ASSESSING TREATMENT EFFECT HETEROGENEITY IN THE CITALOPRAM FOR AGITATION IN ALZHEIMER'S DISEASE CLINICAL TRIAL: A SUBGROUP ANALYSIS TO GUIDE PERSONALIZED TREATMENT SELECTIONS

by

Lisa E. Rein

A thesis submitted to Johns Hopkins University

in conformity with the requirements for the degree of

Master of Science

Baltimore, Maryland

April 2014

Abstract

Alzheimer's disease (AD) is a devastating neurodegenerative condition with symptoms of cognitive decline, behavioral disturbances, and ultimately mortality. As there is currently no cure, improvements for management of AD symptoms are desperately needed. The Citalopram for Agitation in Alzheimer's Disease (CitAD) study examined off-label use of citalopram, a selective serotonin reuptake inhibitor, for management of agitation symptoms. The primary analysis showed a greater average decrease in agitation symptoms and an increase in a potentially serious adverse event with citalopram compared to placebo. Physicians want to know if the treatment effect is heterogeneous, and if so, which patients have the greatest potential to benefit; given the risks, it may be unethical to prescribe the drug to patients with little chance of benefit. Subgroup analyses are employed to assess heterogeneity of effect across subgroups defined by categorical baseline covariates. This is typically done by calculating subgroup treatment effects in a stratified dataset and testing for the interaction between treatment and the baseline covariate. This approach is not very comprehensive, as it only examines one covariate at a time while patients hold multiple characteristics simultaneously. Another limitation of this approach is the use of parametric models which carry the assumption of correct mean model specification. For our subgroup analysis, we employed the two-stage estimation method developed by Cai et al. [3]. In the first stage, parametric working models are used to calculate

the approximate treatment effect, the difference in potential outcomes under citalopram versus placebo, for each participant based on multiple baseline covariates. This predicted treatment effect is called the index score; patients with the same combination of baseline covariates have the same index score and are considered a subgroup. In the second stage, patients are grouped by index score allowing non-parametric estimation of subgroup treatment effects using observed data. Using this approach, we found evidence for treatment effect heterogeneity. CitAD participants with the largest predicted treatment effects were more likely to be living outside long-term care facilities, within the middle age range (ages 76-82), with minimal cognitive impairment (MMSE 21-30), within the middle baseline agitation range (NBRS-A 6-8), and not taking lorazepam.

ScM Thesis Committee:

Dr. Constantine Frangakis (Thesis Advisor)
Professor, Department of Biostatistics

Dr. Jeannie-Marie Leoutsakos (Thesis Reader)
Assistant Professor, Department of Mental Health

Acknowledgements

I would like to express my sincerest gratitude to my advisor, Dr. Constantine Frangakis, for his guidance in completing this project. I would like to thank Dr. Jeannie-Marie Leoutsakos for her comments and encouragement. I would also like to thank Dr. Constantine Lyketsos for this research opportunity - it has been a wonderful learning experience. Many thanks also to Dr. Lon Schneider, Dr. Lea Drye, Dr. David Shade, and the rest of the CitAD team for their insight and assistance.

Lastly, I would like to thank my husband, Alex Rein, and my parents, Jim and Nancy Egner, for their support.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Alzheimer's disease (AD) is a devastating neurodegenerative disease which currently has no cure. The majority of people with AD are 65 or older; less than 5% of people with the disease have "early onset", characterized by diagnosis before age 65 [1]. AD is widely prevalent and will increase with the growing elderly population in the developed and developing worlds. The estimated prevalence of AD was 5.3 million in the United States in 2010 [1]. Worldwide, roughly 18 million are living with the disease [13]. The worldwide prevalence is projected to reach 80 million by the year 2050 [7]. Without a cure, improvements in patient care and management of AD symptoms are desperately needed to accommodate our aging population.

Alzheimer's disease is a type of dementia. Dementia is a clinical syndrome characterized by symptoms of severe cognitive and functional impairment [6]. Dementia can be caused by an event like a stroke, or can be the manifestation of an underlying disease like Parkinson's, Creutzfeldt-Jakob, or Alzheimer's [1]. AD is by far the largest attributable cause of dementia, accounting for an estimated 60-80% of all cases [1]. The precise etiology of AD is unknown but is thought to have both genetic and environmental components [5] [9]. Cognitive symptoms of AD include memory loss, confusion, difficulty recognizing people or places, impaired judgment, and

problems with written and verbal communication [1]. In addition to cognitive impairment, behavioral and psychological problems are also frequent among AD patients. These symptoms can include verbal and physical aggression, agitation, anxiety, hallucinations, paranoid delusions, depression, and other mood disorders [2]. It has been estimated that more than 90% of AD patients will develop at least one behavioral or psychological symptom within a 5 year period [2].

These behavioral and psychological symptoms of AD can be particularly difficult for the patient and burdensome for the family and caregivers. Behavioral symptoms are often the impetus for moving AD patients into long term care facilities [14]. Current pharmacological treatments for AD symptoms of agitation and aggression include antipsychotics, cholinesterase inhibitors, memantine, antidepressants, and anticonvulsants [2]. Antipsychotics have been widely used in the past, however recent findings suggest limited efficacy and increased risk of serious adverse events including mortality [2]. Alternatives such as antidepressants are encouraging, but are still lacking data regarding safety and efficacy for this population. Citalopram is one such antidepressant, a selective serotonin reuptake inhibitor (SSRI), which has been suggested as an alternative for the treatment of agitation in AD patients.

1.1 The CitAD clinical trial

The Citalopram for Agitation in Alzheimer's Disease (CitAD) study design was previously described in detail by Drye et al. [4]. In summary, the CitAD study is a randomized, multi-center, parallel, placebo controlled, double blinded clinical trial. There were a total of eight clinical sites, located within the United States and Canada. Consenting participants with probable Alzheimer's disease and without depression were randomized to either citalopram or placebo at a ratio of 1:1, stratified by center. Each patient had a caregiver who also participated in the study by accompanying the patient to all visits and providing information for patient evaluation. All caregivers

received a structured psychosocial therapy which consisted of counseling sessions with a trained study clinician. Patients were followed for a total of 9 weeks with assessments made at enrollment and follow-up after 3, 6, and 9 weeks. Baseline covariates including demographic information, severity of symptoms, and use of concomitant medications were recorded for all participants at the enrollment visit. The primary efficacy outcomes were agitation as measured by the modified Alzheimer Disease Cooperative Study-Clinical Global Impression of Change (mADCS-CGIC) and the Neurobehavioral Rating Scale agitation subscale (NBRS-A). Secondary efficacy outcomes include cognition, functional impairment, mobility, and psychiatric symptoms. There were 186 total participants with 92 randomized to placebo and 94 randomized to citalopram. Loss to follow-up was relatively low; the week 9 primary efficacy outcomes are available for 167 patients (81 randomized to placebo and 86 randomized to citalopram).

1.2 Purpose

The purpose of this project is to examine treatment effect heterogeneity across participants in the CitAD clinical trial and to identify potential predictors of citalopram response. The primary safety and efficacy results of the CitAD trial were previously presented by Porsteinsson et al. [12]. Patients receiving citalopram showed a significantly greater improvement in symptoms of agitation compared to those receiving placebo, as measured by both primary outcomes; the difference in average change in NBRS-A (week 9 - baseline) was -0.93 [-1.80, -0.06] favoring citalopram, and the odds of improvement according to the mADCS-CGIC were 2.13 [1.23, 3.69] times larger in the citalpram arm. The citalopram arm also showed a significant increase in a potentially serious cardiovascular adverse event compared to placebo. Citalopram was associated with a 18.1 [6.1, 30.1] millisecond greater increase in QTc interval than placebo; 3 of 94 patients in the citalopram arm and 1 of 92 patients in the placebo

arm experienced clinical QTc prolongation.

Although the average treatment effect estimate is informative, physicians ultimately need to make treatment decisions on a patient-by-patient basis. Our post-hoc subgroup analysis is motivated by the physician's need to identify candidates who have the most potential to benefit for prescribing purposes. Given the risks, it would be unethical to prescribe citalopram to patients who have little chance of benefit. However, the positive findings suggest that this drug may still be a good option for a subset of Alzheimer's patients. In this thesis, we propose methods for assessing heterogeneity of affect and for characterizing the subset or subsets or patients who would be most likely to benefit from treatment with citalopram.

1.3 Research questions

The CitAD study lends itself to the following three research questions:

1. What is the average treatment effect?
2. Are there subgroups of people for whom the treatment effect is larger than the estimated average effect?
3. Is there evidence that there are groups that have different true average effects?

1.4 Subgroup analyses

Subgroup analyses are very common in randomized clinical trials. A self-study of the New England Journal of Medicine found that 59 out of 97 total trials published from July 2005 to July 2006 reported some form of subgroup analyses [8]. Similarly, another study showed that 35 out of 50 trials sampled from four major medical journals in 1997 reported results from subgroup analyses [11]. Subgroup analyses appeal to the current trends towards evidence based medicine. There may be true differences in the risk or benefit of a treatment within different subgroups of people; knowing these

differences would be very helpful to physicians in planning a course of treatment for a particular patient. Typical subgroup analyses involve using parametric regression models to look for interactions between treatment and each baseline covariate of interest.

Subgroup analyses are often criticized for lack of power and high false positive rates among reported subgroup effects. The sample size for a clinical trial is determined by the number of participants needed to detect a minimum relevant effect size for a given power. The power to detect the true treatment effect is greatly reduced after dividing the total sample into smaller subgroups. Tests of interaction, such as likelihood ratio tests, are recommended to evaluate heterogeneity among levels of a baseline covariate [16] [11]. Multiplicity is also of great concern in subgroup analyses. The null hypothesis of no treatment effect is tested multiple times, once for each subgroup of interest, inflating the overall type I error rate. It is important to correct for multiplicity in the analysis and also to limit the choice of subgroups to those which have the most a priori biological or clinical support. If possible, possible subgroups should be defined before the trial has begun to avoid the discovery and reporting of spurious subgroup effects which have no biological basis.

Another limitation of the typical approach to subgroup analyses is the separate evaluation of each baseline covariate. Subgroups are typically defined as the presence of a single covariate level such as male or female gender. We would like to estimate the treatment effect for a patient who carries several prognostic baseline traits simultaneously. One could consider each combination of baseline covariate levels as a separate subgroup which is more realistic, however this becomes complicated as the number of subgroups increases dramatically. An additional limitation of typical subgroup analysis approaches is the use of parametric models which require numerous distributional assumptions. Ideally, we would like to relax these assumptions and use a non-parametric approach which relies more on the observed data.

Cai, Tian, Wong, and Wei introduced a novel two-stage subgroup analysis method

which addresses both the need to evaluate combinations of several baseline covariates simultaneously and the desire to obtain empirical estimates of the subgroup treatment effects [3]. In the first stage, a parametric working model is chosen to describe the approximate relationship between the outcome and several baseline covariates which are thought to be prognostic. This parametric framework is used to generate an index score for each patient which is the predicted treatment effect according to the working models. Each index score can be considered its own subgroup corresponding to a specific combination of baseline covariates. In the second stage, the index scores are used to group participants in a way that allows the use of non-parametric methods to ultimately estimate the subgroup treatment effects. We applied this two-stage estimation procedure (with several modifications) to identify baseline predictors of citalopram response and assess heterogeneity of treatment effect.

