**Protocol CSP  556**

**The Effectiveness of rTMS in Depressed VA Patients**

**STATISTICAL ANALYSIS PLAN**

**Prepared by:**

**VA CSPCC**
**CSPCC  151E, Research Bldg. 4**
**Boiler House Rd**
**PO BOX 1010**
**Perry Point, MD 21902**

## STATISTICAL ANALYSIS PLAN APPROVAL

| Responsibility | Name | Signature | Date (DDMMMYYYY) |
|---|---|---|---|
| Principal Investigaor | Jerome A. Yesavage MD | *signature* | 2/3/2017 |
| Co-chair | J. Kaci Fairchild Ph.D | *kaci fairchild* | 02.03.17 |
| Co-chair | Mark S. George MD | *signature* | 2/13/17 |
| Biostatistician | Zhibao Mi  MD Ph.D | *signature* | 2/13/17 |
| CSPCC Director | Kousick Biswas Ph.D | *signature* | 2/13/17 |
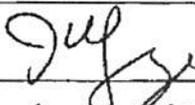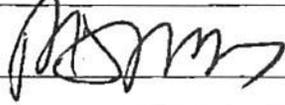
# TABLE OF CONTENTS

(Attachment: Table Shells)

Table 3.2.1 Subject Disposition (ITT)

Table 3.2.2 Listing of Early Terminated Subjects (ITT)

Table 3.2.3 Subject Disposition by Site and Treatment (ITT)

Table 3.3.1 Summary of Analysis Populations

Table 3.4.1 Treatment Exposure (ITT)

Table 3.4.2 Summary of Protocol Non-compliance (ITT)

Table 3.4.3 Summary of Treatment Compliance (ITT)

Table 3.5.1a Summary of Demographics (ITT)

Table 3.5.1b Summary of Demographics (Completers)

Table 3.5.1c Summary of Demographics (Fully Compliant)

Table 3.5.2a Summary of Baseline Assessment (ITT)

Table 3.5.2b Summary of Baseline Assessment (Completer)

Table 3.5.2c Summary of Baseline Assessment (Fully Compliant)

Table 3.5.3a Summary of Subjects with Significant Medical History (ITT)

Table 3.5.3b Summary of Subjects with Significant Medical History (Completer)

Table 3.5.3c Summary of Subjects with Significant Medical History (Full Compliant)

# LIST OF ABBREVIATIONS

| Abbreviation | Definition |
| --- | --- |
| ACLS | Advanced Cardiac Life Support |
| ACNP | American College of Neuropsychopharmacology |
| ADE | Adverse Device Effects |
| AE | Adverse Event |
| ALP | Alkaline Phosphatase |
| ALT/SGPT | Alanine Aminotransferase |
| AST/SGOT | Aspartate Aminotransferase |
| ATHF | Antidepressant Treatment History Form |
| BCLS | Basic Cardiac Life Support |
| BDI | Beck Depression Inventory |
| BHS | Beck Hopelessness Scale |
| BOMC | Blessed Memory Orientation Concentration Test |
| BSS | Beck Scale for Suicidal Ideation |
| BUN | Blood Urea Nitrogen |
| CAPS | Clinician Administered PTSD Scale |
| CBC | Complete Blood Count |
| CBOC | Community Based Outpatient Clinic |
| CFR | Code of Federal Regulations |
| COWA | Controlled Oral Word Association |
| CPR | Cardio Pulmonary Resuscitation |
| CPRS | Computerized Patient Record System |
| CRADO | Chief Research and Development Officer |
| CRF | Case Report Form |
| CRP | Clinical Research Pharmacist |
| CSP | Cooperative Studies Program |
| CSPCC | Cooperative Studies Program Coordinating Center |
| CSPCRPCC | Cooperative Studies Program Clinical Research Pharmacy Coordinating Center |
| CSSEC | Cooperative Studies Scientific Evaluation Committee |
| C-SSRS | Columbia – Suicide Severity Rating Scale |
| CT | Computerized Tomography |
| DAST | Drug Abuse Screening Test |
| DMC | Data Monitoring Committee |
| DSM-IV | Diagnostic and Statistical Manual of Mental Disorders |
| ECT | Electroconvulsive Therapy |
| EDC | Electronic Data Capture |
| EMG | Electromyogram |
| FDA | Food and Drug Administration |
| FWA | Federal Wide Assurance |
| GCP | Good Clinical Practice |

| | |
|---|---|
| GCPMG | Good Clinical Practice Monitoring Group |
| GCPSRG | Good Clinical Practice Standards and Resource Group |
| HIPAA | Health Insurance Portability and Accountability Act |
| HRC | Human Rights Committee |
| HRQL | Health-related Quality of Life |
| HRSD | Hamilton Rating Scale for Depression |
| HSS | Human Subjects Subcommittee |
| IRB | Institutional Review Board |
| ITTRS | Interactive Touch Tone Randomization System |
| JLO | Judgment of Line Orientation |
| LDH | Lifetime Drinking History |
| LDLPFC | Left Dorsolateral Prefrontal Cortex |
| LFT | Liver Function Test |
| LSC-R | Life Stressor Checklist - revised |
| MADRS | Montgomery-Asberg Depression Rating Scale |
| MAOIs | Monoamine Oxidase Inhibitors |
| MAST | Michigan Alcoholism Screening Test |
| MCS | Mental Component Summary |
| MDD | Major Depression Disorder |
| MIRECC | Mental Illness Research, Education and Clinical Center |
| MRI | Magnetic Resonance Imaging |
| MT | Motor Threshold |
| MUSC | Medical University of South Carolina |
| NAART | North American Adult Reading Test |
| NIH | National Institute of Health |
| NIMH | National Institute of Mental Health |
| NP | Nurse Practitioner |
| ORO | Office of Research Oversight |
| PCC | Pharmacy Coordinating Center |
| PCL | PTSD Checklist |
| PCS | Physical Component Summary |
| PDR | Physician's Desk Reference |
| PTSD | Post-Traumatic Stress Disorder |
| PV | Protocol Violation |
| QIDS-C | Quick Inventory of Depressive Symptomatology |
| QOL | Quality of Life |
| RAVLT | Rey Auditory Verbal Learning Test |
| RCT | Randomized Controlled Trial |
| R&D | Research & Development |
| rTMS | Repetitive Transcranial Magnetic Stimulation |
| SA | Substance Abuse |
| SACL | Substance Abuse Checklist |
| SAE | Serious Adverse Event |
| SC | Study Coordinator |
| SCID-I | Structured Clinical Interview for DSM-IV Axis I Disorders |

| | |
|---|---|
| SDMT | Symbol Digits Modalities Test |
| SI | Site Investigator |
| SMART | Site Monitoring, Auditing and Resource Team |
| SSRI | Selective Serotonin Reuptake Inhibitors |
| STAXI-2 | State-Trait Anger Expression Inventory -2 |
| TBI | Traumatic Brain Injury |
| THQ | Trauma History Questionnaire |
| TLFB | Alcohol / Drug of Choice Timeline Followback |
| TMT | Trail Making Test |
| TRMD | Treatment-Resistant Major Depression |
| TSH | Thyroid-Stimulating Hormone |
| UADE | Unanticipated Adverse Device Effects |
| ULN | Upper Limits of Normal |
| VA | Veterans Affairs |
| VR-36 | Veterans RAND 36 Item Health Survey |
| VISN | Veterans Integrated Service Network |

## 1.0   SUMMARY

This statistical analysis plan (SAP) is developed after review of the latest amended CSP-556 study protocol and case report forms (CRFs), but before any analysis of the data has begun. Detailed information is given to aid in the production of the statistical output and the statistical section of the final study report, and potentially manuscripts for publication. This document provides background of the study based on the protocol and describes the populations that will be analyzed. All subject characteristics and the efficacy and safety parameters that will be evaluated, along with the specific statistical methods, are described.

## 2.0   INTRODUCTION

### 2.1 Background

Major Depression (MD) is prevalent in about 10% of American medical outpatients in any given year (Kaplan and Sadock 1996).  Among these patients, as many as 20% respond incompletely or not at all to successive trials of multiple classes of antidepressant and mood stabilization medications, and of psychotherapy (Keller et al. 1992; Thase 2004).  Thus, within the VA population, there are roughly 100,000 patients with Treatment-Resistant Major Depression (TRMD). In such cases, the general treatment strategy is usually to advance treatment delivery in a way that increases response rates, albeit at the expense of increased risks and increased side effects.  One example would be the use of monoamine oxidase inhibitors (MAOIs). Another preferred treatment modality for TRMD is electroconvulsive therapy (ECT) (Anonymous 2002; Kaplan and Sadock 1996; Olfson et al. 1998).  However, despite being the most effective antidepressant in the acute setting, ECT usage is limited by post-treatment amnesia and confusion, the medical risks of general anesthesia, the high costs associated with inpatient hospitalization, general apprehension about the procedure among candidate patients, and some administrative impediments (Martin et al. 2003).  Such approaches may be reasonable for those depressed patients who are suicidal or who have the most severe symptoms.  However, for the majority of patients

with TRMD whose symptoms are more moderate, the decision to escalate treatment decisions is more difficult. Thus, new TRMD treatments are needed, preferably without major safety concerns or side effects as seen with aggressive polypharmacy or ECT. rTMS is a method of delivering brain stimulation without the seizures or risks associated with ECT, nor the potential side effects and risks of MAOI therapy.  It may offer a viable alternative to ECT.  Several studies have reported response of TRMD to rTMS (Avery et al. 1999; George et al. 1997; Loo et al. 1999).  Systematic review and meta-analysis of the studies to date, which are typically of a small scale, appear to show a positive effect in TRMD (Martin et al. 2003).  With a minimal side effect profile, and the rarity of untoward events and side-effects (Pascual-Leone et al. 1993; Wassermann 1997), safety concerns regarding the use of rTMS are considerably less than with ECT.  Importantly, rTMS is much less expensive to administer than ECT (largely due to not requiring anesthesia) (Kozel et al. 2004), and rTMS produces no detrimental cognitive side effects (Little et al. 2000; Triggs et al. 1999).  Thus, there is the potential for a significant advance in care, with associated cost savings, if rTMS were to be shown effective in treating TRMD in VA patients.

## 2.2 Objective

The primary objective is to assess the efficacy of rTMS in veterans to bring about remission of TRMD.

## 2.3 Study Design

This study is a phase II, randomized and balanced within each stratum, double-blind, sham-controlled, intent-to-treat, 2-arm, parallel design, multicenter study in the United States.  Three hundred and sixty veterans diagnosed with TRMD are targeted to be  enrolled at 9 VA Medical Centers over a three year period.  Participants are randomized into a double blind clinical trial to left prefrontal rTMS treatment or to sham (control) rTMS treatment (1:1 ratio) for up to 30 treatment sessions.  All participants are evaluated on a wide variety of measures including cognitive, neurological and functional parameters.  All will meet DSM-IV criteria for Major Depression and all will

have failed at least two prior pharmacological interventions as defined by the Antidepressant Treatment History Form (ATHF) (Sackeim et al. 1990), i.e., they are TRMD patients.  Veterans with PTSD or history of substance abuse are not excluded but detailed histories regarding these disorders are obtained.  Participants are not required to stop using anti-depressant medication.  The primary dependent measure is remission rate of Hamilton Rating Scale for Depression (HRSD24 ≤ 10), and secondary analyses will be conducted on other depression and neuropsychological indices.  Comparisons between the rTMS (active) and the sham groups will be made at the end of the acute treatment phase to test the primary hypothesis. The overview of the experimental design and procedures is presented in Appendix 5.1.

## 2.4 Study Outcome Variables

### 2.4.1 Primary Outcome Variable

The primary outcome is a proportion of participants achieving remission from depression based on the HRSD24 ≤ 10 at the end of the acute treatment phase.

### 2.4.2  Secondary Outcome Variables

There are five secondary outcome measures based on their use in previous studies in major depression, where appropriate.

1.  Depression measured by Montgomery-Asberg Depression Rating Scales (MADRS)
2.  Suicide Ideation measured by Beck Scale for Suicide Ideation (BSS)
3.  Depression measured by Beck Depression Inventory (BDI)
4.  Quality of Life measured by the VR-36
5.  Cognitive Function as measured by a neuropsychological battery

### 2.4.3.  Safety Outcome Variables

The following safety measures are performed during the course of the study: physical examinations, AEs/SAEs, clinical evaluation triggers, and psychological assessments.

## 2.5 Study Assessments Used in the Analysis

### 2.5.1. Basic Information

Participants' demographics, medical history, physical exam, laboratory, toxicology, pregnancy test and medication use are collected.

*2.5.1.1 Demographics/Medical History/Laboratory/Toxicology/Medication Use.*

Relevant demographics are collected as to age, gender, racial/ethnic grouping, military history and income. In addition, a standard medical history and physical examination are completed as well as laboratory tests including a complete blood count (CBC), electrolytes (chemistry), thyroid panel and a liver function test. An alcohol test and urine drug toxicology screen are conducted prior to randomization and also randomly during the following time points of study participation:  a) acute treatment phase: 2nd, 4th, and 6th (if still in acute treatment) weeks ; b) taper phase: 2nd week; c) follow-up phase: 1st, 3rd, and 5th months.

*2.5.1.2 Physical Exam*

A physical exam of the oral cavity, head, eyes, ears nose and throat, cardiovascular system, lungs, abdomen (liver/spleen), extremities, skin, neuropsychiatric mental status and sensory/motor status, musculoskeletal system and general appearance will be performed during the screening/baseline phase and at week 9.  Height will be recorded during the screening/baseline phase only.

