Task Force for the Handbook of Psychiatric Measures. *Handbook of Psychiatric Measures*. American Psychiatric Association, Washington, DC, 2000.

52. National Institute for Health and Clinical Excellence. Depression: the treatment and management of depression in adults. In *National Clinical Practice Guideline*, number 90. National Institute for Health and Clinical Excellence, London, 2009.

53. M. Zimmerman, I. Chelminski, and M. Posternak. A review of studies of the Hamilton depression rating scale in healthy controls: implications for the definition of remission in treatment studies of depression. *J. Nerv. Ment. Dis.*, 192(9):595–601, Sep 2004.

54. L. Kriston and A. von Wolff. Not as golden as standards should be: interpretation of the Hamilton Rating Scale for Depression. *J Affect Disord*, 128(1-2):175–177, Jan 2011.

55. R. C. Young, J. T. Biggs, V. E. Ziegler, and D. A. Meyer. A rating scale for mania: reliability, validity and sensitivity. *Br J Psychiatry*, 133:429–435, Nov 1978.

56. P. S. Masand, J. Eudicone, A. Pikalov, R. D. McQuade, R. N. Marcus, E. Vester-Blokland, and B. X. Carlson. Criteria for defining symptomatic and sustained remission in bipolar I disorder: a post-hoc analysis of a 26-week aripiprazole study (study CN138-010). *Psychopharmacol Bull*, 41(2):12–23, 2008.

57. S. Gopal, D. C. Steffens, M. L. Kramer, and M. K. Olsen. Symptomatic remission in patients with bipolar mania: results from a double-blind, placebo-controlled trial of risperidone monotherapy. *J Clin Psychiatry*, 66(8):1016–1020, Aug 2005.

58. M. Tohen, C. A. Zarate, J. Hennen, H. M. Khalsa, S. M. Strakowski, P. Gebre-Medhin, P. Salvatore, and R. J. Baldessarini. The McLean-Harvard First-Episode Mania Study: prediction of recovery and first recurrence. *Am J Psychiatry*, 160(12):2099–2107, Dec 2003.

59. P. D. Harvey. Defining and achieving recovery from bipolar disorder. *J Clin Psychiatry*, 67 Suppl 9:14–18, 2006.

60. M. S. Bauer, P. Crits-Christoph, W. A. Ball, E. Dewees, T. McAllister, P. Alahi, J. Cacciola, and P. C. Whybrow. Independent assessment of manic and depressive symptoms by self-rating. Scale characteristics and implications for the study of mania. *Arch. Gen. Psychiatry*, 48(9):807–812, Sep 1991.

61. P. C. Whybrow, P. Grof, L. Gyulai, N. Rasgon, T. Glenn, and M. Bauer. The electronic assessment of the longitudinal course of bipolar disorder: the ChronoRecord software. *Pharmacopsychiatry*, 36 Suppl 3:S244–249, Nov 2003.

62. M. Bauer, T. Wilson, K. Neuhaus, J. Sasse, A. Pfennig, U. Lewitzka, P. Grof, T. Glenn, N. Rasgon, T. Bschor, and P. C. Whybrow. Self-reporting software for bipolar disorder:

validation of ChronoRecord by patients with mania. *Psychiatry Res*, 159(3):359–366, Jun 2008.

63. J. T. Cho, S. Bone, D. L. Dunner, E. Colt, and R. R. Fieve. The effect of lithium treatment on thyroid function in patients with primary affective disorder. *Am J Psychiatry*, 136(1):115–116, Jan 1979.

64. R. W. Cowdry, T. A. Wehr, A. P. Zis, and F. K. Goodwin. Thyroid abnormalities associated with rapid-cycling bipolar illness. *Arch. Gen. Psychiatry*, 40(4):414–420, Apr 1983.

65. J. H. Lazarus. Lithium and thyroid. *Best Pract. Res. Clin. Endocrinol. Metab.*, 23(6):723–733, Dec 2009.

66. W. M. Tunbridge, D. C. Evered, R. Hall, D. Appleton, M. Brewis, F. Clark, J. G. Evans, E. Young, T. Bird, and P. A. Smith. The spectrum of thyroid disease in a community: the Whickham survey. *Clin. Endocrinol. (Oxf)*, 7(6):481–493, Dec 1977.

67. *Diagnostic and statistical manual of mental disorders.* American Psychiatric Association, Washington, DC, 3rd edition, 1980.

68. H.M. Taylor and S. Karlin. *An Introduction to Stochastic Modeling.* San Diego, California, 3rd edition, 1998.

69. C. A. Sugar, G. M. James, L. A. Lenert, and R. A. Rosenheck. Discrete state analysis for interpretation of data from clinical trials. *Med Care*, 42(2):183–196, Feb 2004.

70. B. Efron. Estimating the error rate of a prediction rule: Improvement on cross-validation. *Journal of the American Statistical Association*, 78(382):316–331, 1983.

71. C. M. Grinstead and J. L. Snell. *Introduction to Probability.* Providence, RI, 2nd edition, 1997.

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

73. M. R. Chernick. *Bootstrap Methods: A Guide for Practitioners and Researchers*. Hoboken, NJ, 2nd edition, 2007.

74. Nathaniel Schenker. Qualms about bootstrap confidence intervals. *Journal of the American Statistical Association*, 80(390):360–361, 1985.

75. Patrick Billingsley. Statistical methods in markov chains. *The Annals of Mathematical Statistics*, 32(1):12–40, 1961.

76. J. Kennedy and R. Eberhart. Particle swarm optimization. In *IEEE International Conference on Neural Networks*, volume 4, pages 1942–1948, 1995.

77. S. Lin and B. W. Kernighan. An effective heuristic algorithm for the traveling-salesman problem. *Operations Research*, 21(2):498–516, 1973.

78. Yuhui Shi and R. Eberhart. A modified particle swarm optimizer. In *IEEE World Congress on Computational Intelligence*, pages 69–73, 1998.

79. M. Clerc. Standard Particle Swarm Optimization. *http://clerc.maurice.free.fr/pso/SPSOdescriptions.pdf*, 2012.

80. T. J. Hastie and R. J. Tibshirani. *Generalized additive models*. Chapman & Hall, London, 1990.

81. S. Kim and L. Li. A novel global search algorithm for nonlinear mixed-effects models using particle swarm optimization. *J Pharmacokinet Pharmacodyn*, 38(4):471–495, Aug 2011.