1.5 Summary

This thesis focuses on the application of the two-stage estimation procedure to the CitAD study. The methods span several chapters. Chapter 2 includes a description of the data and the traditional bivariate subgroup analysis methods. Chapter 3 provides a step-by-step illustration of the two-stage estimation procedure introduced by Cai et al [3]. Chapter 4 includes description of the hypothesis tests used to assess overall heterogeneity of effect. Results specifically addressing each research question are presented and discussed in Chapter 5. Concluding remarks are made in Chapter 6 including implications of these results for clinical practice and recommendations for future research.

Chapter 2

Data Description & Traditional Subgroup Analyses

2.1 Data

2.1.1 Primary outcomes

The first of two primary outcomes of the CitAD trial was agitation as measured by the modified Alzheimer's Disease Cooperative Study-Clinical Global Impression of Change (mADCS-CGIC). The mADCS-CGIC is a measure of how each participant's agitation symptoms at follow-up compare to his or her baseline symptoms, as evaluated by the study clinician. The mADCS-CGIC score ranges from 1-7 as follows:

1. Marked improvement
2. Moderate improvement
3. Minimal improvement
4. No change
5. Minimal worsening
6. Moderate worsening
7. Marked worsening

The study investigators are particularly interested in predicting which patients will achieve either marked or moderate improvement for prescribing purposes. We have thus created a mADCS-CGIC response variable which is an indicator for either marked or moderate improvement (mADCS-CGIC 1 or 2) at week 9; this binary variable is the outcome we used for the subgroup analyses. The sample size is 167 participants (86 randomized to citalopram and 81 randomized to placebo) for which we have mADCS-CGIC response at week 9 and baseline covariate information. Of these participants, 55 (34 randomized to citaloram and 21 randomized to placebo) had a mADCS-CGIC response of either marked or moderate improvement at week 9.

The second primary outcome was agitation as measured by the Neurobehavioral Rating Scale agitation subscale (NBRS-A). The NBRS-A subscale ranges from 0-18 where larger scores indicate more severe symptoms. In the CitAD study, baseline NBRS-A scores ranged from 1-14, with mean and median scores of 7.6 and 8 respectively. The NBRS-A scores at week 9 ranged from 0-16, with mean and median scores of 4.7 and 4 respectively. A reduction of 50% from baseline NBRS-A is considered a clinically relevant response by the study investigators. Of the available 167 participants at week 9, 75 (48 randomized to citalopram and 27 randomized to placebo) had at least 50% reduction in NBRS-A from baseline.

Our subgroup analyses focused on predicting mADCS-CGIC response rather than NBRS-A improvement, as the study investigators feel that mADCS-CGIC response is the more clinically relevant outcome. We have however, evaluated our predictive model (based on mADCS-CGIC response) using both measures for comparison. We would expect to observe any true subgroup effects consistently across closely related outcomes [15].

2.1.2 Pre-specified subgroups

There were five baseline covariates which were pre-specified for subgroup analyses in the CitAD protocol. The pre-specified baseline covariates are residency, presence

of delusions and/or hallucinations, functional impairment, cognition, and agitation. Residency was recorded at baseline as whether the patient resided in his own home, a caregivers' home, assisted living, or nursing facility. For the subgroup analysis, we made a binary indicator for living in a long term care facility (including either assisted living or nursing facilities). It should be noted that very few participants from long term care facilities enrolled in the study (13 total, 12 of which were followed through week 9). The Neuropsychiatric Inventory (NPI) is a survey completed by the caregiver which assesses neuropsychiatric and behavioral symptoms. To assess the presence of delusions and hallucinations, we used the NPI Delusions (NPI-D) and NPI Hallucinations (NPI-H) subscales; each scale ranges from 0-12 where higher scores indicate more severe symptoms. We created a binary indicator for presence of hallucinations (NPI-H > 0) and/or delusions (NPI-D > 0) at baseline. Functional impairment was measured by the Activities of Daily Living Inventory (ADL) survey as completed by the caregiver. The ADL score ranges from 0-78 where higher scores indicate less functional impairment. We created three subgroups defined by tertiles of baseline ADL scores. Cognition was measured by the Mini Mental State Examination (MMSE) which is a test administered by study personnel. MMSE scores range from 0-30 where higher scores indicate better functioning. We separated the participants into three groups (mild to no impairment, moderate, and severe impairment) based on MMSE cutoffs from the literature [10]. Baseline agitation was measured by the Neurobehavioral Rating Scale agitation subscale (NBRS-A). The NBRS-A ranges from 0-18 where higher scores indicate more severe symptoms. We created three subgroups defined by tertiles of the baseline NBRS-A scores. The number and percentages of participants in each pre-specified subgroup are provided in Table 2.1.

	Total		Citalopram		Placebo	
Total Randomized	186		94		92	
Residence						
Home or relative	173	(93%)	86	(91%)	87	(95%)
Long term care	13	(7%)	8	(9%)	5	(5%)
Neuropsychiatric Inventory (NPI)						
No hallucinations nor delusions	97	(52%)	52	(55%)	45	(49%)
Hallucinations and/or delusions	89	(48%)	42	(45%)	47	(51%)
Activities of Daily Living (ADL)						
Largest tertile, 54-74	65	(35%)	39	(41%)	26	(28%)
Middle tertile, 31-53	64	(34%)	28	(30%)	36	(39%)
Smallest tertile, 6-30	57	(31%)	27	(29%)	30	(33%)
Mini-Mental State Examination (MMSE)						
Mild to no impairment, 21-30	54	(29%)	32	(34%)	22	(24%)
Moderate, 11-20	81	(44%)	45	(48%)	36	(39%)
Severe, 0-10	51	(27%)	17	(18%)	34	(37%)
Neurobehavioral Rating Scale Agitation Subscore (NBRS-A)						
Smallest tertile, 1-5	52	(28%)	29	(31%)	23	(25%)
Middle tertile, 6-8	60	(32%)	34	(36%)	26	(28%)
Largest tertile, 9-14	74	(40%)	31	(33%)	43	(47%)

Table 2.1: Summary of pre-specified baseline subgroups

2.1.3 Post-hoc subgroups

An additional six baseline covariates were selected for the subgroup analysis. These post-hoc covariates are age, gender, and use of memantine, lorazepam, trazodone, and cholinesterase inhibitors. Participant ages ranged from 47 to 92. Three age subgroups were defined by tertiles of baseline age. Gender consists of male and female subgroups; participants were roughly evenly divided between males and females. The remaining post-hoc subgroups are indicators for the use of concomitant Alzheimer's medications within three weeks of enrollment. Use of these medications at enrollment may serve as a proxy for severity of disease. These medications are typically used in severe cases, especially lorazepam and trazodone which were also used as the rescue drugs throughout the trial. The memantine, lorazepam, trazodone, and cholinesterase inhibitor covariates each include two subgroups, one for users of each medication and another for non-users. Users of cholinesterase inhibitors include patients who have used donepezil, rivastigimine, and/or galatamine within three weeks of enrollment. The number and percentages of participants in each post-hoc subgroup are provided in Table 2.2.

	Total		Citalopram		Placebo	
Total Randomized	186		94		92	
Age						
Smallest tertile, 47-75	57	(31%)	30	(32%)	27	(29%)
Middle tertile, 76-82	60	(32%)	29	(31%)	31	(34%)
Largest tertile, 83-92	69	(37%)	35	(37%)	34	(37%)
Gender						
Male	101	(54%)	50	(53%)	51	(55%)
Female	85	(46%)	44	(47%)	41	(45%)
Memantine						
No memantine use	108	(58%)	53	(56%)	55	(60%)
Memantine use	78	(42%)	41	(44%)	37	(40%)
Lorazepam						
No lorazepam use	171	(92%)	88	(94%)	83	(90%)
Lorazepam use	15	(8%)	6	(6%)	9	(10%)
Trazodone						
No trazodone use	167	(90%)	83	(88%)	84	(91%)
Trazodone use	19	(10%)	11	(12%)	8	(9%)
Cholinesterase Inhibitors						
No cholinesterase inhibitor use	58	(31%)	32	(34%)	26	(28%)
Use of cholinesterase inhibitor(s)	128	(69%)	62	(66%)	66	(72%)

Table 2.2: Summary of post-hoc baseline subgroups

2.2 Traditional subgroup analyses

We performed a series of simple bivariate analyses as a first look at potential subgroup effects, and to later inform our covariate selections for the two-stage estimation procedure. We estimated the treatment effect, the odds ratio of mADCS-CGIC response in the citalopram group versus placebo group, within each subgroup using logistic regression. In each logistic regression model, the log odds of mADCS-CGIC response is regressed on treatment as a single predictor, as shown in Equation 2.1, where $E[Y|Z = z]$ is the probability of response for participants randomized to treatment Z taking values of 0 for placebo and 1 for citalopram. The coefficient β_1 is the log odds ratio of response on citalopram versus placebo.

$$\text{logit}\left(E[Y|Z = z]\right) = \beta_0 + \beta_1 z \tag{2.1}$$

The results are provided in Figures 2.1 and 2.2 in the form of forest plots. The forest plots provide the estimate and confidence intervals for the treatment effect (odds ratio of response on citalopram versus placebo) for each covariate stratum. Tests of interaction were done using likelihood ratio tests. The full model including the interaction between the baseline categorical variable and treatment was compared to the reduced model without this interaction. Equations 2.2 and 2.3 provide examples of the full and reduced models respectively for a two-level covariate, X taking values of 0 for the baseline level and 1 for the second level. The degrees of freedom for each interaction test is equal to the number of covariate levels (subgroups) minus one.

$$\text{logit}\left(E[Y|Z = z, X = x]\right) = \beta_0 + \beta_1 z + \beta_2 x + \beta_3 (z \times x) \tag{2.2}$$

$$\text{logit}\left(E[Y|Z = z, X = x]\right) = \beta_0 + \beta_1 z + \beta_2 x \tag{2.3}$$