*2.5.1.3 Pregnancy Test*

A pregnancy test is conducted on all female patients of childbearing potential (that is, all women except for those who are post menopause for > 2 years or who have a history of hysterectomy or surgical sterilization) prior to randomization and every four weeks during the study.  Information on medication use (prescription, natural food products, and "over the counter") will be collected at screening and updated after each block of five sessions during the treatment phase and every four weeks during the follow-up phase.

*2.5.1.4. Antidepressant Treatment History Form (ATHF)*

The ATHF provides a uniform and rigorous method of eliciting and recording a patient's past experience using antidepressant medications (Sackeim et al., 1990). The

ATHF provides detailed information about which treatments the patient has received during the index episode and over his/her lifetime. Specific criteria are used to evaluate the adequacy (e.g., dose and duration) of each treatment trial, and a determination is made, for each trial, whether the patient manifested treatment resistance (did not satisfactorily respond at adequate dose and duration) or whether the stringent dose and duration requirements could not be met (treatment intolerance).

### 2.5.1.5  Concomitant medications

All medications taken by the subject during the screening/baseline phase, during the treatment phase and during the follow-up phase must be pre-approved by the site PI or his/her designee whenever possible to avoid interactions with the study treatment.

### 2.5.2. Measures of Substance Abuse and Post-traumatic Stress

To perform moderator analyses determining if these comorbid conditions are associated with differential response to treatment, relevant measures will be collected. Since all subjects, at the time of the protocol, will not be abusing substances, the most relevant measures will be history of duration and severity of substance abuse, in particular alcohol abuse.  Additional measures will quantify relevant aspects of PTSD.

### 2.5.2.1. Lifetime Drinking History

Lifetime alcohol consumption will be assessed using the Lifetime Drinking History (LDH) instrument as designed by Skinner and Sheu (1982) and refined by Sobell and colleagues (1988, 1990).  LDH is the state-of-the-art validated assessment instrument for obtaining quantitative data on the frequency, amount, duration, and pattern of lifetime alcohol consumption beginning from the onset of regular drinking. Aggregate indices for total lifetime drinking can be assessed with moderate to high reliability (Skinner & Sheu,1982;  Sobell et al., 1990, 1988).  It is recognized that the pattern of drinking behavior (i.e. chronic regular drinking vs. binge drinking) may affect outcome measures.  Therefore, drinking assessment will include measures of total alcohol consumption, typical and maximum alcohol consumption per occasion, average daily and average monthly intake (measures of drinking intensity reflecting both frequency of drinking occasions and dose per occasion), both for the last six

months (current drinking) and for lifetime (lifetime drinking history).

*2.5.2.2. Michigan Alcohol Screening Test (MAST)*

The MAST is self-report measure for the detection of alcoholism. It consists of 25 yes-no questions that are differentially weighted depending on the severity of the symptom addressed in each item. The score ranges for interpretation of the MAST are as follows: 0-4 = absence of alcoholism; 5-6 = possible alcoholism; 7 and up = probable alcoholism. The measure will be used to assess alcohol abuse at baseline and then at the end of acute treatment (intervention) and the follow-up phase.

*2.5.2.3. Drug Abuse Screening Test (DAST)*

The DAST is a self-report measure for the detection of drug abuse or dependence on a range of psychoactive substances, other than alcohol. The DAST was adapted from the MAST and shares a similar item structure (Skinner, 1982). A score of 5 or higher is indicative of a possible drug use disorder.

*2.5.2.4. Structured Clinical Interview for DSM-IV (SCID)*

The Structured Clinical Interview for DSM-IV Axis I Disorders (SCID) is a semi-structured interview that is used to make the major DSM-IV Axis I diagnoses in the most uniform manner possible (First et al., 2002).

*2.5.2.5. Clinician Administered PTSD Scale (CAPS)*

CAPS will determine lifetime and current PTSD. The CAPS measures frequency and intensity of PTSD-related symptoms (Blake et al., 1995). Possible scores range from 0 to 136. In a recent review of studies utilizing the CAPS, (Weathers et al., 2001) propose the following severity score ranges for interpreting the CAPS, which are as follows: 0–19 = Asymptomatic/few symptoms; 20-39 = Mild PTSD/subthreshold; 40-59 = Moderate PTSD/threshold; 60-79 = Severe PTSD symptomatology; $\geq$ 80 = Extreme PTSD. Using these recommendations, in the proposed study, PTSD positive subjects will be positive for lifetime PTSD, related to any lifetime traumatic experience, and will meet CAPS criteria for current, chronic PTSD if they have a current CAPS score of > 40; PTSD negative subjects will be those with a current CAPS score of < 20.

*2.5.2.6. Trauma History Questionnaire (THQ)*

This is a 24-item self-report inventory which has been modified to provide data

on childhood trauma such as sexual or physical assault.  This scale is used to determine the presence or absence of childhood sexual or physical abuse prior to age 13 and to better characterize the trauma histories of our participants.  Responses to items 18-23 focus specifically on the age of occurrence of sexual and physical assault. The Trauma History Questionnaire has been shown to have good test-retest stability (Green, 1996).

### 2.5.2.7. Life Stressor Checklist-Revised (LSC-R)

This is a 30-item structured clinical interview for lifetime exposure to stressful life events.  The scale emphasizes a number of different potentially traumatic events and assesses the participant's emotional reaction to the stressors and the time period in which the stressors occurred.  The LSC-R is reported to have sound psychometric qualities within various PTSD populations (Wolfe & Kimerling, 1997).  This measure along with the Trauma History Questionnaire will be used to assess the trauma histories of our participants.

### 2.5.2.8. PTSD Checklist (PCL-M)

This is a 17 item self-report.  Its limitation is that it is keyed to a single traumatic event; e.g., the patient's worst experience in the military (Blanchard et al., 1996). Nonetheless, it has adequate reliability and has been shown to correlate well with scores from the CAPS.  It will be used for follow-up of symptoms to assess change with treatment.

### 2.5.2.9. Blessed Orientation Memory Concentration Test (BOMC)

The BOMC (Katzman, Brown, Fuld, Peck, Schechter, & Schimmel, 1983) is a 6-item screening measure of cognitive impairment which takes approximately 5 minutes to administer.   The measure consists of 3 orientation questions, listing months backwards, a name and address memory phrase, and counting backwards from 20 to 1. This measure is fairly sensitive to milder levels of cognitive impairment.

### 2.5.3. Efficacy Assessment

### 2.5.3.1. Hamilton Rating Scale for Depression (HRSD)

This measure is the primary outcome measure and is completed after each block of 5 sessions throughout the study.  The HRSD is the "gold standard" of

randomized clinical trials for depression, and the primary measure of most rTMS studies to date.  This study will utilize the 24-item version of this instrument (HRSD24) to evaluate depressed mood, vegetative and cognitive symptoms of depression, and comorbid anxiety symptoms.   It provides ratings on current DSM-IV symptoms of depression, with the exceptions of hypersomnia, increased appetite, and concentration/indecision. The Planning Committee, in making this recommendation, examined the outcome measures used in current NIH and industry trials as well as the ACNP criteria for "remission" (Rush et al., 2006).  The ACNP recommends that outcome measures reflect all major criteria used in the diagnosis of MDD.  Neither the HRSD or the MADRS include all criteria however, they are widely used in clinical trials and the ACNP report notes that the field has not followed this recommendation and that if one were to use these measures for reasons of comparability (as this Planning Committee recommends), the ACNP report suggests the use of other metrics to assure that remission is complete.   For that reason, the Planning Committee recommended additional use of the Quick Inventory of Depressive Symptomatology (QIDS-C16), which meets the ACNP criteria. Certification of Ratings of HRSD:  We plan to follow the NIH protocol procedures for administration and certification of the HRSD ratings.  This will include the use of a prepared script to help administer the HRSD.   Certification of all raters at a participating site will be verified prior to enrollment.  This will be done by shipping recordings of mock interviews (non-patient) to the sites where trained raters have determined a "gold standard HRSD score".  Site raters will then submit their scores. Following NIH procedures, large deviations will be noted, and a rater can have an additional test.  This can be repeated for a total of 3 times until the site is told they must find another rater. Longitudinal Quality Control for HRSD:  Following NIH procedures, to ensure that HRSD do not "drift" over time, one HRSD recording will be circulated to evaluators at all participating sites every 6 months.  The evaluators will be asked to rate this recording and to return their ratings. Evaluators who drift greater than 3 points on the HRSD total score will receive telephone consultation followed by one additional HRSD recording.

### 2.5.3.2. Montgomery Asberg Depression Rating Scale (MADRS)

As another measure of depression, the Montgomery-Asberg Depression Rating Scale (MADRS) has been used with increasing frequency in recent years to measure outcome in antidepressant efficacy trials (Montgomery et al 1979). It offers an alternative view of depressive illness, and may be sensitive to depressive symptoms that are not easily captured in the context of the HRSD, such as hypersomnia, increased appetite, and concentration/indecision. The MADRS is a 10-item clinician rating of depressive symptoms. Each item is scored on a 7-point scale (0 to 6) (range 0–60). Anchors are provided for even numbered scale points. Higher scores represent higher levels of depression.

### 2.5.3.3. Beck Depression Inventory (BDI)

This measure is a 21-item self-report test presented in a multiple choice format which measures presence and extent of depression. Each of the 21 items addresses a specific symptom or attitude that pertains to depressed patients, and which are consistent with descriptions of the depression within the peer-reviewed literature. While generally deemed less reliable than scales score by a trained rater (for example, the HRSD), the Beck scale is easy to administer, and provides convenient means by which patients can effectively communicate their own perception of their mood state. When the test is scored, a value of 0 to 3 is assigned for each answer and then the total score is compared to a key to determine the depression's severity.

### 2.5.3.4. Beck Scale for Suicide Ideation (BSS)

To help clinicians screen psychiatric patients for suicidal ideation, the Beck Scale for Suicide Ideation (Beck and Steer, 1991) was developed, and is herein referred to as the BSS. This self-report measure consists of 21 items and is one of the most thorough assessments of both active and passive suicidal ideation. Respondents are asked to rate the severity of each item on a 3-point scale with scores ranging from 0 to 2. The first five items on the BSS are regarded as a screener for suicidal ideation and assess one's desire to live, desire to die, reasons to live and reasons to die, and suicidal ideation. The remainder of the BSS assesses the duration and frequency of suicidal ideation, ambivalence regarding suicidal ideation, reasons for living /

deterrents for suicide, suicide plan / opportunity to enact plan, expectations for following through after an attempt, preparations that have been made to ready for a suicide, past suicide attempts, and wish to die during past suicide attempts.  It should be noted that the psychiatrist responsible for the assessment of the patient is responsible for performing a more detailed assessment of any patient showing an increase in BSS score.  The BSS will be given at screening, weekly during acute treatment and taper phases and then monthly during the follow-up phase.

Suicidal ideation, hopelessness, agitation, aggression, and depressive symptoms can also be derived from the HRSD and the MADRS, which is also being rated on each patient.  The advantages of the HRSD and the MADRS are that they are interviewer-scored scales that focus not only on ideation, but also on psychotic symptoms, and somatic symptoms.  Thus, this study, in contrast to the other multisite trials of rTMS, will not only be different because of the population (veterans with TRMD) that it treats, but also because it incorporates five scales that enable evaluation of both ideation and intent (CSSRS and BSS), hopelessness (BHS), and mood, psychotic, and somatic symptoms (HRSD, MADRS).

*2.5.3.5.  Health Services: Veterans RAND 36 Item Health Survey (VR-36)*

The VR-36 (formerly known as the SF-36V) is a self-administered survey that measures eight dimensions of health: physical functioning, role limitations due to physical health, bodily pain, general health perceptions, vitality, social functioning, role limitations due to emotional problems, and mental health. It yields scale scores for each of these eight health domains, and two summary measures of physical and mental health: the Physical Component Summary (PCS) and Mental Component Summary (MCS). The VR-36 includes two additional items that assess how much physical and emotional health has changed over the previous year. The VR-36 is a generic measure, as opposed to one that targets a specific age, disease, or treatment group. Thus, it has been useful in assessing the health of general and specific populations, comparing the relative burden of diseases, differentiating the health benefits produced by a wide range of treatments, and screening individual patients. The applicability of the VR-36 is apparent from its widespread use.

### 2.5.3.6. *Neuropsychological Battery*

We propose to use a battery that is sensitive to the potential cognitive effects of rTMS. rTMS may improve cognitive function as depression is lifted, or it could have the potential for impairing function. A battery has been designed to be sensitive to such potential effects and has been used in previous studies of the effectiveness of rTMS. We propose to use measures that have been used in previous studies of rTMS as they have proven to be sensitive and it also provides a basis for comparison of the VA patients entering this study with patients who have participated in other studies of rTMS. The cognitive assessments will be administered at screening, at the end of the acute treatment phase and at the end of the 24 week follow-up phase. These measures are widely used in the literature and have been shown to be effective when working with severely depressed patients Testing includes measures of Executive Function, Attention, Memory, Visuospatial Ability, Processing Speed, Psychomotor Function, and Premorbid Intelligence. Executive functioning is assessed using the Controlled Oral Word Association (COWA) Test, which is a test of verbal fluency. Participants are asked to produce as many words that begin with a specific letter (F, A, or S) as they can within one minute. The participant is then asked to name as many animal names as possible within one minute (Spreen and Strauss, 2006). Attention is assessed using the Stroop Color and Word test (Golden, 1978). This measure consists of three pages: a Word page with 100 color words printed in black ink; a Color page with 100 X's printed in colored ink; and a Color-Word page that contains words from the Word page printed in colors (the word and the color do not match). Participants are asked to read as many words or name as many colors as possible in 45 seconds. The Rey Auditory-Verbal Learning Test (RAVLT; Rey, 1964; Lezak et al., 2004) is used to assess verbal learning and memory. The measure consists of 15 nouns that are read aloud for five trials. After each trial, the participant is asked to recall as many words as they can from the list. Another list of words, an interference list, is read after the fifth trial and the participant is asked to recall the words from that list. Immediately after that recall, the participant is asked to recall as many words from the original list of 15 nouns. This is then followed by a 20 minute delay, during which

other measures of the cognitive assessment will be administered. The participant is asked to recall the original list of 15 words after this 20 minute delay. Finally the participant is asked to identify the original 15 words after being a read a story that contains all of the original 15 words. The Judgment of Line Orientation (JLO; Benton et al., 1994) is used to assess visuospatial ability. There are two alternate forms that each consist of 30 items with an additional 5 practice items. Items are presented in a spiral bound booklet with stimuli appearing in the upper part of the booklet and the multiple choice card appearing in the lower part. The participant is asked to indicate on the multiple choice card the lines that match the direction of the lines on the stimulus card. Processing speed is assessed using the Symbol Digit Modalities Test (SDMT; Smith, 1991). Participants are presented with a coding key consisting of nine abstract symbols. They must scan the coding key and record the corresponding number as quickly as possible. The participants are given 90 seconds to complete the task. Psychomotor functioning is assessed using the Trail Making Test: Parts A and B (TMT; Reitan & Wolfson, 1993). In Part A, the participant is asked to connect, in order, 25 encircled numbers that are dispersed randomly on a page. The participant is then asked to connect 25 encircled numbers and letters in an alternating order in Part B. Both Part A and Part B include practice exercises to ensure the participant understands the nature of the task and all tasks are timed. The North American National Adult Reading Test (NAART; Blair & Spreen, 1989) is used as an estimate of premorbid intellectual functioning. This measure consists of 61 items that are presented in two columns on a page for the participant to read. Participants are asked to read each word aloud as the examiner marks the errors on a score card.

All of the measures in the cognitive assessment are paper and pencil measures that will be administered by research staff. Staff will be trained in the proper administration and scoring of the cognitive assessment. The cognitive assessment is expected to last approximately one hour at screening as well as at each follow-up.

### 2.5.3.7. *Quick Inventory of Depressive Symptomatology (QIDS-C16)*

The ACNP recommends that outcome measures reflect all major criteria used in the diagnosis of MDD. For that reason the Planning Committee recommended

additional use of the Quick Inventory of Depressive Symptomatology (QIDS-C16), which meets the ACNP criteria (Rush et al., 2003). The HRSD does not measure hypersomnia, weight gain or problems with concentration or decision making.

### 2.5.3.8. Beck Hopelessness Scale (BHS)

The Beck Hopelessness Scale is a self-report measure consisting of 20 "yes/no" items (Beck et al., 1974). A total severity of hopelessness is calculated from summing the 20 items and guidelines for interpretation for scores are as follows: 0-3 = minimal hopelessness; 4-8 = mild hopelessness; 9-14 = moderate hopelessness; and 15-20 = severe hopelessness. The BHS is given at screening, weekly during acute treatment and taper phases and then monthly during the follow-up phase.

## 2.5.4. Safety Assessment and Triggers

### 2.5.4.1. Columbia – Suicide Severity Rating Scale (C-SSRS)

Suicide is a rare event. As such, suicide rates cannot be used as an outcome measure for an rTMS study. Similarly, the study is of too short a duration to expect to find a significant difference in numbers or lethality of suicide attempts between treatment and placebo (sham rTMS) groups. Nonetheless, there are two areas that can be expected to change with successful rTMS treatment: preoccupation with suicidal ideations or plans. Because this study uses both a lead in period prior to treatment and a sham rTMS treatment group, we are able to compare the rate of parasuicidal behavior in these patients, who are at more serious risk of a suicide completion. One of the newer methods of monitoring patients at risk for suicide is the C-SSRS. The C-SSRS assesses suicidal ideation as well suicidal behavior over a specified time period and is frequently employed by the Food and Drug Administration in research to determine if suicidality is an adverse effect. The form is collected at multiple time points in during the course of the study. Initially, it is collected at baseline to serve as a screener for persons reporting suicidal ideation or behaviors in the past six months. It is also completed weekly during acute treatment and then monthly during the follow-up phase to monitor for the presence of suicidal ideation or behaviors. The sensitivity of this instrument will allow us to identify even "minor" suicide "gestures" as well as more serious attempts.

### 2.5.4.2.  State-Trait Anger Expression Inventory -2 (STAXI-2)

The STAXI-2 is a 57-item self-report measure of the experience, expression, and control of anger.  The measure consists of six scale, five subscales, and Anger Expression Index which provides a measure of total anger expression.  The STAXI-2 is frequently included in risk assessments for violence and used as such in the current protocol.

### 2.5.4.3.  Adverse events (AEs)

An AE is any untoward medical occurrence in a patient or clinical investigation subject administered a pharmaceutical product and that does not necessarily have a causal relationship with this treatment. AEs will be identified and documented on the AE CRF in appropriate medical terminology.   The severity of the AE and the relationship to the study medication will be determined and reported on the CRF (see below). The severity of each AE will be characterized and then classified into one of three clearly defined categories as follows:

- Mild - the AE did interfere in a significant manner with the subject's normal functioning level.  It may have been an annoyance.
- Moderate - the AE produced some impairment of functioning, but was not hazardous to health.  It was uncomfortable or an embarrassment.
- Severe - the AE produced significant impairment of functioning or incapacitation and was a definite hazard to the subject's health.

These three categories are based on the investigator's clinical judgment, which, in turn, depends on consideration of various factors such as the subject's report and the PI's observations and/or prior experience. The relationship of the AE to the study medication should be specified by the investigator, using the following definitions:

- Not Related: Concomitant illness, accident, or event with no reasonable association with treatment.
- Unlikely:  The reaction has little or no temporal sequence from administration  of the study drug, and/or a more likely alternative etiology exists.
- Possibly Related: The reaction follows a reasonable temporal sequence from administration of the drug and follows a known response pattern to the

suspected drug; the reaction could have been produced by the study drug or could have been produced by the subject's clinical state or by other modes of therapy administered to the subject.

- Probably Related: The reaction follows a reasonable temporal sequence from administration of study drug; is confirmed by discontinuation of the study drug or by rechallenge; and cannot be reasonably explained by the known characteristics of the subject's clinical state.

- Unknown: Use only if the cause is not possible to determine.

### 2.5.5. Control Questionnaire

A questionnaire will be used before and after the first treatment session, and at the end of the final study visit to elicit patient perception of whether they were on active or sham rTMS treatment.

## 2.6 Sample Size Consideration

The primary hypothesis of the study is that in VA patients with TRMD, rTMS will result in a greater remission rate at the end of acute treatment than sham rTMS. The primary outcome measure in this study is success or failure to achieve remission from depression as defined by a score on the Hamilton Rating Scale for Depression (HRSD24) of 10 or less. The primary analysis will be done as an "intent-to-treat" analysis, i.e. patients will be analyzed in the groups to which they were randomized and drop-outs will be considered treatment failures. The primary hypothesis will be addressed using a logistic regression model with PTSD diagnosis, history of substance abuse, and site as covariates. Based on review of the studies reported in the Introduction Section, the Planning Committee felt that a 10% difference between treatments would be of clinical relevance given the severity of the illness. With a sample size of 180 per group, the proposed study will have a power of 81% to detect an absolute difference between groups of 10% in the percentage of those participants who remit (6% sham and 16% rTMS). Thus, a total of 360 patients are required. This goal of 360 patients, larger than any previous study, also provides a measure of protection should some assumptions be wrong.

# 3.0   STATISTICAL METHODS

## 3.1 Statistical Handling Policy

### 3.1.1 Missing Data and Imputations

For subjects who drop out during the treatment phase will be considered as treatment failures for the purpose of the primary analysis and the missing values will not be imputed. Multiple imputation (MI) method may be used for certain secondary analyses. Multiple imputations will be based on Rubin's procedure (Rubin 1987) using SAS PROC MI and PROC MIANALYZE if needed.   Sensitivity analysis will be performed to compare the results from the imputed data based on the two methods, and the complete data without imputation.

### 3.1.2 Analysis Conventions

This section details general policies to be used for the statistical analyses. Departures from these general policies may be given in the specific detailed sections of this statistical analysis plan. When this situation occurs, the rules set forth in the specific section take precedence over the general policies. The following policies will be applied to all data presentations and analyses.

- All statistical tests will use a significance level of $\alpha = 0.05$. Two-tailed tests will be performed for all analyses that use statistical testing.

- All p-values will be rounded to 3 decimal places. All p-values that round to 0.000 will be presented as '<0.001' and p-values that round to 1.000 will be presented as '>0.999'. Any p-value $\leq \alpha$ will be considered statistically significant.

- Summary statistics will consist of the number and percentage of responses in each category for discrete variables, and the mean, median, standard deviation (SD), minimum, and maximum for continuous variables.

- All mean and median values will be formatted to one more decimal place than the measured value. Standard deviation values will be formatted to two more decimal places than the measured value.

- All percentages will be rounded to one decimal place. The number and percentage of responses will be presented in the form XX (XX.X), where the

percentage is in the parentheses. The decimal of the percentage may be dropped due to space constraints when creating a table.

- For a continuous variable, if it is normally distributed or approximately normally distributed, Student t test and/or the general linear regression will be performed to test the group difference as proposed in the analysis plan, otherwise, Wilcoxon test and/or a quantile regression will be used an alternative analysis method; for a categorical variable, Pearson $\chi^2$ test and/or a corresponding generalized linear regression will be performd to test the group difference.

- All listings will be sorted for presentation in order of treatment group, site number, subject number, and date of procedure or event.

- All analysis and summary tables will have the population sample size for each treatment group in the column heading.

- Calculating change from baseline to a visit will be done as follows: change = visit –baseline.

- Baseline is defined as the last data point before the first treatment is administered. If baseline data are not available, screening data will be used.

- Baseline measurements and other important covariates will be adjusted if they are not balanced between the two treatment groups unless they are specified in the particular analyses.

- Version 9.4 of SAS® or higher will be the statistical software package used to produce all summaries, listings, statistical analyses, and graphs.

- Updated version of MedDRA will be used for adverse event and pre-treatment coding.

- The current version of the World Health Organization (WHO) drug dictionary will be used for the coding of medications.

## 3.2 Subject Disposition

Subject disposition will be summarized for the ITT population. The following data will be presented:

- The number and percentage of subjects who completed or discontinued prematurely from the study by treatment group. The number and percentage of subjects who discontinued for each reason will be presented for each treatment group (Table 3.2.1). The number and percentage of subjects who completed or discontinued prematurely in each treatment group will also be displayed graphically (Figure 3.2.1).

- A listing of subjects that discontinued prematurely from the study. The listing will include information about treatment, study center, subject number, age, race, number of sessions on treatment, and reason for discontinuation (Table 3.2.2), and the percentages of the completion and discontinuation by site will be presented (Table 3.2.3).

  The number of subjects that were enrolled at each study center and the number and percentage of subjects that completed or discontinued at each study center will be summarized for each treatment group and for all subjects.

  The end of trial CRF will be used to determine who discontinued prematurely from the study.

## 3.3 Analysis Populations

**Intent-to-Treat (ITT) –** This population includes all subjects who were randomized to the study. Subjects will be assigned for analysis according to the group to which they were randomized.

**Completers –** This population includes all ITT subjects who patients who were treated according to the protocol and had fewer than 4 TMS sessions  not completed during acute treatment phase (George et al 2010).

**Fully Compliant –** This population includes all ITT subjects who had fewer than 2 sessions not completed and had no other treatment related protocol violations during acute treatment phase (George et al 2010).

**Safety –** This population includes all subjects who received at least one session of the rTMS treatment. The number of subjects in each population will be summarized for each treatment group and for all subjects (Table 3.3.1).

## 3.4 Study Treatment Session and Compliance

Study rTMS treatment session and compliance will be summarized by treatment group and overall for the ITT population.

### 3.4.1 Study Treatment Session

The following information will be presented for each treatment group and overall (Table 3.4.1):

- Summary statistics (number of subjects with data, mean, median, SD, minimum, and maximum) for the number of sessions of rTMS treatment.
- The number and percentages of subjects in each of the following categories of the acute treatment: 0-5, 6-10, 11-15, 16-19, $\geq$20 sessions.

### 3.4.2 Study Treatment Compliance

The following information will be presented for each treatment group and overall for compliance:

- Summary of protocol noncompliance (Table 3.4.2)
- Summary statistics (number of subjects, mean, median, SD, minimum, and maximum) for treatment compliance (Table 3.4.3)
- The number and percentage of subjects in each of the following categories of compliance: ≤70%, >70%-80%, >80%-90%, and >90% (Table 3.4.3)

## 3.5 Subject Demographics and Pre-Treatment Characteristics

Subject demographics and pre-treatment characteristics will be summarized for the Intent-to-Treat population. The demographics and pre-treatment characteristics will be summarized for each treatment group and for all subjects.

### 3.5.1 Demographics

The summary of demographics at baseline will include:

- The number and percentage of subjects with each category of race, gender, ethnicity, marital status, education, military history, location of military service, branch of service, work history, type of job, type of payment, income, income past four weeks, and pregnancy test results.

- The sample size, mean, median, SD, minimum, and maximum values for age, weight, and height. Age will be calculated as follows:
  - Age = Largest Integer ≤ [(Screening Visit Date – Date of Birth +1)/365.25]
  - Weight at screening from physical examination CRF will be used for the summary
  - BMI = (Weight/Height$^2$)*100$^2$. The unit of weight will be kilogram and the unit of height will be centimeter.

Summary of the demographics characteristics of the subject population will be presented for ITT, completer, and full compliance populations (Tables 3.5.1a-c), as well as other chosen subgroups which may be presented if the populations are very different from ITT. Significance tests of the balances between the two treatment groups will be performed and p-values will be presented.