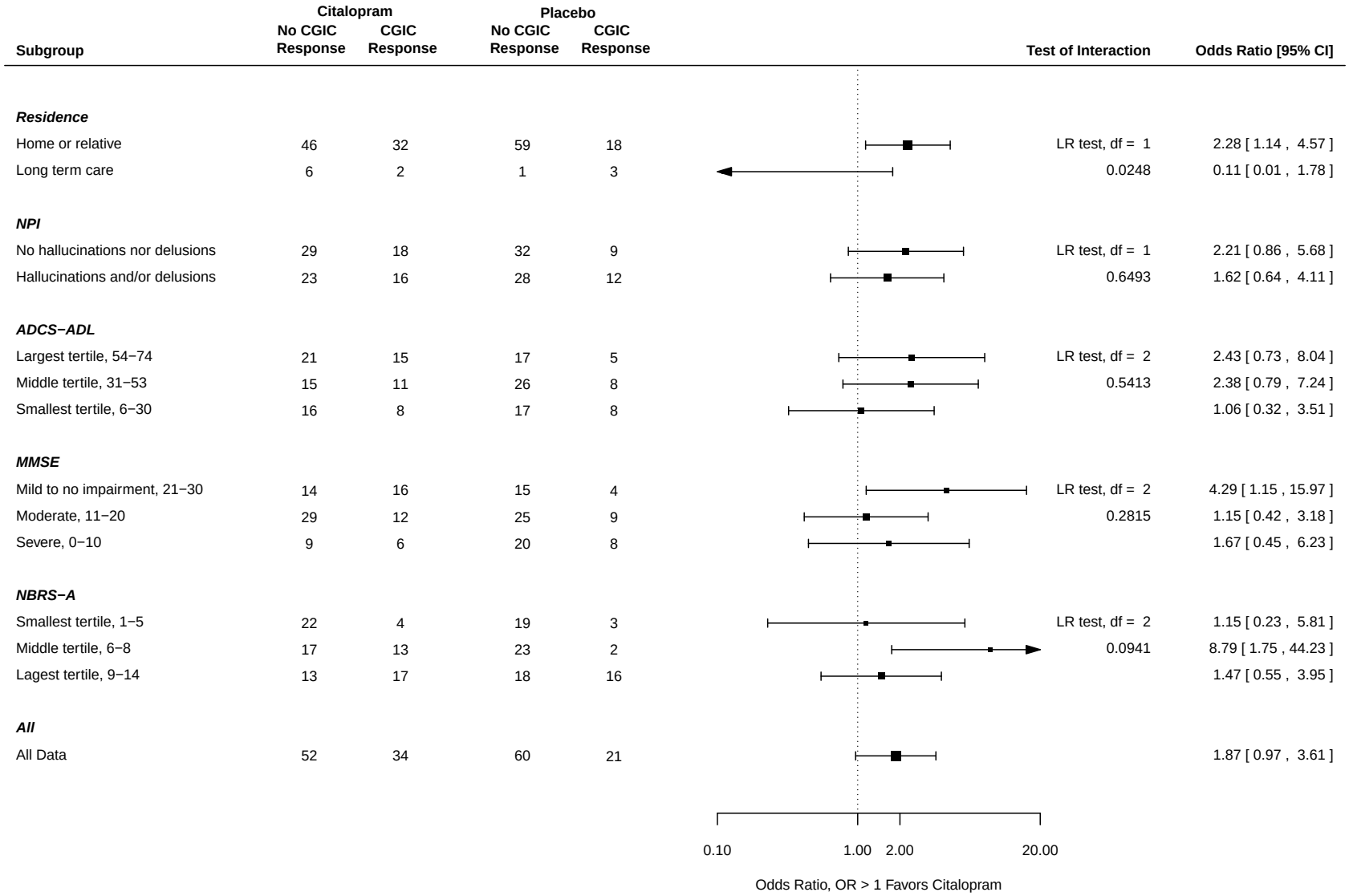

Figure 2.1: Bivariate subgroup analyses for pre-specified covariates

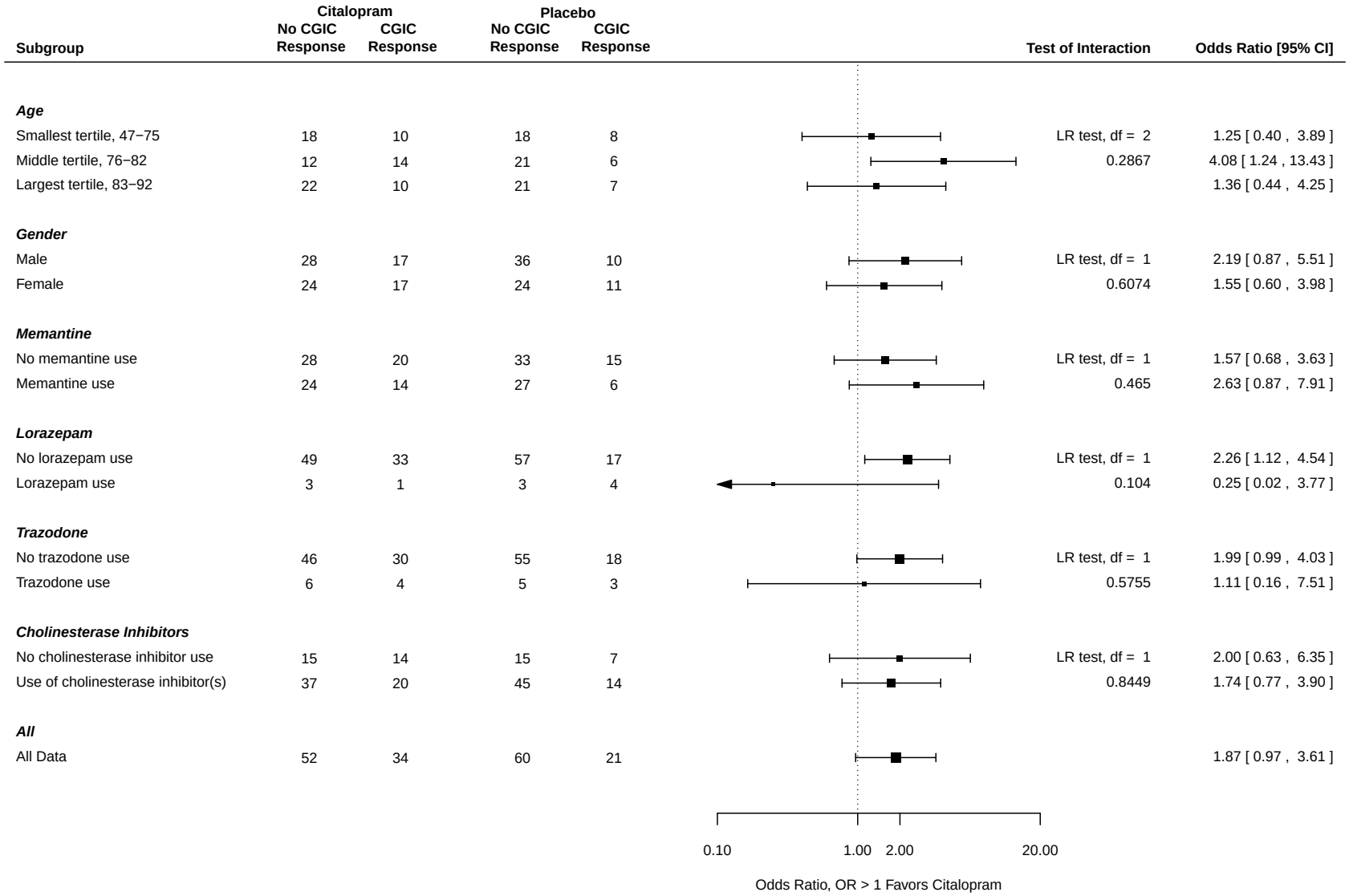

Figure 2.2: Bivariate subgroup analyses for post-hoc covariates

Chapter 3

Two-Stage Estimation Procedure

This two-stage estimation method was introduced by Cai, Tian, Wong, and Wei in Biostatistics, 2011 [3]. In the first stage, a parametric working model (such as a generalized linear model) is chosen as an approximation to the predictive true regression of the dependent variable (outcome) on covariates which are thought to be prognostic. The model is fitted separately for each treatment group using methods that assume it is correct, for example, using weighted least squares. An index score is then calculated for each patient as the difference in the fitted expected outcomes, under assignment to treatment versus control group. In the second stage, a non-parametric method is used to model the observed outcome as a function of the index score for each treatment group separately. The difference in the two non-parametric curves is a non-parametric estimate of the treatment effect for subgroups of patients with the same index scores. In this chapter, we illustrate each of these steps using the CitAD dataset. In addition to applying this procedure to our dataset, we have made several modifications which aide in interpretation of results and also provide a means for evaluating performance of this method.

3.1 Stage 1: Generating the index scores

3.1.1 Covariate selection

To generate the index scores, we first had to decide which prognostic factors to include in the working models. We selected the most promising baseline covariates based on the results from the traditional bivariate subgroup analyses (Figures 2.1 & 2.2). We decided to include baseline covariates for which the fold change in the treatment effect (odds ratio, citalopram versus placebo) for any two levels is greater than three. The five baseline covariates which meet this criteria are residency, MMSE, NBRS-A, age, and lorazepam.

3.1.2 Defining the index score

For each prospective patient, we would like to predict his or her treatment effect based on his or her baseline characteristics. In a causal inference framework, the true treatment effect is the difference in the patient's potential outcome if he or she were given citalopram and the potential outcome if he or she were given placebo. In mathematical terms, the treatment effect for a subgroup with the same baseline covariates is

$$s(\mathbf{x}) = E_{approx}\left[Y_{(1)}|\mathbf{X}=\mathbf{x}\right] - E_{approx}\left[Y_{(0)}|\mathbf{X}=\mathbf{x}\right] \tag{3.1}$$

$$= g_1(\boldsymbol{\beta_1}\mathbf{x}) - g_0(\boldsymbol{\beta_0}\mathbf{x}) \tag{3.2}$$

where $Y_{(1)}$ is the potential outcome if the patient were given citalopram, $Y_{(0)}$ is the potential outcome if the patient were given placebo, and $\mathbf{X}=\mathbf{x}$ is a vector of baseline covariates (or functions of these values). As shown in Equation 3.2, the expected value of each potential outcome can be approximated by a parametric model with

design matrix, $\mathbf{X}$=$\mathbf{x}$, and covariate matrix, $\boldsymbol{\beta}$. The functions g_1 and g_2 are smooth link functions which relate the expected potential outcome to the linear predictor.

As defined in Cai et. al [3], the index score is an estimate for the true subgroup treatment effect given by

$$\hat{s}(\mathbf{x}) = g_1(\hat{\boldsymbol{\beta}}_\mathbf{1}\mathbf{x}) - g_0(\hat{\boldsymbol{\beta}}_\mathbf{0}\mathbf{x}) \tag{3.3}$$

The vector of coefficients, $\hat{\boldsymbol{\beta}}_\mathbf{1}$, is estimated by fitting the working parametric model with data from all participants randomized to citalopram. Similarly, $\hat{\boldsymbol{\beta}}_\mathbf{0}$ is estimated by fitting the working model with data from all participants randomized to placebo. Each model is fitted separately for each treatment group using weighted least squares. The index score is the difference in the fitted values from the citalopram model and the placebo model, and is the predicted treatment effect for each person according to the working models. Each unique combination of baseline covariates corresponds to a unique index score. Persons with the same index score can be thought of as a subgroup. Persons with index scores which are close to one another have similar predicted treatment effects.

3.1.3 Fitting the working models

The form of the working model for the binary mADCS-CGIC response outcome is a logistic regression with five categorical predictors: residency, MMSE, NBRS-A, age, and lorazepam. The estimated regression coefficients and their standard error estimates are shown in Table 3.1 for citalopram and in Table 3.2 for placebo. These coefficients would be used to calculate the index score for a new patient given his or her baseline characteristics.