### 3.5.2 Outcome Measures at Baseline

Baseline comparability among the treatment groups will be evaluated with respect to baseline values of outcome measures besides the demographic variables. The sample size, mean, median, SD, minimum, and maximum values for the psychological and cognitive functional measures will be summarized for each treatment group and all subjects (Tables 3.5.2a-c). Significance tests of the balances between the two treatment groups will be performed and p-values will be presented.

### 3.5.3 Medical History

The number and percentage of subjects reporting a medical history will be summarized by the following medical conditions: Allergies, HEENT disorder, Cardiovascular disorder, Renal disorder, Hepatic disorder, Pulmonary disorder, Gastrointestinal disorder, Musculoskeletal disorder, Neurological disorder, Psychiatric disorder, Dermatologic disorder, Metabolic disorder, Hematologic disorder, Endocrine disorder, Genitourinary disorder, Reproductive system disorder, and Traumatic brain (Tables 3.5.3a-c).

### 3.5.4 Pre-randomization Signs and Symptoms (Adverse Events)

Pre-randomization signs and symptoms will be coded using MedDRA updated version. Any adverse event that has a start date prior to the date of the randomization

will be summarized as pre-randomization adverse events. The number and percentage of subjects reporting each pre-randomization adverse events will be summarized by body system. The number and percentage of subjects reporting any pre-randomization adverse events will also be reported (Tables 3.5.4a-c). If a subject reports the same pre-randomization adverse event more than once, then that subject is only counted once for the summary of that pre-randomization adverse event, using the most severe intensity.

## 3.6 Prior and Concomitant Medications

Concomitant medications are recorded every week in acute and taper treatment phases, and every other four weeks in the follow-up phase. Each summary below will be done for each treatment group and for all subjects:

- Prior medications – Prior medications are considered to be any medication that was stopped prior to the date of the randomization
- Concomitant medications – Concomitant medications are considered to be any medication that was taken on or after the date of the first dose of study drug
- All medications recorded on the CRF are coded to the therapeutic drug classes and generic drug names using the World Health Organization (WHO) drug classifications
- The current version of the World Health Organization (WHO) drug dictionary are used for the coding of medications

Each summary will give the number and percentage of subjects who took medications that were coded to each generic drug name and therapeutic drug class, as well as the number and percentage of subjects that took any medication.

Summaries for prior medications will be done for ITT medications (Table 3.6.1), while summaries for concomitant medication will be done for ITT (Table 3.6.2a), completer (Table 3.6.2b), and fully compliant (Table 3.6.2c), populations.

## 3.7 Efficacy Analyses

The primary endpoint analysis will be done for the ITT population. All secondary and other efficacy analyses will be done for the ITT, completer, and fully compliant populations, unless otherwise specified. Potential difficulties of convergence of model fit, which arise in efficacy analyses that include covariates, will be handled using the method described in Section 3.7.1.

At the completion of the follow-up phase for all subjects in the clinical trial, the efficacy endpoints analyses will be performed. The primary efficacy analysis is based on the measurements of Hamilton Rating Scale for Depression (HRSD). The other key efficacy analyses will include:

- Depression measured by Montgomery-Asberg Depression Rating Scales (MADRS)

- Depression measured by Beck Depression Inventory (BDI)

- Suicide Ideation measured by Beck Scale for Suicide Ideation (BSS)

- Quality of Life measured by the VR-36

- Cognitive Function as measured by a neuropsychological battery

All endpoints will be summarized for all visits once the study is complete.

### 3.7.1 Primary Efficacy Analysis

*3.7.1.1 Primary Endpoint Analysis*

The primary endpoint is the proportion of subjects in each treatment group who achieve "remission" from depression at the end of the acute treatment phase (after a maximum of 30 sessions). "remission" is assessed by a HRSD score $\leq 10$ at the end of the acute treatment session (after a maximum of 30 sessions). For patients who dropout before completing the acute treatment phase will be considered treatment failures for the primary analysis on remission rates. The primary null hypothesis is that there is no difference between the treatment and sham groups in the proportion of subjects who are remitted during the acute treatment phase. The primary variable is defined in Section 2.4.1.

Logistic regression will be used for the primary endpoint ($Y = 1$ if remitted,

otherwise $Y = 0$) analysis on the ITT population with treatment group as the testing factor ($x$). The following covariates will be included in the model: PTSD diagnosis (yes/no as assessed by SCID, $z_1$), History of substance abuse (yes/no as assessed by SCID, $z_2$), and Site (9 participating sites, $z_3$). Given the remission probability $p = \Pr(Y = 1 \mid x, z_1, z_2, z_3)$ the basic model is defined as follows:

$$\log it(p) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x + \beta_3 z_1 + \beta_4 z_2 + \beta_3 z_3$$

In the event that the logistic regression does not converge when all covariates are included, the model will be refit by dropping the covariates from the model one at a time in the following order until convergence is obtained.

    a. History of substance abuse

    b. PTSD diagnosis

    c. Site

The number and percentage of subjects in remission at the end of the acute treatment will be summarized and compared using Pearson $\chi^2$ test (Table 3.7.1a). Odds ratio and 95% confidence interval (CI) will be presented using SAS PROC GENMOD. If the coefficient for treatment effect is significant (i.e., the confidence interval for the odds ratio does not include 1), then the null hypothesis will be rejected (McCullagh and Nelder, 1989). The significant level for each covariate coefficient in the model will best determined by Wald test(Table 3.7.1b).

*3.7.1.2 Additional Analyses for Primary Measurement*

Besides the primary endpoint analysis, other analyses will be performed based on HRSD measurements.

3.7.1.2.1 Durability of Remission

The durability of remission is the subjects in each treatment group who sustain "remission" from depression after 24 weeks post-treatment at the end of the follow-up phase. "Remission" is assessed by a HRSD score $\leq 10$ at the end of the follow-up phase. The null hypothesis is that there is no difference between the treatment and sham groups in the time to recurrence of subjects who sustain remitted after 24 weeks

post-treatment at the end of the follow-up phase. If a subject is no longer remitted at a certain follow-up time point $t$ based on the criteria, then the subject fails and the failure time is the study day of failure ($X_{ij}$). Durability is calculated by subtracting the date which the subject first becomes remitted from the date of recurrence. Similarly, if a subject remains remitted at a certain follow-up time point, then the subject is censored and the censoring time ($C_{ij}$) is the date of censoring minus the study day on which the subject became remitted. Group differences will be tested based on the Log-rank test (Mantel 1966). Both Kaplan-Meier estimates (Kaplan and Meier 1958) and the Log-rank statistics will be obtained using PROC LIFETEST (Tables 3.7.2a-c). The survival distribution of durability of remission between the two treatment groups will be plotted using Kaplan-Meier curves (Figures 3.7.2a-c).

3.7.1.2.2 Response to Treatment

This secondary analysis of the primary efficacy is the proportion of subjects in each treatment group who respond to the treatment. "response" is defined as > 50% decrease of HRSD score at the end of the acute treatment phase. The null hypothesis is that there is no difference between the treatment and sham groups in the proportion of subjects who respond to the treatment at the end of the acute treatment phase. The primary variable is defined in Section 2.4.1. The logistic model will include three covariates, PTSD diagnosis, History of substance abuse, and Site. The number and percentage of subjects in remission at the end of the acute treatment will be summarized (Table 3.7.3a-c). Odds ratio and 95% confidence interval (CI) will be presented using SAS PROC GENMOD. If the coefficient for treatment effect is significant (i.e., the confidence interval for the odds ratio does not include 1), then the null hypothesis will be rejected(Table 3.7.4a-c).

3.7.1.2.3. Remission Rates from Baseline to the Acute Treatment (Short term efficacy)

These endpoints are the proportions of subjects in each treatment group who are remitted at end of acute treatment. The null hypothesis is that there is no difference between the rTMS and sham groups in the proportions of subjects who are remitted across those time points after the acute treatment. The criteria to define the

remission are described in Section 3.7.1. The number and percentage of subjects' remission at the baseline and end of each five sessions the acute treatment will be summarized (Table 3.7.5). Odds ratio and 95% confidence interval (CI) will be presented using SAS PROC GENMOD. SAS PROC GENMOD will be used to run logistic regressions and odds ratios and their 95% confidence intervals will be presented. Further, the generalized linear mixed model will be applied to test the differences of the proportions of subjects' remission across all the time point (Breslow and Clayton 1993). The link function will be logit. SAS PROC GLIMMIX will be used to analyze the data (Tables 3.7.5a-c). The  generalized linear mixed model will include three covariates, PTSD diagnosis, History of substance abuse, and Site.

### 3.7.1.2.4. Remission Rates from End of the Acute Treatment to End of the Follow-up (Long term efficacy)

These endpoints are the proportions of subjects in each treatment group and the data collected from the end of acute treatment to the end of follow-up. The analytic technique is parallel to those described above in Section 3.7.1.2.3. The null hypothesis is that there is no difference between the rTMS and sham groups in the proportions of subjects who are remitted across those time points from the end of the acute treatment to the end of follow-up. The results will be shown in Tables 3.7.6a-c).

### 3.7.1.2.5. HRSD Score Over Time

HRSD score at the baseline, end of each treatment block, and end of every four week follow-up time will be tested between the two treatment groups using SAS PROC GLM at each visit. A mixed regression model will be used to test the main effects of treatment using SAS PROC MIXED across all time visits. The model will include week, treatment and treatment by week interaction adjusted for the baseline. A means model will be assumed. An unstructured covariance model will be used in the analysis. If the model fails to converge a compound symmetry covariance matrix will be assumed. The restricted maximum likelihood (REML) method will be used to estimate the covariance parameters. If the treatment by week interaction is not statistically significant then this term will be removed from the final model. The repeated factor will be visit; all other variables will be considered as fixed effects. The model is defined as follows:

$$Y_{ijk} = \mu + \tau_k + B_j + \varepsilon_{ijk}$$

where $Y_{ijk}$ is the HRSD scores measured on $k^{th}$ treatment and on $j^{th}$ visit for $i^{th}$ subject, $\mu$ is the overall mean, $\tau_k$ is the mean effect of the treatment and the interaction, $B_j$ is the random effect, and $\varepsilon_{ijk}$ is the error (Laird et al 1982). The sample sizes, least-squares means, and standard deviations will be presented at each visit by treatment arm for each study visit (Tables 3.7.7a-c). The assumptions relating to the mixed-effects model will be reviewed by examining residual graphs. Scatter plots of residuals will be used to examine deviations from normality and to assess lack-of-fit. The least square means (LS-means) and their standard errors at each visit will be plotted by treatment groups (Figures 3.7.3a-c).

### 3.7.2 Secondary Efficacy Analyses

The secondary efficacy analyses include outcome measures of Montgomery-Asberg Depression Rating Scales (MADRS), Beck Scale for Suicide Ideation (BSS), Beck Depression Inventory (BDI), Quality of Life measured by the VR-36, and Cognitive Function as measured by a neuropsychological battery.

*3.7.2.1 Montgomery-Asberg Depression Rating Scales (MADRS)*

An important secondary outcome variable is MADRS which is defined in Section 2.5.3.2. Higher MADRS score indicates more severe depression, and each item yields a score of 0 to 6. The overall score ranges from 0 to 60. At each time point of the baseline, end of acute treatment, and end of follow-up, mean and standard deviation of each item will be tabulated by treatment group (Tables 3.7.8a-c).

<u>3.7.2.1.1. MADRS Score Overtime</u>

MADRS score at the baseline, end of each treatment block, and end of every four week follow-up time will be tested between the two treatment groups using SAS PROC GLM at each visit (Tables 3.7.9a-c). A mixed regression model will be used to test the main effects of treatment using SAS PROC MIXED. The model will include visit, treatment and treatment by visit interaction adjusted for the baseline. A means model will be assumed. An unstructured covariance model will be used in the analysis. If the model fails to converge a compound symmetry covariance matrix will be

assumed. The restricted maximum likelihood (REML) method will be used to estimate the covariance parameters. If the treatment by visit interaction is not statistically significant then this term will be removed from the final model. The model is described in Section 3.7.1.2.5. The sample sizes, least-squares means, and standard deviations will be presented at each visit by treatment arm for each study visit (Tables 3.7.9a-c). The assumptions relating to the mixed-effects model will be reviewed by examining residual graphs. Scatter plots of residuals will be used to examine deviations from normality and to assess lack-of-fit.  The least square means (LS-means) and their standard errors at each visit will be plotted by treatment groups
(Figures 3.7.4a-c).

### 3.7.2.1.2. Depression Severity based on MADRS Score

According to Herrmann et al (Herrmann et al 1998), MADRS score can be categorized into four depression severity groups, i.e. normal (0-6), mild depression (7-19), moderate depression (20-34), and severe depression (>34). The severity of the depression will be tested at end of the acute treatment, and end of the following–up using proportional odds model adjusted for baseline and SAS PROC LOGISTIC will be used to run the model and p-values related to the treatment effect will be presented. Further, the generalized linear mixed model will be applied to test the differences of the proportions of subjects' remission across all the time point. The link function will be cumulative logit. SAS PROC GLIMMIX will be used to analyze the data (Tables 3.7.10a-c).

### *3.7.2.2 Beck Depression Inventory (BDI)*

BDI is described in Section 2.5.3.3. Higher BDI score indicates more severe depression, and each item yields a score of 0 to 3. The overall score of the 21 item test ranges from 0 to 63. At each time point of the baseline, end of acute treatment, and end of follow-up, mean and standard deviation of each category will be tabulated by treatment group (Table 3.7.11 a-c).

### 3.7.2.2.1. BDI Score Overtime

BDI score at the baseline, end of each treatment block, and end of every four week follow-up time will be tested between the two treatment groups using SAS PROC

GML at each visit. A mixed regression model will be used to test the main effects of treatment using SAS PROC MIXED. The model will include visit, treatment and treatment by visit interaction adjusted for the baseline. A means model will be assumed. An unstructured covariance model will be used in the analysis. If the model fails to converge a compound symmetry covariance matrix will be assumed. The restricted maximum likelihood (REML) method will be used to estimate the covariance parameters. If the treatment by visit interaction is not statistically significant then this term will be removed from the final model. The model is described in Section 3.7.1.2.5. The sample sizes, least-squares means, and standard deviations will be presented at each visit by treatment arm for each study visit (Tables 3.7.12a-c). The assumptions relating to the mixed-effects model will be reviewed by examining residual graphs. Scatter plots of residuals will be used to examine deviations from normality and to assess lack-of-fit.  The least square means (LS-means) and their standard errors at each visit will be plotted by treatment groups (Figures 3.7.5a-c).

### 3.7.2.2.2. Depression Severity based on BDI Score

According to Beck et al (Beck et al 1988), BDI score can be categorized into four depression severity groups, i.e. minimal depression (0-9), mild depression (10-18), moderate depression (19-29), and severe depression (30-63). The severity of the depression will be tested at end of the acute treatment, and end of the following–up using proportional odds model and SAS PROC LOGISTIC will be used to run the model and p-values related to the treatment effect will be presented. Further, the generalized linear mixed model will be applied to test the differences of the proportions of subjects' remission across all the time point. The link function will be cumulative logit. SAS PROC GLIMMIX will be used to analyze the data (Tables 3.7.13a-b).

### 3.7.2.3 Beck Scale for Suicide Ideation (BSS)

BSS is described in Section 2.5.3.4. Higher BSS score indicates higher tendency of suicidal ideation, and each item yields a score of 0 to 2. The overall score of the 21 item test ranges from 0 to 38 (last two items not counted). At each time point of the baseline, end of acute treatment, and end of follow-up, mean and standard deviation of each item will be tabulated by treatment group (Table 3.7.14 a-c). BSS

score at the baseline, end of each treatment block, and end of every four week follow-up time will be tested between the two treatment groups using SAS PROC GLM at each visit adjusted for the baseline. A mixed regression model will be used to test the main effects of treatment using SAS PROC MIXED. The model will include visit, treatment and treatment by visit interaction adjusted for the baseline. A means model will be assumed. An unstructured covariance model will be used in the analysis. If the model fails to converge a compound symmetry covariance matrix will be assumed. The restricted maximum likelihood (REML) method will be used to estimate the covariance parameters. If the treatment by visit interaction is not statistically significant then this term will be removed from the final model. The model is described in Section 3.7.1.2.5. The sample sizes, least-squares means, and standard deviations will be presented at each visit by treatment arm for each study visit (Tables 3.7.15a-c). The assumptions relating to the mixed-effects model will be reviewed by examining residual graphs. Scatter plots of residuals will be used to examine deviations from normality and to assess lack-of-fit.  The least square means (LS-means) and their standard errors at each visit will be plotted by treatment groups (Figures 3.7.6a-c).

### 3.7.2.4. Veterans RAND 36 Item Health Survey (VR-36)

As described in the section 2.5.3.5, the VR-36 measures eight concepts of health using 12 questions with 37 items: physical functioning, role limitations due to physical health, bodily pain, general health perceptions, vitality, social functioning, role limitations due to emotional problems, and mental health. It yields scale scores for each of these eight health domains. Scoring of all eight VR-36 is based on Ware et al (1994) standard scoring algorithm, which includes a linear transformation from a raw score so that scores range from 0-100, where 100 denotes the best health. The eight scales are summarized into two components, physical and mental component summaries(PCS and MCS) and each is scored using weights derived from a national US probability sample (Ware et al 1994). The two summaries make an important contrast between the physical and psychological health status.  At each time point of the baseline, end of acute treatment, and end of follow-up, mean and standard deviation of each health domain will be tabulated by treatment group (Table 3.7.16a-c).

Physical, mental and total scores at the baseline, end of each treatment block, and end of every four week follow-up time will be tested between the two treatment groups using SAS PROC GLM at each visit adjusted for the baseline. A mixed regression model will be used to test the main effects of treatment using SAS PROC MIXED. The model will include visit, treatment and treatment by visit interaction adjusted for the baseline. A means model will be assumed. An unstructured covariance model will be used in the analysis. If the model fails to converge a compound symmetry covariance matrix will be assumed. The restricted maximum likelihood (REML) method will be used to estimate the covariance parameters. If the treatment by visit interaction is not statistically significant then this term will be removed from the final model. The model is described in Section 3.7.1.2.5. The sample sizes, least-squares means, and standard deviations will be presented at each visit by treatment arm for each study visit (Tables 3.7.17a-c). The assumptions relating to the mixed-effects model will be reviewed by examining residual graphs. Scatter plots of residuals will be used to examine deviations from normality and to assess lack-of-fit. The least square means (LS-means) and their standard errors at each visit will be plotted by treatment groups (Figures 3.7.7a-c, and Figures 3.7.7d-f,).

### 3.7.2.5. Cognitive Function Analysis (Neuropsychological Battery)

Cognitive functions are assessed at screening, at the end of the acute treatment phase and at the end of the 24 week follow-up phase, including measures of executive function, attention, memory, visuospatial ability, processing speed, psychomotor function, and premorbid intelligence.

### 3.7.2.5.1.Self-Rating Scale Memory Function

This is an 18 item self-rating scale of memory function test which is constructed to ask subjects to compare their memory now to their memory during the period before hospitalization. For each item, subjects rated themselves on a 9 point scale from $-4$ (worse than ever before), through 0 (same as before), to 4( better than ever before).

At each time point of the baseline, end of acute treatment, and end of follow-up, mean and standard deviation of each item rating scale and overall rating scale will be tabulated by treatment group (Table 3.7.18a-c). The overall scale will be analyzed at

each measurement time point using SAS PROC GLM adjusted for baseline, age, dominant hand, year of education, medical history, neurological illness, learning disability, substance abuse and test familiarity; and the mixed model will be performed to test overall scale using SAS PROC MIXED adjusted for the above covariates as well as visit. The sample sizes, least-squares means, and standard deviations will be presented at each visit by treatment arm for each study visit (Tables 3.7.19a-c). The assumptions relating to the mixed-effects model will be reviewed by examining residual graphs. Scatter plots of residuals will be used to examine deviations from normality and to assess lack-of-fit.  The means (LS-means) and their standard errors at each visit will be plotted by treatment groups (Figures 3.7.8a-c).

### 3.7.2.5.2 Rey Auditory Verbal Learning Test (RAVLT)

The RAVLT measures a wide range of functions such as short-term auditory-verbal memory, rate of learning, learning strategies, retroactive, and proactive interference, presence of confabulation of confusion in memory processes, retention of information, and differences between learning and retrieval. Detailed method is described in Section 2.5.3.6. The test raw scores including Trials I-V, Trial B, Trial VI, Delayed recall, Recognition memory, False positives, and Trials I-V total are collected. At each time point of the baseline, end of acute treatment, and end of follow-up, mean and standard deviation of each raw score will be tabulated by treatment group (Table 3.7.20a-c). The test raw scores at each time point  will be analyzed using SAS PROC GLM adjusted for baseline, age, dominant hand, year of education, medical history, neurological illness, learning disability, substance abuse and test familiarity; and the mixed model will be performed to test the each score over time using SAS PROC MIXED adjusted for the same covariates. The sample sizes, least-squares means, and standard deviations will be presented at each visit by treatment arm for each study visit (Tables 3.7.21a-c). The assumptions relating to the mixed-effects model will be reviewed by examining residual graphs. Scatter plots of residuals will be used to examine deviations from normality and to assess lack-of-fit.

### 3.7.2.5.3 Symbol Digit Modalities Test (SDMT)

The SDMT developed by Aaron Smith (Smith, 1991) and measuring cerebral dysfunction in children and adults, is a test for divided attention, but requires complex visual scanning and tracking, perceptual speed, motor speed, and memory. The score in the written administrations of the SDMT test is the number of correct substitutions in each 90-second interval. The total number of correct responses can be found by counting the number of responses that correctly match the number printed above each box. The score is recorded as a proportion of the total number of responses. The total score provides a measure of the speed and accuracy of symbol-digit substitutions. At each time point of the baseline, end of acute treatment, and end of follow-up, mean and standard deviation of each correct responses, total responses, and the proportion of the total number of responses will be tabulated by treatment group (Table 3.7.22a-c). The test  scores at each time point  will be analyzed using SAS PROC GLM adjusted for baseline, age, dominant hand, year of education, medical history, neurological illness, learning disability, substance abuse and test familiarity; and the mixed model will be performed to test the each score over time using SAS PROC MIXED adjusted for the same covariates. The sample sizes, least-squares means, and standard deviations will be presented at each visit by treatment arm for each study visit (Tables 3.7.23a-c). The assumptions relating to the mixed-effects model will be reviewed by examining residual graphs. Scatter plots of residuals will be used to examine deviations from normality and to assess lack-of-fit.

### 3.7.2.5.4 Trail Making Test (TMT)

TMT assesses visual attention and task switching, consisting of two parts (A and B) in which the examinee is instructed to connect  a set of 25 dots as fast as possible while still maintaining accuracy. The completing time and number of errors will be collected as outcome measures for both part A and part B, and tabulated by treatment group (Table 3.7.24a-c). The test  outcomes at each time point  will be analyzed using SAS PROC GLM adjusted for baseline, age, dominant hand, year of education, medical history, neurological illness, learning disability, substance abuse and test familiarity; and the mixed model will be performed to test the outcomes over time using SAS PROC MIXED adjusted for the same covariates. The sample sizes,

least-squares means, and standard deviations will be presented at each visit by treatment arm for each study visit (Tables 3.7.25a-c). The assumptions relating to the mixed-effects model will be reviewed by examining residual graphs.

### 3.7.2.5.5. Judgement of Line Orientation (JLO)

As standardized test of visuospatial skills, JLO assesses a subject's ability to match the angle and orientation of lines in space. The 30 item test requires the examinee to match two angled lines to a set of 11 lines arranged in a semicircle and separated 18 degrees from each other. The number of total correct answers are collected and tabulated by treatment group (Table 3.7.26a-c). The test outcomes at each time point will be analyzed using SAS PROC GLM adjusted for baseline, age, dominant hand, year of education, medical history, neurological illness, learning disability, substance abuse and test familiarity; and the mixed model will be performed to test the outcomes over time using SAS PROC MIXED adjusted for the same covariates. The sample sizes, least-squares means, and standard deviations will be presented at each visit by treatment arm for each study visit (Tables 3.7.27a-c). The assumptions relating to the mixed-effects model will be reviewed by examining residual graphs. Scatter plots of residuals will be used to examine deviations from normality and to assess lack-of-fit.

### 3.7.2.5.6. Controlled Oral Word Association Test (COWA)

COWA test assesses subjects' verbal fluency. examinees are asked to produce as many words that begin with a specific letter (F, A, or S) as they can within one minute. The examinees are then asked to name as many animal names as possible within one minute. The test is scored for specific letter F, A, S, total of FAS (F+A+S), and animal naming in four aspects, i.e. number of correct, perservations, intrusions, and variants. The raw scores will be tabulated by treatment group (Table 3.7.28a-c). Every score at each time point will be analyzed using SAS PROC GLM adjusted for baseline, age, dominant hand, year of education, medical history, neurological illness, learning disability, substance abuse and test familiarity; and the mixed model will be performed to test the raw scores over time using SAS PROC MIXED adjusted for the same covariates. The sample sizes, least-squares means, and standard deviations will

be presented at each visit by treatment arm for each study visit (Tables 3.7.29a-c). The assumptions relating to the mixed-effects model will be reviewed by examining residual graphs. Scatter plots of residuals will be used to examine deviations from normality and to assess lack-of-fit.

### 3.7.2.5.7. Stroop Color and Word Test (SCWT)

SCWT assesses cognitive processing and helps diagnosis on brain dysfunction, cognition, and psychopathology. The test is based on the observation which subjects read words much faster than they identify and name colors. The Stroop Test yields three basic scores: raw word score, raw color score, and the raw color-word score. These scores will be tabulated by treatment group (Table 3.7.30a-c). Every score at each time point  will be analyzed using SAS PROC GLM adjusted for baseline, age, dominant hand, year of education, medical history, neurological illness, learning disability, substance abuse and test familiarity; and the mixed model will be performed to test the raw scores over time using SAS PROC MIXED adjusted for the same covariates. The sample sizes, least-squares means, and standard deviations will be presented at each visit by treatment arm for each study visit (Tables 3.7.31a-c). The assumptions relating to the mixed-effects model will be reviewed by examining residual graphs. Scatter plots of residuals will be used to examine deviations from normality and to assess lack-of-fit.

### 3.7.2.5.8. North American National Adult Reading Test (NAART)

NAART estimates subjects' verbal intellectual ability. The use of a pronunciation guide is required to facilitate scoring for the NAART. If a subject pronounces a word incorrectly, put a checkmark in the "No" box and if a patient correctly pronounces a word, put a check mark in the "Yes" box on the scoring sheet. Each incorrectly pronounced word counts as one error. The total number of errors will be tabulated by treatment group (Table 3.7.32a-c). Every score at each time point  will be analyzed using SAS PROC GLM adjusted for baseline, age, dominant hand, year of education, medical history, neurological illness, learning disability, substance abuse and test familiarity; and the mixed model will be performed to test the raw scores over time using SAS PROC MIXED adjusted for the same covariates. The sample sizes, least-

squares means, and standard deviations will be presented at each visit by treatment arm for each study visit (Tables 3.7.33a-c). The assumptions relating to the mixed-effects model will be reviewed by examining residual graphs. Scatter plots of residuals will be used to examine deviations from normality and to assess lack-of-fit.