	Estimate	Std. Error	Z score	p-value
(Intercept)	-1.2657	0.7005	-1.81	0.0708
Residence: Long term care	-0.7566	0.9423	-0.80	0.4221
MMSE: Moderate, 11-20	-1.3868	0.6067	-2.29	0.0223
MMSE: Severe, 0-10	-1.3019	0.7900	-1.65	0.0994
NBRS-A: Middle tertile, 6-8	1.5236	0.7315	2.08	0.0373
NBRS-A: Largest tertile, 9-14	2.3663	0.7579	3.12	0.0018
Age: Middle tertile, 76-82	0.7359	0.6389	1.15	0.2494
Age: Largest tertile, 83-92	0.1930	0.6485	0.30	0.7660
Lorazepam: User	-0.0242	1.3041	-0.02	0.9852

Table 3.1: Citalopram working model coefficients

	Estimate	Std. Error	Z score	p-value
(Intercept)	-1.6780	0.9513	-1.76	0.0777
Residence: Long term care	3.4785	1.5643	2.22	0.0262
MMSE: Moderate, 11-20	0.0268	0.8604	0.03	0.9752
MMSE: Severe, 0-10	-0.6216	0.9321	-0.67	0.5048
NBRS-A: Middle tertile, 6-8	-0.7819	1.0874	-0.72	0.4721
NBRS-A: Largest tertile, 9-14	1.9738	0.8238	2.40	0.0166
Age: Middle tertile, 76-82	-0.3326	0.7423	-0.45	0.6541
Age: Largest tertile, 83-92	-1.1348	0.8402	-1.35	0.1768
Lorazepam: User	2.2256	1.2938	1.72	0.0854

Table 3.2: Placebo working model coefficients

Calculating the index score for CitAD participants

The next step of the procedure is to calculate an index score for each of the 167 CitAD participants. We have adjusted the calculation of the index score from that published in Cai et al. [3] to avoid possible overfitting. Overfitting is a concern because we use the participant's observed response to fit the working model for the group that the participant was assigned to in the trial. To avoid overfitting, we employ a leave-one-out approach; we fit the working model for the assigned group with data from all participants in that group except the participant for which the index score is being calculated. The working model for the group which the participant was not assigned is still fitted with the complete data from all participants in that group (same coefficients as either Table 3.1 or Table 3.2). For participant i assigned to citalopram,

the index score calculation is shown in Equation 3.4 where $\hat{\boldsymbol{\beta}}_{\mathbf{1},-\boldsymbol{i}}$ is the vector of estimated regression coefficients when the citalopram working model is fitted without the i^{th} participant. Similarly, the index score for participant i assigned to placebo is given in Equation 3.5.

$$\hat{s}(\mathbf{x}) = g_1(\hat{\boldsymbol{\beta}}_{\mathbf{1},-\boldsymbol{i}}\mathbf{x}) - g_0(\hat{\boldsymbol{\beta}}_{\mathbf{0}}\mathbf{x}) \tag{3.4}$$

$$\hat{s}(\mathbf{x}) = g_1(\hat{\boldsymbol{\beta}}_{\mathbf{1}}\mathbf{x}) - g_0(\hat{\boldsymbol{\beta}}_{\mathbf{0},-\boldsymbol{i}}\mathbf{x}) \tag{3.5}$$

There were 68 unique index scores populated among the 167 participants ranging from -0.920 to 0.693. Participants in the same treatment group with the same combination of baseline covariates received the same index score; as these participants are essentially exchangeable, leaving one out resulted in the same regression coefficient estimates and associated index score. The distribution of the index scores for the CitAD participants is shown in Figure 3.1.

3.2 Stage 2: Non-parametric estimation of the treatment effect

3.2.1 Visualizing the data

For each participant in the trial, we now have an index score, the observed mADCS-CGIC (0 for non-response or 1 for response), and his or her treatment assignment. To visualize this, we plotted each participant as a data point with the index score on the x-axis and observed response on the y-axis as shown in Figure 3.2. Since the values on the y-axis can only be 1 or 0, the points have been jittered to show all of the data. Each point is colored according to the participant's treatment assignment, with

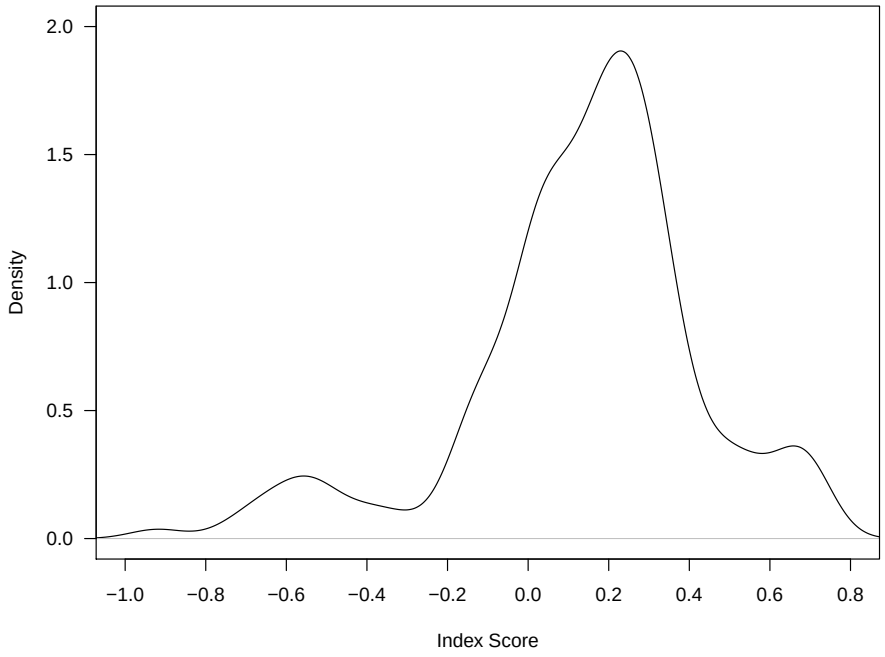

Figure 3.1: Density of index scores

citalopram in red and placebo in blue. A loess smoother for each treatment group is also shown to help visualize the distribution of responses across index scores. As expected, the response probability increases with increasing index score in the citalopram group. We might have expected a flat line near zero for the placebo group, as the placebo should not have increased or decreased the probability of response. However, there is an apparent negative association between index score and response probability in the placebo group. There was a low level of response in the placebo group (21 out of 81 participants) which is likely attributable to increased level of patient care and counseling which was provided to all caregivers. As shown in Table 3.2, placebo response is associated with several working model covariates including baseline severity of agitation. The index score is designed to predict differences between the predicted responses (citalopram minus placebo), and so we observe higher placebo response rates at low index scores and higher citalopram response rates at high index scores. At the point where the curves cross, there is roughly no observed difference

in response rates between patients randomized to citalopram versus placebo.

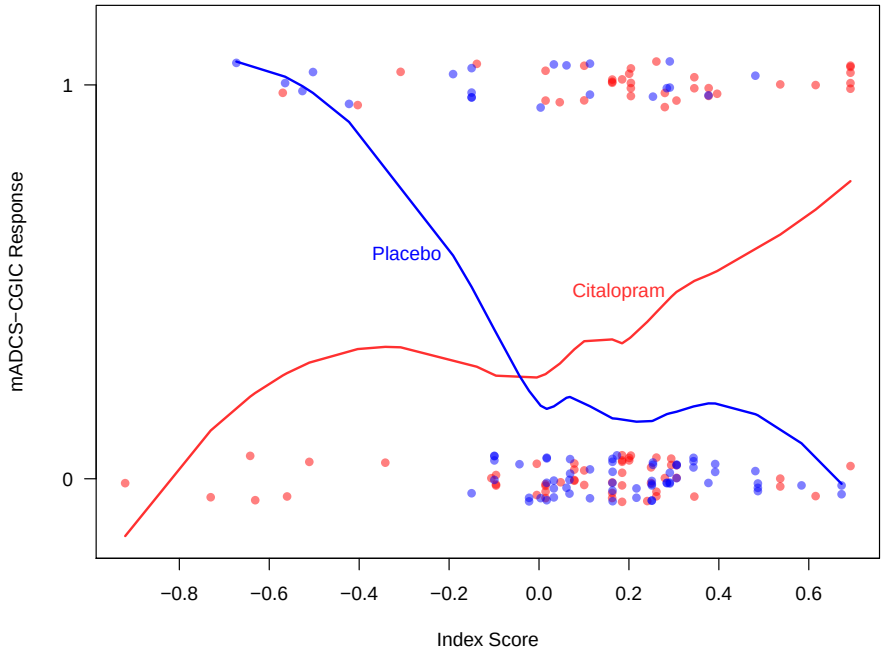

Figure 3.2: Plot of mADCS-CGIC response by index score

3.2.2 Non-parametric smoothing technique

We next use a smoothing technique to obtain a non-parametric estimate of the response probability for each treatment group at each index score. Unlike the smoother used in Cai et al. [3], we used a cumulative smoothing technique because it makes the analysis easily interpretable. The non-parametric estimate of the response probability for a particular index score is the proportion of persons with a response out of all persons with an index score greater than or equal to the selected score. The left-most estimate (corresponding to index score -0.920 or greater) is the average probability of response for all participants. The right-most estimate is the proportion of persons with a response at the largest index score, 0.693, alone.

A plot of the non-parametric estimates of the response probability by treatment group is shown in Figure 3.3. The last step to obtaining the non-parametric estimates

of the treatment effect is to subtract the estimates for the placebo group (shown in blue) from the estimates for the citalopram group (shown in red) at each index score. The resulting values are the non-parametric estimates of the treatment effect shown in black in Figure 3.3. As an example of how to read and interpret the plots in Figure 3.3, participants with index scores of 0.2 or greater comprise a subgroup for which the estimated response probability under placebo is 0.167 and the estimated response probability under citalopram is 0.525; the estimated treatment effect for this subgroup is the difference in those probabilities, 0.358.

3.3 Confidence intervals

3.3.1 Point-wise confidence intervals

We created 1,000 bootstrap samples by sampling with replacement from the original 167 participants with available week 9 mADCS-CGIC data. Each bootstrap dataset consists of 167 entries each with treatment assignment, the observed mADCS-CGIC response, and index score (calculated from the original dataset). For each bootstrapped dataset, we re-ran the cumulative smoother to obtain a new estimate of the treatment effect at each index score. The empirical distributions of the 1,000 bootstrap estimates at each index score are the basis for point-wise 95% confidence intervals as shown in Figure 3.5. The 0.025 quantile of each distribution is the lower bound and the 0.975 quantile is the upper bound of the 95% confidence interval for the treatment effect estimate at each index score.