*3.7.2.6 Secondary Endpoint Analysis Results*

The secondary endpoints are defined as severity improvements of MADRS, BSS, BDI, VR-36, and Cognitive function from baseline to end of the acute treatment. Given the several simultaneous hypothesis testing, multiplicity adjustment will be applied. The significance level for the secondary endpoint analyses will be 0.01. The analysis results can be extracted from the above analysis at the time point of end of acute treatment as shown in Table 3.7.34a-c.

**3.7.3 Other Efficacy Analyses**

*3.7.3.1. Quick Inventory of Depressive Symptomatology (QIDS-C16)*

Since the HRSD does not measure hypersomnia, weight gain or problems with concentration or decision making, the Quick Inventory of Depressive Symptomatology (QIDS-C16), which meets the ACNP criteria is used for the additional MDD measurement (Rush et al., 2003). This is a clinician rated 16 item depressive assessment. Items 1-4 assess sleeping problem and the score is the highest of any 1 of the 4 items (max 3); items 6-9 assess weight problems and the score is the highest of any 1 of the 4 items (max 3); items 15-16 assess psychomotor problems and the score is the higher one (max 3); the remaining 6 items assess other MDD related functions and each item has one score (0-3). Thus, the total score for the nine MDD symptom domains (sleep, weight, psychomotor changes, depressed mood, decreased interest, fatigue, guilt, concentration, and suicidal ideation) range from 0 to 27. Higher score indicates more severe depression. At each time point of the baseline, end of acute treatment, and end of follow-up, mean and standard deviation of each item will be tabulated by treatment group (Tables 3.7.35a-c).

3.7.3.1.1. QIDS-C16 Score Overtime

The QIDS-C16 score at the baseline, end of each treatment block, and end of every four week follow-up time will be tested between the two treatment groups using

SAS PROC GLM at each visit. A mixed regression model will be used to test the main effects of treatment using SAS PROC MIXED. The model will include visit, treatment and treatment by visit interaction adjusted for the baseline. A means model will be assumed. An unstructured covariance model will be used in the analysis. If the model fails to converge a compound symmetry covariance matrix will be assumed. The restricted maximum likelihood (REML) method will be used to estimate the covariance parameters. If the treatment by visit interaction is not statistically significant then this term will be removed from the final model. The model is described in Section 3.7.1.2.5. The sample sizes, least-squares means, and standard deviations will be presented at each visit by treatment arm for each study visit (Tables 3.7.36a-c). The assumptions relating to the mixed-effects model will be reviewed by examining residual graphs. Scatter plots of residuals will be used to examine deviations from normality and to assess lack-of-fit. The least square means (LS-means) and their standard errors at each visit will be plotted by treatment groups (Figures 3.7.9a-c).

### 3.7.3.1.2. Depression Severity based on QIDS-C16 Score

QIDS-C16 score can be categorized into five depression severity groups, i.e. no depression (0-5), mild depression (6-10), moderate depression (11-15), severe depression (16-20), and very severe depression (21-27). The severity of the depression will be tested at end of the acute treatment, and end of the following–up using proportional odds model and SAS PROC LOGISTIC will be used to run the model and p-values related to the treatment effect will be presented. Further, the generalized linear mixed model will be applied to test the differences of the proportions of subjects' remission across all the time point. The link function will be cumulative logit. SAS PROC GLIMMIX will be used to analyze the data (Tables 3.7.37a-c).

### 3.7.3.2. Beck Hopelessness Scale (BHS)

As described in Section 2.5.3.8, the BHS is a 20-item self-report inventory developed by Dr. Aaron T. Beck that was designed to measure three major aspects of hopelessness: feelings about the future, loss of motivation, and expectations, consisting of 20 "yes/no" items with possible scores from 0 to 20 (Beck 1974). Higher score indicates more severe pessimism.

### 3.7.3.2.1. BHS Overtime

BHS at the baseline, end of each treatment block, and end of every four week follow-up time will be tested between the two treatment groups using SAS PROC GLM at each visit. A mixed regression model will be used to test the main effects of treatment using SAS PROC MIXED. The model will include visit, treatment and treatment by visit interaction adjusted for the baseline. A means model will be assumed. An unstructured covariance model will be used in the analysis. If the model fails to converge a compound symmetry covariance matrix will be assumed. The restricted maximum likelihood (REML) method will be used to estimate the covariance parameters. If the treatment by visit interaction is not statistically significant then this term will be removed from the final model. The model is described in Section 3.7.1.2.5. The sample sizes, least-squares means, and standard deviations will be presented at each visit by treatment arm for each study visit (Tables 3.7.38a-c). The assumptions relating to the mixed-effects model will be reviewed by examining residual graphs. Scatter plots of residuals will be used to examine deviations from normality and to assess lack-of-fit.  The least square means (LS-means) and their standard errors at each visit will be plotted by treatment groups (Figures 3.7.10a-c).

### 3.7.3.2.2. Hopelessness Severity based on BHS

A total severity of hopelessness is calculated from summing the 20 items and guidelines for interpretation for scores are as follows: 0-3 = minimal hopelessness; 4-8 = mild hopelessness; 9-14 = moderate hopelessness; and 15-20 = severe hopelessness.. The severity of hopelessness will be tested at end of the acute treatment, and end of the following–up using proportional odds model and SAS PROC LOGISTIC will be used to run the model and p-values related to the treatment effect will be presented. Further, the generalized linear mixed model will be applied to test the differences of the proportions of subjects' remission across all the time point. The link function will be cumulative logit. SAS PROC GLIMMIX will be used to analyze the data (Tables 3.7.39a-c).

## 3.8 Safety Analyses

All safety analyses will be done for the safety population and reported in tabular forms. These analyses include adverse events/serious adverse events, vital signs, physical examinations, birth control/pregnancy test, safety trigger tests, and drug abuse and alcoholism tests.

### 3.8.1 Adverse Events

An adverse event is any symptom, sign, illness, or experience that develops or worsens in severity and/or frequency after the start of the study drug. Incidence of adverse events will be summarized for each treatment group by body system and MedDRA term. The number and percentage of subjects with each body system and MedDRA term will be presented for each treatment group. Pearson $\chi^2$ test and/or Fisher Exact test were used to compare the frequency difference of AEs and SAEs between the treatment groups in System of Body (SOC) and Preferred Terms (PT) levels. Tables to summarize the incidence rates will be created for each of the following groups:

- Adverse events (Table 3.8.1)
- Adverse events by relationship to study treatment and device (Tables 3.8.2a-b)
- Adverse events by intensity (Table 3.8.3)
- Adverse events leading to premature discontinuation (Table 3.8.4)
- Adverse events presented in descending order of frequency by MedDRA term (no body systems shown) (Table 3.8.5)
- Serious adverse events (Table 3.8.6)

Adverse events that led to premature discontinuation from the study will be listed. Serious adverse events will also be listed. These listings will contain details about the adverse event such as intensity and relationship to study treatment. Other supportive data, such as the subject's age, will be given. All adverse events will be coded with MedDRA (updated version). The adverse events and severe adverse events will be listed by subject (Tables 3.8.7-8).

### 3.8.2. Vital Signs

Vital sign measurements include height (inches), weight (pounds), systolic and diastolic blood pressure (mmHg), pulse rate (beats/min), and oral temperature (in

Fahrenheit). Vital signs are measured at screening/baseline and the final treatment block. Vital signs mean changes from baseline to end of the treatment will be summarized. The summary will present data for the baseline visit and the mean change from baseline to the final evaluation. For each visit, the sample size, mean, SD, median, minimum, and maximum values for each treatment will be presented for each parameter (Table 3.8.9), and tested using SAS PROC TTEST. If vital signs are not normally distributed  the non-parametric method will be applied using SAS PROC NPAR1WAY.

### 3.8.3. Physical Examination

Physical examinations will be performed at screening/baseline and the final treatment block. Each body system will be categorized as abnormal, normal, or not done. Abnormal physical examination will be recorded in text. Physical examination changes from baseline to  will be categorized as improved, no change, or worsened for each body system. The number and percentage of subjects in each category of change will be given at each visit for each treatment group and body system (Table 3.8.10). SAS PROC GENMOD will be used to test these physical parameters adjusted for the baselines.

### 3.8.4. Birth Control and Pregnancy Test

Birth control and pregnancy test data will be collected within 7 days prior to randomization and every 4 weeks thereafter throughout the study. The number and percentage of subjects for each category of method of birth control use and pregnancy test results (N/A (male), Not done, Negative, and Positive) will be presented at each test for each treatment groups and all subjects Table 3.8.11 and Table 3.8.12 respectively.

### 3.8.5. Trigger Tests

#### 3.8.5.1. Columbia – Suicide Severity Rating Scale (C-SSRS)

The C-SSRS is a measure of suicidal ideation and behavior. The primary outcomes of the measurement are suicidal ideation (1-5 items), suicidal behavior (6-10 items),  and self-injurious behavior without suicidal intent; and the secondary outcomes are number of suicidal behaviors (total number of each type of suicidal behavior, i.e.

suicide attempts, aborted attempts, and interrupted attempts) and lethality rating (rated1-5) for suicidal behavior, and severity rating (rated 0-5) and intensity (scored 0-25) for suicidal ideation. The baseline scores are compared with those of past six month, while the follow-up scores are compared with those from last visit.

The number and percentage of subjects for each category of suicidal ideation, suicidal behavior,  completed suicide, as well as self-injurious behavior without suicidal intent, will be presented at each measurement by the treatment groups and all subjects (Table 3.8.13). The difference of  suicidal ideation, suicidal behavior, suicidal ideation or behavior, completed suicide and self-injurious behavior without suicidal intent between the treatment groups will be tested at each measurement time point using Fisher exact test in SAS PROC FREQ when events are rare. Further, the generalized linear mixed model will be applied to test the differences of the proportions across all the time point. SAS PROC GLIMMIX will be used to analyze the data. For the generalized linear mixed models will include visit, treatment and the visit and treatment interaction adjusted for the baseline measurement. then, Intensity of ideation will also be analyzed based on the most severe ideation, frequency, duration, controllability, deterrents, and reasons for ideation which are tabulated in Table 3.8.14a. The difference of  these intensity items will be tested at each measurement time point using Fisher exact test. The intensity scores for each category and the total scores of all five categories at the baseline, end of each treatment block, and end of every four week follow-up time will be tested between the two treatment groups using SAS PROC GLM at each visit. A mixed regression model will be used to test the main effects of treatment using SAS PROC MIXED. The model will include visit, treatment and treatment by visit interaction adjusted for the baseline as shown Table 3.8.14b.

### 3.8.5.2.  State-Trait Anger Expression Inventory-2 (STAXI-2)

The STAXI-2 assesses various areas of anger and the traits of experiencing anger. The 57 item assessment with a 4-point score ("Not at all" to "Almost Always") includes three sections or scales, State Anger ( subscaled S-Ang, S-Ang/F, S-Ang/V,and  S-Ang/P), Trait Anger (T-Ang, Tang/T,  and  T-Ang/R),  and  Anger Expression and Anger Control (AX-O, AX-I, AC-O, AC-I, and AX Index). For each scale

and subscale, raw scores and percentiles are calculated based on the four point score for each item. At each time point of the baseline and each visit after treatment and during follow-up, mean and standard deviation, as well as median and interquartile range, of each scale/subscaled raw scores and will be summarized by the treatment groups (Table 3.8.15a-b). The raw scores and percentile for each scale/subscale at the baseline, end of each treatment block, and end of every four week follow-up time will be tested between the two treatment groups using SAS PROC GLM at each visit. A mixed regression model will be used to test the main effects of treatment using SAS PROC MIXED. The model will include visit, treatment and treatment by visit interaction adjusted for the baselines. The percentiles of each scale are categorized into $\leq 25\%$, >25% to <75%, and $\geq 75\%$. At each time point of the baseline and each visit after treatment and during follow-up, numbers and percentages of each scale/subscaled percentile for each category, as well as clinical evaluation trigger (i.e. $\leq 25\%$ and $\geq 75\%$) are tabulated by the treatment groups (Table 3.8.16). the difference of each category and trigger for each scale/subscale at each visit between the treatment groups will be tested using Pearson $\chi 2$ or Fisher exact test, whichever it applies. The overall differences across all visits for each scale between the treatment groups will be tested by generalized linear mixed model using SAS PROC GLIMMIX.

### 3.8.6. Drug Abuse and Alcoholism Tests

*3.8.6.1. Drug Abuse Screening Test (DAST)*

The DAST is designed to provide a brief instrument for clinical screening drug abuse. This 28 self-report item questionnaire uses 'yes' or 'no' responses to address various problems related to drug misuse, and yields quantitative scores ranging from 0 to 28 with a higher score endorsed in the direction of increased drug use problems. A score of 5 or higher is indicative of a possible drug use disorder. the test is performed at screening, end of acute treatment, and end of follow-up. The number and percentage of each item will be tabulated by the treatment groups at each measurement time (Table 3.8.17). the difference of each item at each visit between the treatment groups will be tested using Pearson $\chi 2$ or Fisher exact test. At each time point, summary statistics of DAST scores will be tabulated (Table 3.8.18) and tested

between the two treatment groups using SAS PROC GLM at each measurement adjusted for the baseline. A mixed regression model will be used to test the main effects of treatment using SAS PROC MIXED. The model will include visit, treatment and treatment by visit interaction adjusted for the baselines. Further the test scores will be dichotomized into < 5 and $\geq$5 as indication of drug use. The logistic regression will be used test the difference between the treatment groups using PROC LOGISTIC and the overall differences across all visits for each scale between the treatment groups will be tested by generalized linear mixed model using SAS PROC GLIMMIX (Table 3.8.19).