3.3.2 Correcting for multiplicity

Because we are estimating these point-wise 95% confidence intervals for multiple index scores (there are 68 unique index scores among the CitAD participants), we made a correction for multiple comparisons to preserve the overall type I error rate. For each

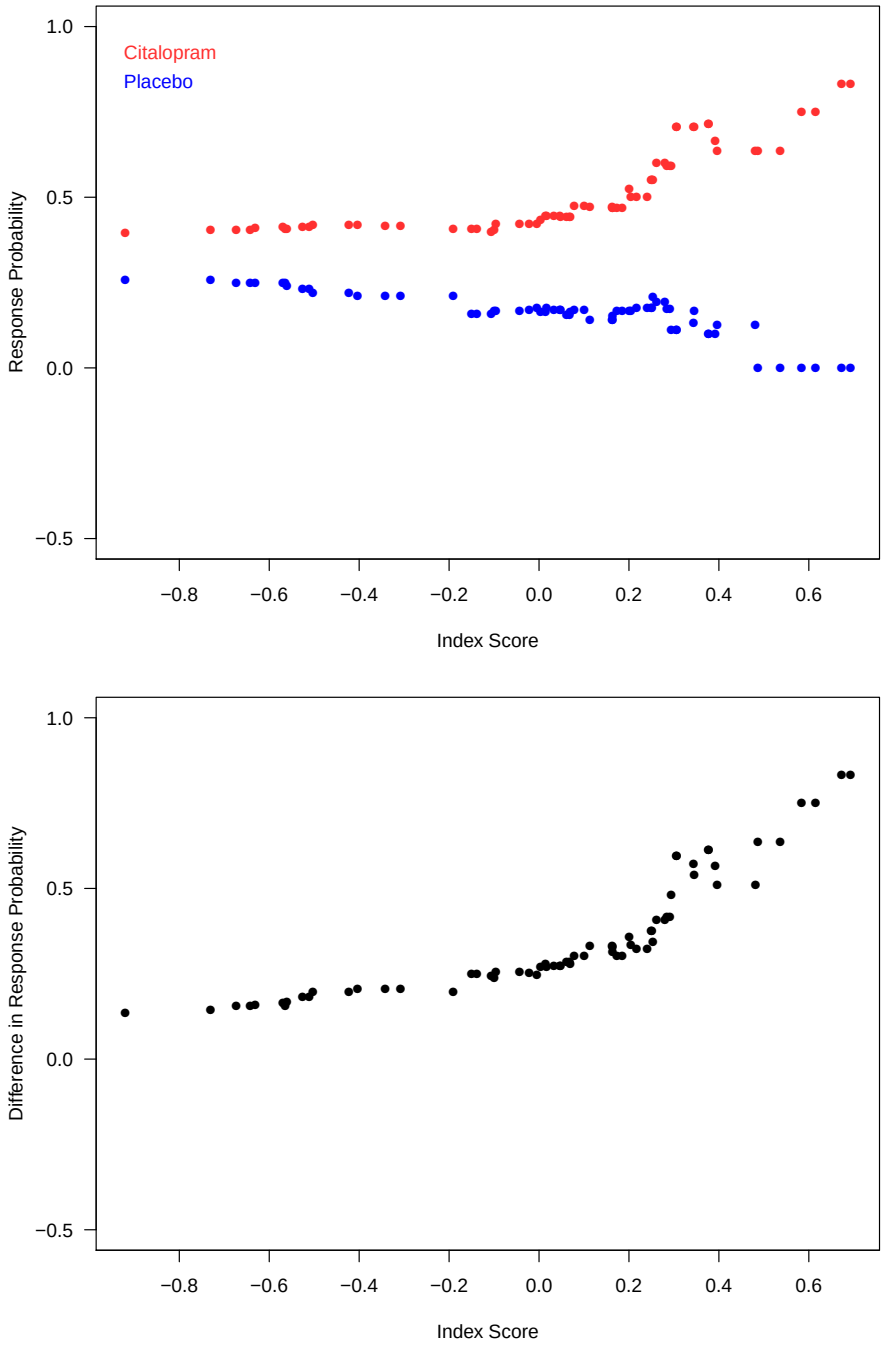

Figure 3.3: Plots of response probability and treatment effect by index score

point-wise confidence interval, there is a 5% probability that the corresponding true treatment effect value lies outside the bounds simply by chance. For the set of all 68 true subgroup treatment effects, the chance that any one of them lies outside the

bounds is then necessarily larger than 5%. We need to apply a correction such that the confidence bands cover all 68 true subgroup treatment effects simultaneously with 95% probability. The corrected confidence bounds will be wider than the point-wise confidence intervals.

In Cai et al. [3], multiplicity is addressed by the calculation of a simultaneous confidence interval in addition to the point-wise interval. The simultaneous confidence interval is given by Equation 3.6 where $\hat{\delta}_i$ is the bootstrap treatment effect estimate at the i^{th} index score, $\hat{\sigma}_i$ is the corresponding bootstrap standard error estimate, and γ is a correction factor.

$$\hat{\delta}_i \pm \gamma \hat{\sigma}_i \tag{3.6}$$

The correction factor, γ, is the 95^{th} percentile of the distribution of maximum standard deviations of the treatment effect estimates from their means in each bootstrap sample. Each bootstrap iteration produces 68 subgroup treatment effect estimates. Each of these estimates can be given a standardized Z-score to represent its distance from the mean (across all bootstrap samples) as shown in Equation 3.7, where δ_{ij} is the treatment effect estimate for index score i and bootstrap sample j, $\hat{\sigma}_i$ is the bootstrap standard error estimate, and $\bar{\delta_{i\cdot}}$ is the mean of the bootstrap treatment effects at index score i. The maximum Z-score for each bootstrap sample is given by Equation 3.8. The correction factor γ is then the 95^{th} percentile of the distribution of 1,000 $Z\text{max}$ values. The distribution of 1,000 maximum Z-scores and correction factor are shown in Figure 3.4. The correction factor after considering multiple (68) looks is larger than 1.96 which means the simultaneous confidence band will be wider than the point-wise confidence intervals. The point-wise and simultaneous confidence bands are shown in Figure 3.5.

$$Z_{ij} = \frac{|\delta_{ij} - \bar{\delta_{i.}}|}{\hat{\sigma}_i} \tag{3.7}$$

$$Z\max_j = \max\{Z_{1j}, Z_{2j}, ..., Z_{68j}\} \tag{3.8}$$

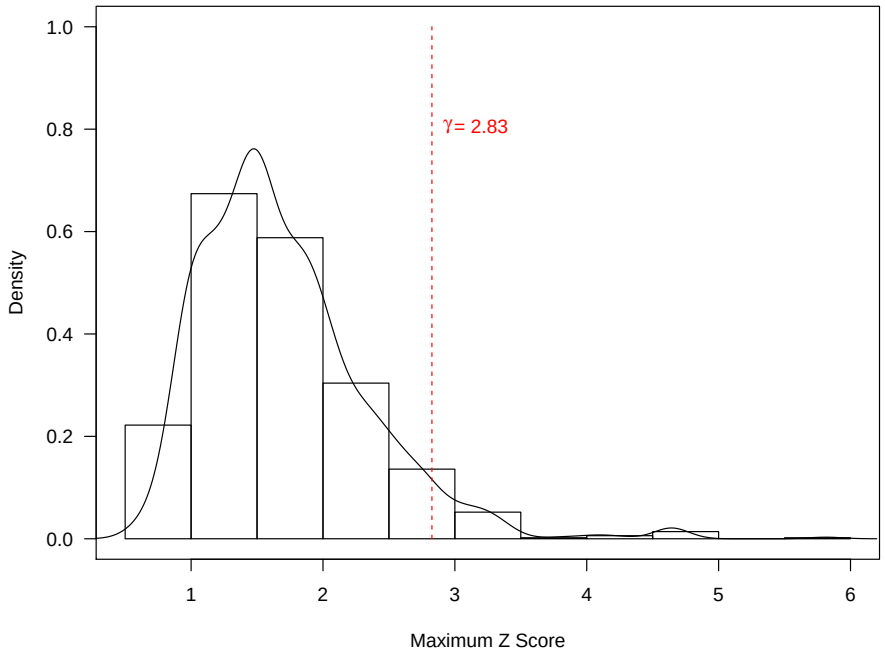

Figure 3.4: Density of maximum Z-scores from 1,000 bootstrapped datasets

Instead of reporting both point-wise and simultaneous confidence intervals, we have instead chosen to report a single 95% confidence interval which corrects for multiple looks moving left to right. We think that the natural interpretation of the data is to observe the average treatment effect (the estimate at the farthest left) and then move right until a subgroup is found to have an estimated treatment effect which significantly exceeds the average treatment effect. The index score for the first estimated treatment effect moving left to right which meets this criteria defines the largest subgroup for which we can say that the treatment effect exceeds the average. Therefore, we have calculated a separate correction factor, γ_i, for each index score

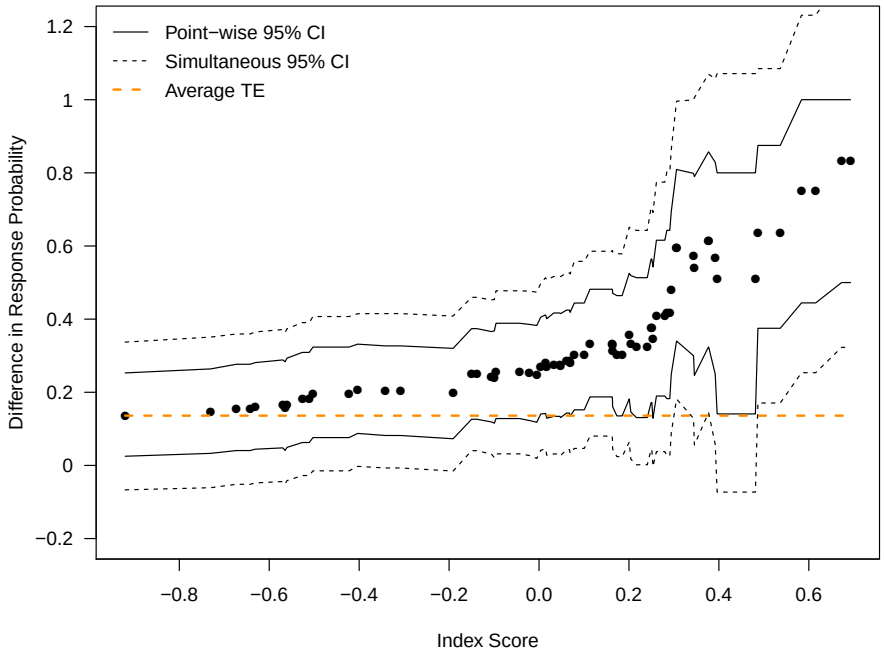

Figure 3.5: Plot of treatment effect with point-wise and simultaneous 95% confidence intervals

which takes into account all previous looks (to the left of that score). The confidence interval to the far left is therefore the same as the point-wise interval and the confidence interval to the far right is the same as the simultaneous interval. The resulting plot is shown in Figure 3.6. The x-axis labels have been changed from the index scores to the percent of CitAD participants with that index score or greater for ease of interpretation. This also is the percentage of participants in each subgroup. The orange line represents the average treatment effect, the probability difference estimated from all participants.

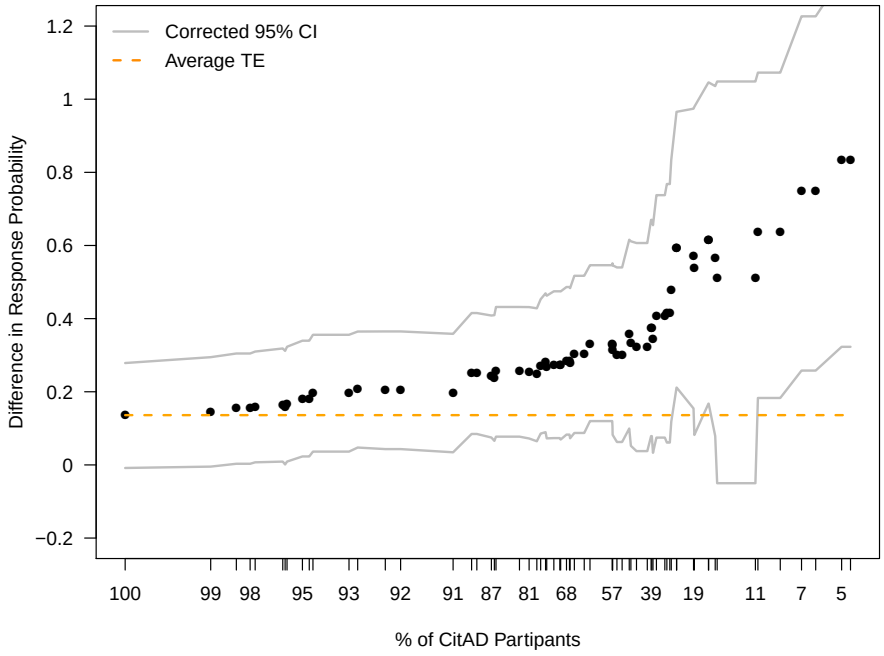

Figure 3.6: Plot of treatment effect and 95% confidence intervals after correcting for multiplicity left to right

Chapter 4

Hypothesis Testing

4.1 Test for heterogeneity by index score deciles

We conducted a hypothesis test for overall treatment effect heterogeneity using the two-stage estimation method without the cumulative smoother. Instead of using the cumulative smoothing technique, we split the data into ten groups based on index score deciles. We then calculated the non-parametric estimate of the treatment effect within each decile; the non-parametric treatment effect estimate is the proportion of observed responders in the citalopram group minus the proportion of responders in the placebo group. We did the same procedure for 1,000 bootstrapped samples to get a distribution of the treatment effect for each group which is shown in Figure 4.1.

The null hypothesis is that the treatment effect (difference in response probability) is the same in all of the ten subgroups. The estimate of the common treatment effect is the inverse variance weighted average of the ten individual treatment effect estimates as shown in Equation 4.1 where θ_i is the mean treatment effect estimate for decile i,

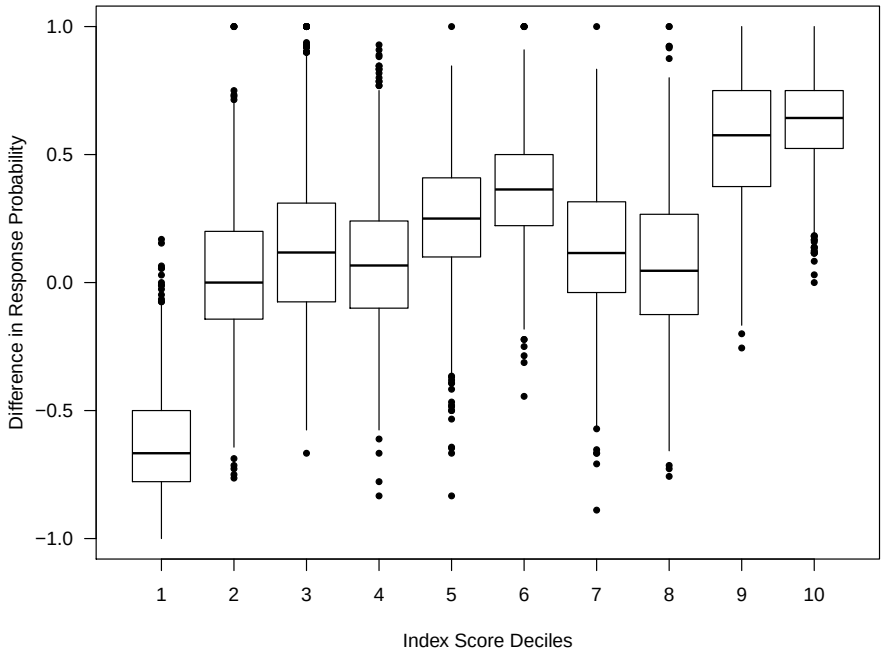

Figure 4.1: Boxplot of bootstrap estimates of treatment effect at index score deciles

and σ_i^2 is the variance of the treatment effect estimates for decile i.

$$\hat{\theta}_0 = \frac{\sum_{i=1}^{10} \left(\frac{\theta_i}{\sigma_i^2} \right)}{\sum_{i=1}^{10} \left(\frac{1}{\sigma_i^2} \right)} = 0.173 \tag{4.1}$$

We then used a likelihood ratio test to compare the null model with a common treatment effect to the full model with a separate treatment effect estimated for each decile. The likelihood ratio is shown in Equation 4.2. In this calculation, we assume the decile estimates are independent; after examining the empirical correlation matrix, this seems to be a reasonable assumption. The test statistic (TS) for the likelihood ratio test is shown in Equation 4.3. This test statistic is compared to a χ^2 distribution with 9 degrees of freedom (the number of groups in the full model minus the number in the reduced model). The resulting p-value (Equation 4.4) is very small suggesting that at least two subgroups, as defined by the index score deciles, have different

treatment effects.

$$LR(\boldsymbol{\theta}) = \frac{\prod_{i=1}^{10} \left(2\pi\hat{\sigma}_i^2\right)^{1/2} \exp\left\{-\frac{(\theta_i - \hat{\theta}_0)^2}{2\hat{\sigma}_i^2}\right\}}{\prod_{i=1}^{10} \left(2\pi\hat{\sigma}_i^2\right)^{1/2} \exp\left\{-\frac{(\theta_i - \hat{\theta}_i)^2}{2\hat{\sigma}_i^2}\right\}}$$

$$= \prod_{i=1}^{10} \exp\left\{-\frac{(\theta_i - \hat{\theta}_0)^2}{2\hat{\sigma}_i^2}\right\} \tag{4.2}$$

$$TS = -2\log\left[LR(\boldsymbol{\theta})\right] = \sum_{i=1}^{10} \frac{(\theta_i - \hat{\theta}_0)^2}{\hat{\sigma}_i^2} = 26.0 \tag{4.3}$$

$$P(TS \geq \chi_9^2) = 0.00206 \tag{4.4}$$

4.2 Comparison of the ten decile mean model with a three mean model.

Looking at Figure 4.1, there appear to be three distinct groups: placebo responders defined by decile 1, non-responders defined by deciles 2-8, and citalopram responders defined by deciles 9-10. We hypothesize that this simpler, three group model is adequate to describe differences in treatment effect across index scores instead of the ten decile model. We used a likelihood ratio test to compare the reduced model with three means to the full model with ten means. The test statistic is calculated as previously shown, but the null common treatment effect estimate is replaced with the inverse variance weighted average effect within each of the three groups. The test statistic for this comparison is 1.46 which is compared to a χ^2 distribution with 7 degrees of freedom (the number of groups in the full model minus the number in the reduced model). The resulting p-value is 0.984, suggesting that the full ten decile

model is not a significant improvement over the proposed three group model. The boxplot for the three mean model is shown in Figure 4.2. The estimated treatment effects for the three groups are -0.629 for decile 1, 0.166 for deciles 2-8, and 0.605 for deciles 9-10.

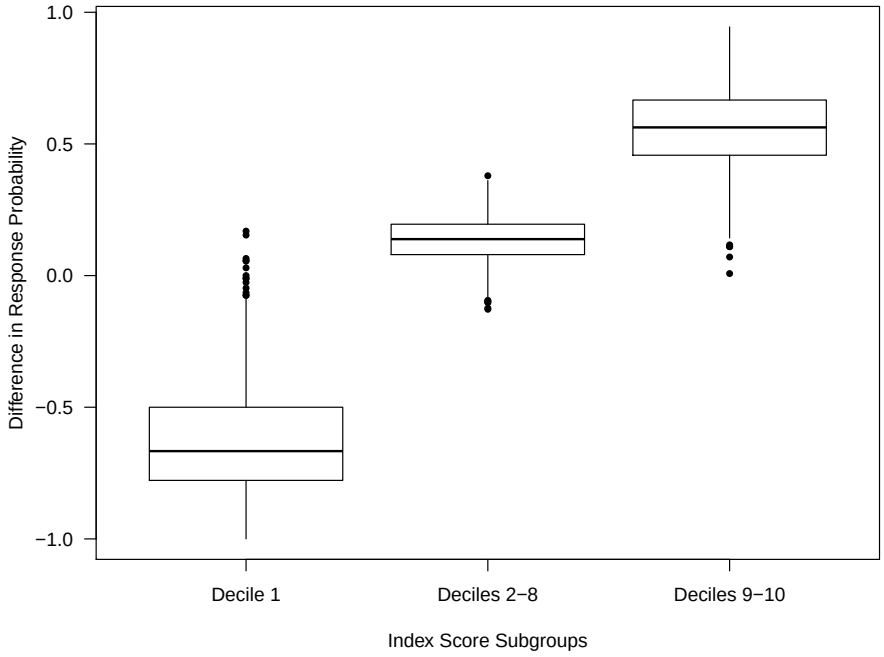

Figure 4.2: Boxplot of bootstrap estimates of treatment effect at three index score subgroups

4.3 Test for consistent heterogeneity of NBRS-A

We would expect true differences across subgroups to be consistent among related outcomes. The two primary outcomes of the CitAD study, mADCS-CGIC and NBRS-A, both measure severity of agitation. To assess consistency of these subgroup effects between the two measures, we did a hypothesis test for heterogeneity in the NBRS-A outcome at the mADCS-CGIC index score deciles. We used the same method as described in Section 4.1, except we calculated the non-parametric estimate of the NBRS-A treatment effect within each decile. The non-parametric treatment effect

estimate is the proportion of observed NBRS-A responders in the citalopram group minus the proportion of responders in the placebo group where NBRS-A response was defined as at least 50% reduction from baseline NBRS-A at week 9. This procedure was repeated for 1,000 bootstrapped samples to obtain a distribution of the NBRS-A treatment effect for each decile as shown in Figure 4.3. The likelihood ratio test for the null hypothesis of a common treatment effect suggests no significant differences in NBRS-A treatment effect across the mADCS-CGIC index score decile subgroups ($p = 0.864$).