*3.8.6.2. Michigan Alcoholism Screening Test (MAST)*

The MAST is parallel self-administered items on the DAST and is a widely used assessment device for alcoholism. This 25 self-report item questionnaire uses 'yes' or 'no' responses to provide screening for clinical and non-clinical settings, and yields quantitative scores ranging from 0 to 50 with a higher score endorsed in the direction of increased alcoholism.  the test score is usually categorized into 0-3 (no apparent problem), 4 (early or middle problem drinker), 5-50 (problem drinker/alcoholic). The test is performed at screening, end of acute treatment, and end of follow-up. The number and percentage of each item will be tabulated by the treatment groups at each measurement time (Table 3.8.20). the difference of  each item at each visit between the treatment groups will be tested using Pearson $\chi2$ or Fisher exact test. At each time point, summary statistics of DAST scores will be tabulated (Table 3.8.21) and tested between the two treatment groups using SAS PROC GLM at each measurement adjusted for the baseline. A mixed regression model will be used to test the main effects of treatment using SAS PROC MIXED. The model will include visit, treatment and treatment by visit interaction adjusted for the baselines. Further the test scores will be dichotomized into 0-3, 4 and 5-50 to assess severity of alcohol drinking. The logistic regression will be used test the difference between the treatment groups using PROC LOGISTIC and the overall differences across all visits for each scale between the treatment groups will be tested by generalized linear mixed model using SAS PROC GLIMMIX (Table 3.8.22).

**3.8.7. Urine Toxicology Screening and Alcohol Test**

Urine rapid toxicology screening for drug and alcohol test is performed at screening phase, acute treatment phase of blocks 2, 4, and 6, taper phase week 2, and follow-up phase of weeks 4, 12 , and 20. The test includes five drugs on dip card A: methamphetamine (MET) and ecstasy (ECS), amphetamine (AMP) and methylenedioxyamphetamine (MDA), marijuana (THC), cocaine (COC), and opiates (OPI); five drugs on dip card B: barbiturates (BAR), benzodiazepines (BZD), phencyclidine (PCP), methadone (MTD), and oxycodone (OXY); and alcohol. All test results are recorded as negative, positive, and no results. At each testing time point, the number and percent of each category of the testing results will be tabulated (Table 3.8.23), and the difference between the treatment groups will be tested using Pearson $\chi 2$ test.

## 3.9 Additional Analyses

**3.9.1 Treatment effect on PTSD**

*3.9.1.1. PTSD Check List (PCL-M)*

The PTSD checklist –military version (PCL-M) used in the study is a 17-item self-report measure reflecting DSM-IV symptoms of PTSD. It has multiple clinical and research purposes including screening individuals for PTSD, helping in diagnostic assessment of PTSD and monitoring change in PTSD symptoms. The PCL-M inquired about symptoms in response to stressful experience in military service based on the 17 items. For each item there were five scales – not at all, a little bit, moderately, quite a bit, and extremely, valued from 1 to 5. The total symptom severity score that ranged from 17-85 was obtained by summing the scores from each of the 17 items. A cutoff score of 44 or above is considered as PTSD positive. At each time point of the baseline, end of acute treatment, and end of follow-up, summary statistics of the total scores will be tabulated by treatment group (Tables 3.9.1a-c), and analyzed at each measurement time point using SAS PROC GLM adjusted for baseline; and the mixed model will be performed to test total score across time point using SAS PROC MIXED. The covariates of the model will include baseline, treatment, time point, and interaction

of time and treatment. The least square means (LS-means) and their standard errors at each visit will be plotted by treatment groups (Figures 3.9.1a-c).PTSD severity is categorized as low (17-33), moderate (34-43), and high ($\geq$44) based on the PCL-M score. The number and percentage of the severity will be tabulated by the treatment groups at each measurement time (Tables 3.9.2a-c). the difference of each category at each visit between the treatment groups will be tested using Pearson $\chi 2$ tests and logistic regression by PORC LOGISITC adjusted for the baseline. The overall differences across all time points for each severity category between the treatment groups will be tested by generalized linear mixed model using SAS PROC GLIMMIX. The model will include the baseline, treatment, time, and interaction of time and the treatment.

### 3.9.1.2. Clinician Administered PTSD Scale (CAPS)

The CAPS is a structured interview for assessing core and associated symptomsof PTSD. It assesses the frequency and intensity of each symptom and yields both continuous and dichotomous scores for current and lifetime PTSD symptoms. Possible scores range from 0 to 136. At each time point of the baseline, end of acute treatment, and end of follow-up, summary statistics of the total scores will be tabulated by treatment group (Tables 3.9.3a-c), and analyzed at each measurement time point using SAS PROC GLM adjusted for baseline; and the mixed model will be performed to test total score across time point using SAS PROC MIXED. The covariates of the model will include baseline, treatment, time point, and interaction of time and treatment. The least square means (LS-means) and their standard errors at each visit will be plotted by treatment groups (Figures 3.9.2a-c). PTSD severity is categorized as asymptomatic (0–19), mild (20-39), moderate (40-59), severe (60-79), and extreme ($\geq$80) based on CAPS score. The number and percentage of the severity will be tabulated by the treatment groups at each measurement time (Tables 3.9.4a-c). the difference of each category at each visit between the treatment groups will be tested using Pearson $\chi 2$ tests and logistic regression by PORC LOGISITC adjusted for the baseline. The overall differences across all time points for each severity category between the treatment groups will be tested by generalized linear mixed model using

SAS PROC GLIMMIX. The model will include the baseline, treatment, time, and interaction of time and the treatment.

### 3.9.2 Prognostic Factor Analysis

Prognostic factors or predictors of treatment efficacy for the primary outcome (HRSD) and major secondary outcome measures ( MADRS, BSS, BDI, and VR-36) at each time point of the baseline, end of acute treatment, and end of follow-up, will be explored using demographic information and baseline assessments such as age, baseline severity, type of comorbidity (PTSD, substance abuse, or both), duration of illness, prior treatment resistance, concomitant medication use, and possibly their interaction with the treatment. The analyses will be performed on all ITT, completer, and fully compliant populations. Linear or generalized linear mixed regressions will be used to model moderate effect with the factors described above on the treatment efficacy, depending on whether the outcome measure is continuous or categorical. The modeling approach is the same as described in the corresponding sections of the primary and secondary outcome analyses (Tables 3.9.5a-c).

In the event that a multiple regression does not converge when all covariates are included, the model will be refit by dropping the problematic covariate from the model and perhaps investigating it in a separate single-covariate model.

### 3.9.3 Additional Baseline Analyses

#### 3.9.3.1 Trauma History Questionnaire (THQ)

THQ is a 24 item self-report measure that examines experiences with potential traumatic events including crime, general disaster, and sexual and physical assault, using a yes/no format. The summary statistics of THQ scores for each event and overall score will be tabulated. Student t test using SAS PROC TTEST or Wilcoxon test using SAS PROC NPAR1WAY will be performed to test score differences between the two treatment groups (Table 3.9.6).

#### 3.9.3.2. Blessed Orientation Memory Concentration Test (BOMC)

The BOMC assesses level of cognitive impairment using 3 orientation questions, counting backwards from 20 to 1, months backwards, and the name and address memory phase. The testing result is measured by a weighted error score with

a maximum 28. The summary statistics of BOMC scores for each event and overall score will be tabulated. Student t test using SAS PROC TTEST or Wilcoxon test using SAS PROC NPAR1WAY will be performed to test score differences between the two treatment groups (Table 3.9.7).

### 3.9.3.3. Life Stressor Checklist-Revised (LSC-R)

The LSC-R is a self-report assessment for traumatic or stressful life events. The questionnaire includes 30 life events of natural disasters, physical or sexual assault, death of a relative and other events using a yes/no format. In addition, a five point intensity scale for each event is used to weight on the event score. The summary statistics of LSC-R score (0-30) and its weighted score (0-150) will be tabulated. Student t test using SAS PROC TTEST or Wilcoxon test using SAS PROC NPAR1WAY will be performed to test score differences between the two treatment groups (Table 3.9.8).

### 3.9.3.4. Lifetime Drinking History (LDH)

As described in Section 2.5.2.1, the LDH is a structured interview designed to provide quantitative data on patterns of alcohol consumption. This assessment is performed only at baseline. The summary statistics of the total alcohol consumption, typical and maximum alcohol consumption per occasion, average daily and average monthly intake both for the last six months (current drinking) and for lifetime (lifetime drinking history) will be tabulated. Student t test using SAS PROC TTEST or Wilcoxon test using SAS PROC NPAR1WAY will be performed to test score differences between the two treatment groups (Table 3.9.9).

### 3.9.3.5. Antidepressant Treatment History Form (ATHF)

The ATHF provides detailed information about antidepressant resistance for a patient to be qualified for the study. The form collects drug name, dosage, blood level, duration, number of weeks in treatment, reason stopped, outcome, overall confidence rating, resistance rating, and adequate trials. The drug resistance criteria are based on adequate trial score. if the score is >2, the patient meets the drug resistant criteria. The summary statistics of the adequate trial score will be tabulated(Table 3.9.10). Student t test using SAS PROC TTEST or Wilcoxon test using SAS PROC NPAR1WAY will be

performed to test score differences between the two treatment groups, and Pearson $\chi2$ test using SAS PROC FREQ will be used to test the patient drug resistant eligibility between the two treatment groups.

### 3.9.3.6. Structured Clinical Interview for DSM-IV (SCID)

The semi-structured interview is to make the major DSM-IV diagnosis criteria for subject eligibility. There are four diagnostic modules (mood disorders, psychotic symptoms, substance use disorder, and anxiety disorders) with two aspects, i.e. lifetime prevalence (absent, subthreshold, threshold), and symptomatic diagnosis criteria in past month (absent, present). The summary statistics of each category of lifetime prevalence and symptomatic diagnosis criteria in past month will be tabulated(Table 3.9.11a-b) for each diagnostic module. Pearson $\chi2$ test using SAS PROC FREQ will be used to test the differences between the two treatment groups.

### 3.9.3.7. Laboratory Data

Hematology, chemistry, liver function, and endocrinology are analyzed at the study site local laboratories  during screening process. For hematology, complete blood counts (CBC) including red blood cells (RBC), white blood cells (WBC) and platelets, quantitative analyses for hemoglobin, hematocrit, and other measurement are performed. For blood chemistry, serum is separated according to standard procedures, and quantitative analysis is performed for the following analytes: sodium, potassium, chloride, carbon dioxide, glucose, creatinine, and blood urine nitrogen (BUN) and urea. For liver function, albumin, total bilirubin, direct bilirubin, alanine aminotransferase (ALT/SGPT), aspartate aminotransferase (AST/SGOT), alkaline phosphatase (ALP), and total protein are measured. For endocrinology, TSH, total T3 and T4 are measured. All these laboratory measurements are captured as normal, abnormal not clinically significant, abnormal clinically significant, and not done. The data are tabulated by treatment groups based on the above categories. (Table 3.9.12).

### 3.9.4 Control Questionnaire Analysis

As described in 2.5.5, the questionnaire is designed to check the possibility that the blindness is bleached during the treatment. The survey is done before first

treatment session, after first treatment session, and end of the study by the local site investigator, the TMS treater, the study coordinator, and the participant. The distribution of the best guess of the treatment assignment and the confidence level of the guess will be tabulated in Table 3.9.13a-b by the treatment groups and tested by Pearson $\chi 2$ test using SAS PROC FREQ. For those with moderate or higher confidence level, the concordance of the true and the guessed treatment assignment will be performed using $\kappa$ statistic as shown in Table 3.9.14.

## 4.0  REFERENCES

Anonymous (2002): Refractory Depression: Current Strategies and Future Directions. In Nolen WA ZJ, Roose SP, Amsterdam JD (ed). New York: Wiley.

Avery DH, Claypoole K, Robinson L, et al. (1999): Repetitive transcranial magnetic stimulation in the treatment of medication-resistant depression: preliminary data. J Nerv Ment Dis 187:114-7.

Beck AT, Steer RA, Garbin MG J (1988). "Psychometric properties of the Beck Depression Inventory Twenty-five years of evaluation". Clin. Psych. Review 8: 77–100.

Beck AT, Steer RA (1991): Manual for Beck scale for suicide ideation. San Antonio, TX: Psychological Corporation.

Beck A, Weissman A, Lester D, Trexler L. (1974). The measurement of pessimism: the hopelessness scale. Journal of Consulting and Clinical Psychology 42:861-865.

Benton, A.L., Sivan, A. B., Hamsher, K. deS., Varney, N. R., & Spreen, O. (1994).Contributions to neuropsychological assessment (2nd ed.). Orlando, FL: Psychological Assessment Resources.

Blair, JR. & Spreen, O. (1989). Predicting premorbid IQ: A revision of the National Adult Reading Test. The Clinical Neuropsychologist 3, 129-136.

Blake DD, Weathers FW, Nagy LM, et al. (1995). The development of a Clinician-Administered PTSD Scale. Journal of Traumatic Stress, 8:75-90.

Blanchard EB, Jones-Alexander J, Buckley TC, Forneris CA. (1996).  Psychometric properties of the PTSD Checklist (PCL).  Behav Res Ther 34:669-73.

Breslow, N. E. and Clayton, D. G. (1993) Approximate Inference in Generalized Linear Mixed Models *JASA*  88, 9–25.

First M, Spitzer R, Gibbon M, Williams J (2002): Structured Clinical Interview for DSM-IV-TR Axis I Disorders, Research Version, Non-Patient Edition. (SCID-I/NP). New York: Biometrics Research, New York State Psychiatric Institute.