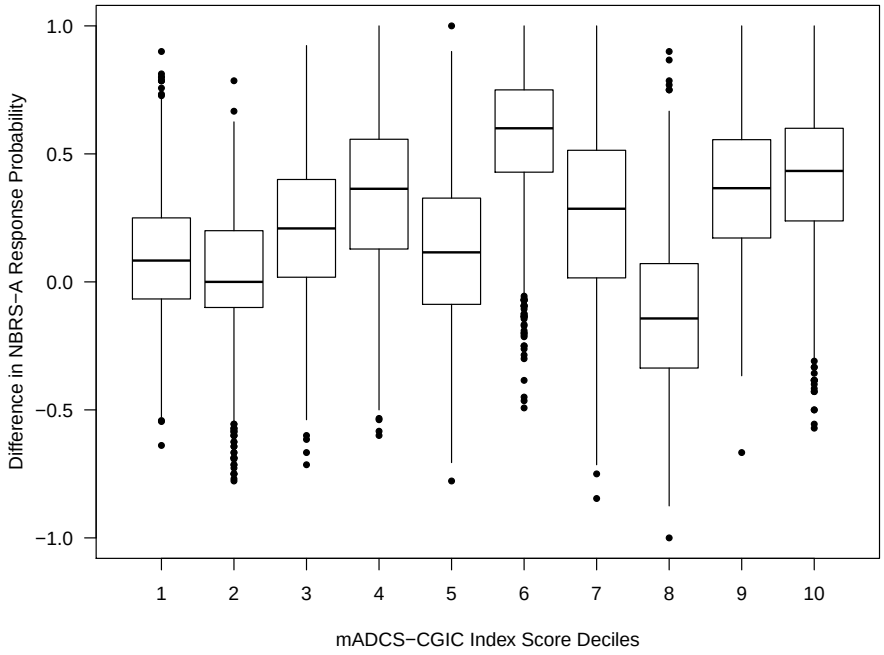

Figure 4.3: Boxplot of bootstrap estimates of NBRS-A treatment effect at mADCS-CGIC index score deciles

Chapter 5

Results & Discussion

In this chapter, we answer the original three research questions and interpret our results. We will also relate the index scores back to the original baseline covariates to propose which factors may be predictive of citalopram response.

5.1 Aim 1: What is the average treatment effect?

There is evidence that citalopram increases mADCS-CGIC response (marked or moderate improvement of agitation symptoms) when compared to placebo. The estimated average effect is a 0.136 difference in response probability (citalopram - placebo) with a 95% confidence interval of [0.0134, 0.297]. The average treatment affect can be seen in Figure 3.6 as the treatment effect estimate at the 100% percentile mark on the x axis. This average treatment effect is an estimate of the difference in the potential outcome if all of the study participants had been assigned to citalopram and the potential outcome if they had all been assigned to placebo.

This finding is consistent with previous estimates of the average mADCS-CGIC treatment effect. In the primary outcomes analysis, the average mADCS-CGIC treatment effect estimate was an odds ratio of 2.13 (odds ratio of being at or better than a given mADCS-CGIC category for citalopram versus placebo) with a 95% confidence

interval of [1.23, 3.69] as estimated using ordinal logistic regression. The estimated treatment effect from this analysis is smaller in magnitude than that reported in the primary analysis; the difference between the two estimates is likely attributable to our treatment of the original seven category mADCS-CGIC score as a binary response variable. In the simple bivariate logistic regression analyses (Figures 2.1 & 2.2), the average treatment effect estimate was an odds ratio of 1.87 (odds ratio of marked or moderate improvement for citalopram versus placebo) with a 95% confidence interval of [0.967, 3.61]. All estimates suggest that citalopram significantly improves mADCS-CGIC scores compared to placebo.

5.2 Aim 2: Is there a subgroup with a larger than average treatment effect?

From Figure 3.6 it is observed that there are several subgroups of patients for which the entire confidence interval is above the average treatment effect of 100% of patients. Moving from the left (the average treatment effect) to the right, the largest subgroup for which the confidence interval almost exceeds the average treatment effect is the subgroup of patients with the largest 60% of the index scores. The largest subgroup for which confidence interval definitely exceeds the average treatment effect is the subgroup with the top 20% of index scores.

This can be interpreted clinically by viewing the distribution of covariates for these patients and comparing to the distribution of all patients. These comparisons are made in Table 5.1 and Table 5.2. The distribution of covariates among all participants is shown in Table 5.1. We then show the distribution of covariates among the top 90% to 10% of index scorers in increments of 10%. The subgroup of the top 60% has a larger distribution of participants living out of long term care, having mild to no cognitive impairment (MMSE), in the middle tertile of baseline agitation

symptoms (NBRS-A), in the middle age range, and not using lorazepam at baseline. These trends become more prominent as you observe the distribution of covariates in subgroups with larger average index scores.

	All Participants		Top 90%		Top 80%		Top 70%		Top 60%	
Residence										
Home or relative	155	(93%)	150	(99%)	134	(100%)	120	(100%)	100	(100%)
Long term care	12	(7%)	1	(1%)	0	(0%)	0	(0%)	0	(0%)
MMSE										
Mild to no impairment 21-30	49	(29%)	46	(30%)	44	(33%)	44	(37%)	38	(38%)
Moderate, 11-20	75	(45%)	69	(46%)	56	(42%)	42	(35%)	36	(36%)
Severe, 0-10	43	(26%)	36	(24%)	34	(25%)	34	(28%)	26	(26%)
NBRS-A										
Smallest tertile, 1-5	48	(29%)	43	(28%)	33	(25%)	19	(16%)	10	(10%)
Middle tertile, 6-8	55	(33%)	54	(36%)	53	(40%)	53	(44%)	51	(51%)
Largest tertile, 9-14	64	(38%)	54	(36%)	48	(36%)	48	(40%)	39	(39%)
Age										
Smallest tertile, 47-75	54	(32%)	49	(32%)	34	(25%)	34	(28%)	23	(23%)
Middle tertile, 76-82	53	(32%)	50	(33%)	50	(37%)	46	(38%)	40	(40%)
Largest tertile, 83-92	60	(36%)	52	(34%)	50	(37%)	40	(33%)	37	(37%)
Lorazepam										
No lorazepam use	156	(93%)	147	(97%)	131	(98%)	117	(98%)	99	(99%)
Lorazepam use	11	(7%)	4	(3%)	3	(2%)	3	(2%)	1	(1%)

Table 5.1: Baseline characteristics of subgroups with top 100% to 60% of index scores

	Top 50%		Top 40%		Top 30%		Top 20%		Top 10%	
Residence										
Home or relative	84	(100%)	68	(100%)	51	(100%)	35	(100%)	17	(100%)
Long term care	0	(0%)	0	(0%)	0	(0%)	0	(0%)	0	(0%)
MMSE										
Mild to no impairment, 21-30	38	(45%)	30	(44%)	23	(45%)	23	(66%)	17	(100%)
Moderate, 11-20	20	(24%)	13	(19%)	6	(12%)	6	(17%)	0	(0%)
Severe, 0-10	26	(31%)	25	(37%)	22	(43%)	6	(17%)	0	(0%)
NBRS-A										
Smallest tertile, 1-5	10	(12%)	7	(10%)	0	(0%)	0	(0%)	0	(0%)
Middle tertile, 6-8	49	(58%)	41	(60%)	31	(61%)	26	(74%)	17	(100%)
Largest tertile, 9-14	25	(30%)	20	(29%)	20	(39%)	9	(26%)	0	(0%)
Age										
Smallest tertile, 47-75	21	(25%)	9	(13%)	6	(12%)	6	(17%)	6	(35%)
Middle tertile: 76-82	36	(43%)	36	(53%)	29	(57%)	21	(60%)	8	(47%)
Largest tertile: 83-92	27	(32%)	23	(34%)	16	(31%)	8	(23%)	3	(18%)
Lorazepam										
No lorazepam use	83	(99%)	68	(100%)	51	(100%)	35	(100%)	17	(100%)
Lorazepam use	1	(1%)	0	(0%)	0	(0%)	0	(0%)	0	(0%)

Table 5.2: Baseline characteristics of subgroups with top 50% to 10% of index scores

5.3 Aim 3: Is the treatment effect truly heterogeneous?

There is evidence that there are groups with different true average effects. The p-value of the hypothesis test for heterogeneity is 0.002 (Equations 4.2-4.4) which suggests that the treatment effect is not homogeneous across all participants. This can also be interpreted clinically by examining the distribution of covariates in each index score subgroup. There are three distinct index score subgroups with different average treatment effects as shown in Figure 4.2. These distribution of covariates in these three subgroups are shown in Table 5.3.

The distribution of covariates in the subgroup with the largest index scores (deciles 9-10) are consistent with the patterns seen in Aim 2. We can also examine the distribution of covariates in the subgroup with the lowest potential for benefit (decile 1); the covariate categories associated with reduced treatment effect are living in long term care facilities, moderate to severe cognitive impairment, severe baseline symptoms of agitation, being in the youngest age tertile (ages 47-75), and baseline use of lorazepam.

	All Participants		Decile 1		Deciles 2-8		Deciles 9-10	
Residence								
Home or relative	155	(93%)	10	(48%)	113	(99%)	32	(100%)
Long term care	12	(7%)	11	(52%)	1	(1%)	0	(0%)
MMSE								
Mild to no impairment, 21-30	49	(29%)	3	(14%)	23	(20%)	23	(72%)
Moderate, 11-20	75	(45%)	11	(52%)	58	(51%)	6	(19%)
Severe, 0-10	43	(26%)	7	(33%)	33	(29%)	3	(9%)
NBRS-A								
Smallest tertile, 1-5	48	(29%)	5	(24%)	43	(38%)	0	(0%)
Middle tertile, 6-8	55	(33%)	1	(5%)	28	(25%)	26	(81%)
Largest tertile, 9-14	64	(38%)	15	(71%)	43	(38%)	6	(19%)
Age								
Smallest tertile, 47-75	54	(32%)	10	(48%)	38	(33%)	6	(19%)
Middle tertile, 76-82	53	(32%)	3	(14%)	29	(25%)	21	(66%)
Largest tertile, 83-92	60	(36%)	8	(38%)	47	(41%)	5	(16%)
Lorazepam								
No lorazepam use	156	(93%)	14	(67%)	110	(96%)	32	(100%)
Lorazepam use	11	(7%)	7	(33%)	4	(4%)	0	(0%)

Table 5.3: Baseline charactestics of three index score subgroups with different average treatment effects

5.4 Discussion

5.4.1 Plausibility

One strength of our analysis is the focus on baseline covariates which were pre-specified in the protocol, although we also considered some plausible post-hoc covariates. All of the subgroup analyses that we have completed for the mADCS-CGIC outcome have been presented in this report; we did not continue on a "fishing expedition" to explore additional baseline covariates which may not have any biological or clinical relevance. The cut-points to categorize/collapse the continuous and ordinal baseline covariates were not pre-specified, but were chosen in a rather agnostic manner. We based our decisions on clinically relevant values from the literature and distribution tertiles, instead of searching for cutpoints which would yield the most significant interactions. All of these aspects of our method enhance the credibility of our results.