George MS, Wassermann EM, Kimbrell TA, et al. (1997): Mood improvement following daily left prefrontal repetitive transcranial magnetic stimulation in patients with depression: a placebo-controlled crossover trial. Am J Psychiatry 154:1752-6.

George MS, Lisanby SH, Avery D, McDonald WM, Durkalski V, Pavlicova M, Anderson B, Nahas Z, Bulow P, Zarkowski P, Holtzheimer PE 3rd, Schwartz T, Sackeim HA. (2010) Daily left prefrontal transcranial magnetic stimulation therapy for major depressive disorder: a sham-controlled randomized trial. Arch Gen Psychiatry. 67(5):507-16

Golden, CJ. (1978). Stroop Color and Word Test: A manual for clinical and experimental uses. Chicago, IL: Stoelting Co.

Green BL (1996): Trauma History Questionnaire. In Stamm BH (Ed.), Measurement of stress, trauma, and adaptation (pp. 366-369). Lutherville, MD: Sidran Press.

Herrmann, N.; Black, S. E.; Lawrence, J.; Szekely, C.; Szalai, J. P. (1998). "The Sunnybrook Stroke Study : A Prospective Study of Depressive Symptoms and Functional Outcome". Stroke 29 (3): 618–624.

Kaplan, E. L.; Meier, P (1958) Nonparametric estimation from incomplete observations. *J. Amer. Statist. Assn.* 53:457–481

Kaplan H, Sadock B (1996): Concise Textbook of Clinical Psychiatry. Baltimore: William & Wilkins. Validation of a short orientation-memory-concentration test of cognitive impairment.  <u>Am Journal of Psychiatry</u>,140, 734-739.

Katzman R, Brown T, Fuld P, Peck A, Schechter R, and Schimmel H (1983): Validation of a short orientation-memory-concentration test of congestive impairment American Journal of Psychiatry 140(6): 734-739

Keller MB, Lavori PW, Mueller TI, et al. (1992): Time to recovery, chronicity, and levels of psychopathology in major depression. A 5-year prospective follow-up of 431 subjects. Arch Gen Psychiatry 49:809-16.

Kozel FA, George MS, Simpson KN (2004): Decision analysis of the cost-effectiveness of repetitive transcranial magnetic stimulation versus electroconvulsive therapy for treatment of nonpsychotic severe depression. CNS Spectr 9:476-82.

Laird, Nan M.; Ware, James H. (1982). Random-Effects Models for Longitudinal Data *Biometrics*  38 (4): 963–974

Lezak, MD, Howieson, DB, Loring, DW, Hannay, HJ, & Fischer, JS (2004).  Neuropsychological Assessment (4th ed.). New York: Oxford University Press.

Little JT, Kimbrell TA, Wassermann EM, et al. (2000): Cognitive effects of 1- and 20-hertz repetitive transcranial magnetic stimulation in depression: preliminary report. Neuropsychiatry Neuropsychol Behav Neurol 13:119-24.

Loo C, Mitchell P, Sachdev P, et al. (1999): Double-blind controlled investigation of transcranial magnetic stimulation for the treatment of resistant major depression. Am J Psychiatry 156:946-8.

Mantel N (1966). Evaluation of survival data and two new rank order statistics arising in its consideration *Cancer Chemotherapy Reports* **50** (3): 163–70=

Martin JL, Barbanoj MJ, Schlaepfer TE, et al. (2003): Repetitive transcranial magnetic stimulation for the treatment of depression. Systematic review and meta-analysis. Br J Psychiatry 182:480-91.

McCullagh, P. and Nelder, J.A. (1989), *Generalized Linear Models*, London: Chapman Hall.

Montgomery SA, Asberg M (1979). "A new depression scale designed to be sensitive to change". British Journal of Psychiatry 134 (4): 382–89

Olfson M, Marcus S, Sackeim HA, et al. (1998): Use of ECT for the inpatient treatment of recurrent major depression. Am J Psychiatry 155:22-9.

Pascual-Leone A, Houser CM, Reese K, et al. (1993): Safety of rapid-rate transcranial magnetic stimulation in normal volunteers. Electroencephalogr Clin Neurophysiol 89:120-30.

Reitan, R. M. & Wolfson, D. (1993). The Halstead-Reitan Neuropsychological Test Battery: Theory and Clinical Interpretation. (2nd ed.). Neuropsychological Press, Tucson, AZ.

Rey, A. (1964). L'examen clinique en psychologie. Presses Universitaires de France: Paris.

Rubin, D. B. (1987). Multiple Imputation for Nonresponse in Surveys. New York: John Wiley & Sons.

Rush AJ, Trivedi MH, Ibrahim HM, et al. (2003): The 16-Item Quick Inventory of Depressive Symptomatology (QIDS), clinician rating (QIDS-C), and self-report (QIDS-SR): a psychometric evaluation in patients with chronic major depression. Biol Psychiatry. 54:573-83.

Rush AJ, Kraemer HC, Sackeim HA, et al. (2006): Report by the ACNP Task Force on response and remission in major depressive disorder. Neuropsychopharmacology. 31:1841-53.

Sackeim HA, Prudic J, Devanand DP, et al. (1990): The impact of medication resistance and continuation pharmacotherapy on relapse following response to electroconvulsive therapy in major depression. J Clin Psychopharmacol 10:96-104.

Skinner H A, Sheu W J (1982): Reliability of alcohol use indices. The Lifetime Drinking History and the MAST. Journal of Studies on Alcohol 43:1157-70.

Smith, A. (1991). Symbol Digit Modalities Test. Los Angeles, CA: Western Psychological Services.

Sobell L C, Sobell M B (1990):Self-reports issues in alcohol abuse: State of the art and future directions. Behavioral Assessment 12:77-90.

Sobell L C, Sobell M B, Riley D M, et al. (1988): The reliability of alcohol abusers' self-reports of drinking and life events that occurred in the distant past. Journal of Studies on Alcohol 49:225-32.
Spreen, O. & Strauss, E. (2006). A Compendium of Neuropsychologica Tests: Administration, Norms and Commendary. (3rd ed.). Oxford University Press, New York.

Thase ME (2004): Therapeutic alternatives for difficult-to-treat depression: a narrative review of the state of the evidence. CNS Spectr 9:808-16, 818-21.

Triggs WJ, McCoy KJ, Greer R, et al. (1999): Effects of left frontal transcranial magnetic stimulation on depressed mood, cognition, and corticomotor threshold. Biol Psychiatry 45:1440-6.

Ware, J. E., Kosinski, M., & Keller, S. D. (1994). SF-36 physical and mental summary scales: A user's manual. Boston, MA: The Health Institute.

Wassermann EM (1997): Report on risk and safety of repetitive transcranial magnetic stimulation (rTMS): Suggested guidelines from the International Workshop on Risk and Safety of rTMS (June 1996). Electroencephalo Clin Neuro. 108:1-16

Weathers F W, Keane T M, Davidson J R (2001): Clinician-administered PTSD scale: a review of the first ten years of research. Depression and anxiety 13:132-56.

Wilcoxon, Frank (1945). "Individual comparisons by ranking methods". *Biometrics Bulletin* **1** (6): 80–83.

Wolfe J, Kimerling R (1997): Gender issues in the assessment of posttraumatic stress disorder.  In J. P. Wilson & T. M. Keane (Eds.), Assessing psychology trauma and PTSD (pp. 192-238). New York, NY: Guilford Press.

## 5.0  APPENDICES

**Appendix 5.1 Study Schema**

**Screening 2-4 Weeks**

Sign Consent Form

↓

Screening Assessments

↓

Randomization

**Acute Treatment Phase 4 - 11 Weeks**

Active rTMS  (x20 Sessions)

Sham rTMS  (x20 Sessions)

Not Remitted (HRSD>10)

Remitted (HRSD≤10)

Remitted (HRSD≤10)

Not Remitted (HRSD>10)

↓

Additional Active rTMS 5-10 Sessions Max

Additional Sham rTMS 5-10 Sessions Max

**Follow-up Phase 24 Weeks**

Not Remitted (HRSD>10)

Remitted (HRSD≤10)

Remitted (HRSD≤10)

Not Remitted (HRSD>10)

Follow-up 24 Weeks No Taper

Treatment Taper and Follow-up 24 Weeks

Follow-up 24 Weeks No Taper

## Appendix 5.2 Schedule of Assessments

| | | 2-4 weeks | Acute Treatment Phase 4 - 11 weeks End of Session Number [1] | | | | | | Follow-up Phase 24 weeks (6 months) PRN Taper weeks | | | Weeks | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Assessment | | Screening/ Baseline | 5 | 10 | 15 | 20 | [25] | [30/ last] | 1 | 2 | 3 | 4 | 8 | 12 | 16 | 20 | 24 |
| 86 | Consent | S* | | | | | | | | | | | | | | | |
| 01 | Screening Form | S | | | | | | | | | | | | | | | |
| 02 | Randomization Form | B* | | | | | | | | | | | | | | | |
| 03 | Baseline | S | | | | | | | | | | | | | | | |
| 04 | Medical History | S | | | | | | | | | | | | | | | |
| 05 | Physical Exam | S | | | | | | x | | | | | | | | | |
| 06 | Labs | S | | | | | | | | | | | | | | | |
| 07 | Structured Clinical Interview for DSM-IV-TR (SCID-I) | S | | | | | | | | | | | | | | | |
| 08 | Current/Past ATHF | S | | | | | | | | | | | | | | | |
| 09 | Lifetime Drinking History | B | | | | | | | | | | | | | | | |
| 10 | Clinician Administered PTSD Scale (CAPS) | B | | | | | | x | | | | | | | | | x |
| 11 | Trauma History Questionnaire (THQ) | B | | | | | | | | | | | | | | | |
| 12 | Life Stressor Checklist-revised (LSC-R) | B | | | | | | | | | | | | | | | |
| 13 | Six-Item Blessed Orientation-Memory-Concentration (BOMC) | S | | | | | | | | | | | | | | | |
| 14 | Pregnancy Test | $S^2$ | Every 4 weeks | | | | | | | | | | | | | | |
| 15 | Medication Use | S | x | x | x | x | x | x | x | x | x | x | x | x | x | x | x |
| 16 | Study Visit Form | | x | x | x | x | x | x | x | x | x | x | x | x | x | x | x |
| 17 | Audiometry | S | | | | | | x | | | | | | | | | x |
| 18 | rTMS Treatment Log | | x | x | x | x | x | x | | | | | | | | | |
| 19 | rTMS Taper Log | | | | | | | | x | x | x | | | | | | |
| 20 | Hamilton Rating Scale for Depression (HRSD) and MADRS | $S^2$ | x | x | x | x | x | x | | | | x | x | x | x | x | x |
| 21 | Beck Depression Inventory (BDI) | B | x | x | x | x | x | x | | | | x | x | x | x | x | x |
| 22 | Quick Inventory of Depressive Symptomatology (QIDS-C16) | B | x | x | x | x | x | x | | | | x | x | x | x | x | x |
| 23 | Columbia-Suicide Severity Rating Scale (C-SSRS) | S | | | | | | | | | | | | | | | |
| 24 | Columbia-Suicide Severity Rating Scale (C-SSRS) F/U | | x | x | x | x | x | x | x | x | x | x | x | x | x | x | x |
| 25 | Beck Scale for Suicidal Ideation (BSS) | S | x | x | x | x | x | x | x | x | x | x | x | x | x | x | x |

| | Assessment | 2-4 weeks Screening/ Baseline | Acute Treatment Phase — 4 - 11 weeks — End of Session Number [1] — 5 | 10 | 15 | 20 | [25] | [30/last] | Follow-up Phase — 24 weeks (6 months) — PRN Taper weeks — 1 | 2 | 3 | Weeks — 4 | 8 | 12 | 16 | 20 | 24 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 26 | Beck Hopelessness Scale | S | x | x | x | x | x | x | x | x | x | x | x | x | x | x | x |
| 27 | Quality of Life (VR 36) | B | | | | | | x | | | | | | | | | x |
| 28 | Neuropsychological Battery | B | | | | | | x | | | | | | | | | x |
| 29 | DAST | S | | | | | | x | | | | | | | | | x |
| 30 | PTSD Checklist (PCL) | B | | | | | | x | | | | | | | | | x |
| 31 | MAST | S | | | | | | x | | | | | | | | | x |
| 32 | STAXI-2 | S | x | x | x | x | x | x | x | x | x | x | x | x | x | x | x |
| TLFB | Alcohol/Drug Timeline Followback (TLFB) | | x | x | x | x | x | x | x | x | x | x | x | x | x | x | x |
| 33 | Urine Tox Screen/Alcohol Test | S | | x | | x | | X | | x | | x | | x | | x | |
| 34 | Termination Form [3] | | | | | | | | | | | | | | | | x |
| 35 | Control Questionnaire | Before and after the 1st treatment | | | | | | | | | | | | | | | x |
| 36 | Protocol Deviation | As required | | | | | | | | | | | | | | | |
| 37 | Adverse Events (AE) and ADE and Serious Adverse Events (SAEs) and UADE | As required | | | | | | | | | | | | | | | |

**\*B= Baseline**
**\*S= Screening**

[1]  Sessions 21-25 (block 5) and 26-30 (block 6) may not be required if patient goes into remission earlier

[2]  Must be conducted within 7 days prior to randomization

[3]  Termination Form will be completed at the end of the study OR when a patient decides to end study participation prior to the study completion date; i.e., a patient decides to withdraw or leave the study for any reason, withdraws consent, or is suspended from the study.

## 6.0 TABLE SHELLS AND FIGURE LIST FOR DATA PRESENTATION

(SEE ATTACHED DOCUMENT)