Although the subgroup covariates were selected before looking at any of the data, the direction of the associations were not explicitly pre-specified. Our collaborators from the study agree that the observed interactions seem plausible and have offered various biological explanations. However, it is difficult to ascertain, after the fact, whether the direction of the observed interactions were consistent with their original expectations based on clinical practice or whether they had been influenced by seeing the results. As these results have not yet been disseminated to all of the investigators, a study has been proposed to conduct a survey to see how the naive clinician would rank the significance and direction of the association between the treatment and each of the covariates.

Of major concern for plausibility is the observed lack of consistency between the mADCS-CGIC and NBRS-A treatment effects across the same subgroups. As both primary outcomes measure severity of agitation, we would expect to see the same

interactions using both measures if these subgroup effects are real. The observed lack of treatment effect heterogeneity in the NBRS-A response across subgroups (Figure 4.3) reduces the credibility of our results somewhat. We are unsure whether the lack of consistency is an indication that the proposed interactions are spurious effects or if the lack of consistency is the product of disagreement between the two outcome measures. As shown in Table 5.4, the cross-tabulation of week 9 mADCS-CGIC response (marked or moderate improvement) versus week 9 NBRS-A response (>= 50% reduction from baseline), the two measures of agitation do not completely agree with one another. This lack of agreement could be a result of how the original ordinal scales were dichotomized. It is possible that we would see better agreement across subgroups if the NBRS-A variable were treated as continuous or categorized in a different way. Given the sample size, we may also not have the power to detect true heterogeneity in the NBRS-A outcome.

mADCS-CGIC Response	NBRS-A Response 0	1
0	78	34
1	14	41

Table 5.4: Agreement between mADCS-CGIC and NBRS-A responses

5.4.2 Contribution to existing methodology

Our contribution to the two-stage estimation procedure is the use of the cumulative smoothing technique to determine the largest group of patients for which the subgroup treatment effect is significantly larger than the average. Additionally, we describe this subgroup as the percent of participants with the highest index scores, providing a means for evaluation of the index scoring system. The calculated index scores depend on the selection of the working models. Currently there is no way to compare the performance of different specifications of the working models, such as inclusion

of different baseline covariates. An ideal index scoring system will efficiently group or rank patients such that there is the largest differentiation between high treatment effect and low treatment effect subgroups. One way of evaluating the performance of the method would be to maximize the percent of participants in the largest subgroup with an above average treatment effect.

5.4.3 Assumptions & limitations

The two-stage estimation procedure is subject to the assumption that the working models are specified such that they are useful in ranking participants by their true treatment effects. The working models are used to generate the index score which is the predicted treatment effect for a subgroup of patients with the same combination of baseline covariates. The non-parametric estimation methods depend on the fact that these subgroups can be efficiently ranked according to their true treatment effects using the index scores. So the working models are not used for inference, but as a tool to facilitate non-parametric estimation. The assumption is that the working models are close enough to the true relationship to efficiently rank the subgroups. This is less limiting than the typical assumption of correct specification required for most parametric methods. We are not certain how sensitive this two-stage method is to gross mis-specification of the working models. Other limitations include complete case analysis which is subject to the assumption that the outcome was missing completely at random, meaning that the missingness did not depend on previous responses or the missing value. This is likely not a valid assumption, however our results should be fairly robust to this assumption as missingness rates were low (10% missing primary outcomes at week 9 and full baseline information on all participants). To relax this assumption, we could impute missing outcomes using observed data from the participant's index score decile.

There are several additional aspects of this analysis which may limit its usefulness

in guiding actual clinical practice. Our collaborators have noted that it would be useful to have subgroups based on concrete measures which are less subject to fluctuation. While age and residency are very concrete measures, the MMSE and NBRS-A scores may be subject to measurement error and random day-to-day fluctuations. This was somewhat out of our control, as we tried to focus on subgroups which were pre-specified in the protocol. Another limitation is that this analysis also does not account for possible risks associated with citalopram. Clinicians must consider both the potential benefits and risks in making prescribing decisions for each patient. This would require the additional modeling of a risk profile for each patient, as the probability of adverse events is likely heterogeneous as well. Our current analysis should be regarded as one piece of a multitude of factors clinicians must take into account.

Chapter 6

Conclusions

In this thesis, we have illustrated how the novel two-stage estimation approach introduced by Cai et al. [3] can be applied to a real clinical trial using the CitAD dataset. Using this approach, we have identified several likely predictors of citalopram response. CitAD participants with the largest predicted treatment effects were more likely to be living outside long-term care facilities, within the middle age range of CitAD participants (ages 76-82), with minimal cognitive impairment (MMSE 21-30), within the middle baseline agitation range (NBRS-A 6-8), and not taking lorazepam. These trends were also seen using traditional bivariate subgroup analysis methods, however we did not have the power to detect any significant individual interactions as the sample size was quite limited. The two-stage estimation procedure has allowed us to consider combinations of multiple baseline factors simultaneously and calculate non-parametric estimates of subgroup treatment effects. This approach has provided more persuasive evidence for true treatment effect heterogeneity among CitAD participants.

Concerns such as the lack of consistency between related outcomes and questionable biological plausibility cast doubt on our findings. With this in mind, we advise the reader to intperpret these results as exploratory or hypothesis generating rather

than confirmatory. Additional experimental or observational data will be required to confirm the proposed interactions. Future research plans include a survey study designed to more objectively assess clinical opinion on the biological plausibility of the observed interactions. Future development of the method may include changing the framework for calculation of the index score to improve prediction.

References

[1] Alzheimer's Association. 2010 Alzheimer's disease facts and figures. *Alzheimer's & Dementia*, 6(2):158–194, March 2010.

[2] Clive G. Ballard, Serge Gauthier, Jeffrey L. Cummings, Henry Brodaty, George T. Grossberg, Philippe Robert, and Constantine G. Lyketsos. Management of agitation and aggression associated with Alzheimer disease. *Nature Reviews Neurology*, 5(5):245–255, May 2009.

[3] Tianxi Cai, Lu Tian, Peggy H. Wong, and L. J. Wei. Analysis of randomized comparative clinical trial data for personalized treatment selections. *Biostatistics*, 12(2):270–282, April 2011.

[4] Lea T. Drye, Zahinoor Ismail, Anton P. Porsteinsson, Paul B. Rosenberg, Daniel Weintraub, Christopher Marano, Gregory Pelton, Constantine Frangakis, Peter V. Rabins, Cynthia A. Munro, Curtis L. Meinert, D. P. Devanand, Jerome Yesavage, Jacobo E. Mintzer, Lon S. Schneider, Bruce G. Pollock, and Constantine G. Lyketsos. Citalopram for agitation in Alzheimer's disease: design and methods. *Alzheimer's & Dementia*, 8(2):121–130, January 2012.

[5] Margaret Gatz, Chandra Reynolds, Laura Fratiglioni, Boo Johansson, James Mortimer, Stig Berg, Amy Fiske, and Nancy L. Pedersen. Role of genes and environments for explaining Alzheimer disease. *Archives of General Psychiatry*, 63(2):168–174, February 2006.

[6] Jacob H. G. Grand, Sienna Caspar, and Stuart W. S. Macdonald. Clinical features and multidisciplinary approaches to dementia care. *Journal of Multidisciplinary Healthcare*, 4:125–147, January 2011.

[7] George T. Grossberg and Abhilash K. Desai. Management of Alzheimer's disease. *Journal of Gerontology: Medical Sciences*, 58A(4):331–353, October 2003.

[8] David M. Kent, Peter M. Rothwell, John P. A. Ioannidis, Doug G. Altman, and Rodney A. Hayward. Assessing and reporting heterogeneity in treatment effects in clinical trials: a proposal. *Trials*, 11(85):1–11, January 2010.

[9] Zaven S. Khachaturian. Diagnosis of Alzheimer's disease. *Journal of the American Medical Association Neurology*, 42(11):1097–1105, 1985.

[10] D. Mungas. In-office mental status testing: a practical guide. *Geriatrics*, 46(7):54–58, 63, 66, July 1991.

[11] Stuart J. Pocock, Susan E. Assmann, Laura E. Enos, and Linda E. Kasten. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. *Statistics in Medicine*, 21(19):2917–2930, October 2002.

[12] Anton P. Porsteinsson, Lea T. Drye, Bruce G. Pollock, D. P. Devanand, Constantine Frangakis, Zahinoor Ismail, Christopher Marano, Curtis L. Meinert, Jacobo E. Mintzer, Cynthia A. Munro, Gregory Pelton, Peter V. Rabins, Paul B. Rosenberg, Lon S. Schneider, David M. Shade, Daniel Weintraub, Jerome Yesavage, and Constantine G. Lyketsos. Effect of citalopram on agitation in Alzheimer disease: the CitAD randomized clinical trial. *Journal of the American Medical Association*, 311(7):682–691, February 2014.

[13] Christiane Reitz, Carol Brayne, and Richard Mayeux. Epidemiology of Alzheimer disease. *Nature Reviews Neurology*, 7(3):137–152, March 2011.

[14] Cynthia Steele, Barry Rovner, Gary A. Chase, and Marshal Folstein. Psychiatric symptoms and nursing home placement of patients with Alzheimer's disease. *American Journal of Psychiatry*, 8(August):1049–1051, 1990.

[15] Xin Sun, Matthias Briel, S. D. Walter, and G. H. Guyatt. Is a subgroup effect believable? Updating criteria to evaluate the credibility of subgroup analyses. *British Medical Journal*, 340:850–854, 2010.

[16] Rui Wang, Stephen W. Lagakos, James H. Ware, David J. Hunter, and Jeffrey M. Drazen. Statistics in medicine: reporting of subgroup analyses in clinical trials. *The New England Journal of Medicine*, 357(21):2189–2194, November 2007.

Curriculum Vitae

Lisa E. Rein

Education

2012 - 2014	ScM Biostatistics, Johns Hopkins University
2003 - 2007	BS Engineering Physics, University of Illinois
	Minors: Mathematics, Bioengineering

Experience

2013 - 2014	Student Consultant, Johns Hopkins Biostatistics Center
2008 - 2012	Research Technologist, Walter Reed Army Institute of Research
2007 - 2008	Research Scientist, ToxServices LLC
2005 - 2007	Serials Clerk, University of Illinois Library
2006	Research Assistant, University of Illinois NASA UROP internship
2004	Research Assistant, L'Institut National Polytechnique de Lorraine

Publications

Chuang, I., Sedegah, M., ..., **Rein, L.**, et al. (2013) DNA prime/Adenovirus boost malaria vaccine encoding P. falciparum CSP and AMA1 induces sterile protection associated with cell-mediated immunity. PloS ONE, 8(2): e55571.

Schwenk R, Lumsden JM, **Rein LE**, et al. (2011) Immunization with the RTS,S/AS malaria vaccine induces IFN-γ+ CD4 T cells that recognize only discrete regions of the circumsporozoite protein and these specificities are maintained following booster immunizations and challenge. Vaccine. 29(48): 8847-8854.

Lumsden JM, Schwenk RJ, **Rein LE**, et al. (2011) Protective immunity induced with the RTS,S/AS vaccine is associated with IL-2 and TNF-α producing effector and central memory CD4+ T cells. PLoS ONE. 6(7): e20775.