The Dissertation Committee for Nishant Verma
certifies that this is the approved version of the following dissertation:

Biomarker for Tracking Progression of Alzheimer's Disease in Clinical Trials

Committee:

Mia K. Markey, Supervisor

Matthew C. Cowperthwaite

Alan C. Bovik

Kristen Grauman

Joydeep Ghosh

Biomarker for Tracking Progression of Alzheimer's Disease in Clinical Trials

by

Nishant Verma, B.Tech.; M.S.E.

DISSERTATION

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT AUSTIN

August 2015

Dedicated to my parents.

Acknowledgments

I would like to express my sincere gratitude to a multitude of people. Firstly, I would like to thank my parents, Dr. Shiv Narayan and Shiromani, and my brother Neerav, who have been a constant source of moral support and inspiration for me. It is hard to express into words the sacrifices made by my parents in order to make sure that I receive the best possible education and environment to succeed. I would have certainly not made this far without their constant support and encouragement.

I would like to sincerely thank my Ph.D. adviser Dr. Mia K. Markey. Mia's research guidance and brainstorming sessions with her have greatly helped in shaping this dissertation work. I am especially thankful to Mia for the amount of independence and support she gave me throughout my Ph.D. to explore new research problems and collaborations. The opportunity to explore new areas and identify research problems has greatly helped me grow as a researcher. Besides Mia, I would also like to thank Dr. Matt Cowperthwaite, whose role during my Ph.D. has been no lesser than as my co-adviser. I have had the opportunity to work with Matt on a variety of different research problems since the first year of my Ph.D. He has always encouraged me in my research work and helped me out with resources that have been critical in completing the work presented in this dissertation.

I would like to thank my committee members Dr. Al Bovik, Dr. Kristen Grauman, and Dr. Joydeep Ghosh for their valuable suggestions on my dissertation research. Besides knowing them as my committee members, I also had the opportunity of take their classes. The knowledge that I gained from their classes has played an extremely important role in my Ph.D. research. I also had the privilege of working with Dr. Bovik on a research project and would like to thank him for his technical feedback, which greatly improved the quality of the research work.

I would also like to thank current and past BMIL members, particularly, Gautam, Gezheng, Juhun, Clement, Nisha, Daifeng, Shuang, Hans and Christos. I am specially thankful to Gautam, who has been like a mentor to me since my first year of Ph.D. I would also like to thank the administrative staff of Biomedical Engineering. In particular, Michael, Brittain, Margo, Carol, and Krystal have helped me through several administrative issues. Finally, I would like to express my sincere thanks to the UT Library System and the Texas Advanced Computing Center, without which this dissertation would not have been possible. I would also like to sincerely thank the NeuroTexas Institute Research Foundation at St. David's HealthCare in Austin, which provided financial support during several years of my Ph.D.

Nishant Verma

The University of Texas at Austin

June 2015

Biomarker for Tracking Progression of Alzheimer's Disease in Clinical Trials

Publication No. _____

Nishant Verma, Ph.D.
The University of Texas at Austin, 2015

Supervisor: Mia K. Markey

Currently, there are no treatments available for mitigating the neurological effects of Alzheimer's disease. All clinical trials of disease-modifying treatments, which showed promise in animal models, have failed to show a significant treatment effect in human trials. The lack of a sensitive outcome measure and the focus on the dementia stage for investigating treatments are believed to be the primary reasons behind the failure of all clinical trials till date. The currently used outcome measure, the Alzheimer's Disease Assessment Scale-Cognitive subscale (ADAS-Cog), suffers from low sensitivity in tracking progression of cognitive impairment in clinical trials. A shift in the focus to the prodromal mild cognitive impairment (MCI) stage may help improve the efficiency of clinical trials. However, even lower sensitivity of the ADAS-Cog and an inability to specifically select progressive MCI patients limit the efficiency of clinical trials in the MCI stage.

Cerebral atrophy measured on structural magnetic resonance (MR) imaging is highly promising for tracking disease progression in clinical trials. However, cerebral atrophy has not been yet approved as a valid biomarker due to the lack of an understanding behind its relationship with cognitive impairment. The focus of this dissertation spans across the two research areas of (i) developing automatic algorithms for analysis of patients' brain MR volumes, and (ii) improving the efficiency of clinical trials of disease-modifying treatments. This dissertation presents a novel knowledge-driven decision theory approach for automatic tissue segmentation of brain MR volumes, which shows better segmentation performance than the existing approaches.

The remaining dissertation contributions focus at improving the efficiency of clinical trials of disease-modifying treatments. An improved scoring methodology is presented for the ADAS-Cog outcome measure, which measures cognitive impairment with better accuracy and significantly improves the sensitivity of the ADAS-Cog in the mild-to-moderate Alzheimer's disease stage. However, the ADAS-Cog continues to suffers from low sensitivity in the MCI stage due to inherent limitations of its items. For improving the efficiency of clinical trials in the MCI stage, a biomarker has been developed that combines the ADAS-Cog with cerebral atrophy for more accurate tracking of Alzheimer's progression and facilitating selection of MCI patients in clinical trials.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background & Significance

Alzheimer's disease is the most common form of age-related dementia and is estimated to affect nearly 5.3 million elderly people in United States in 2015 [72, 1]. The prevalence in the US is expected to dramatically increase ($\sim$13.8 million by 2050) as advances in medicine improve life expectancy and the "baby boomer" generation enters the age range most susceptible to Alzheimer's disease [72]. Alzheimer's disease has significant implications on patients' quality of life and survival. The patients' quality of life is severely degraded due to the neuropsychiatric symptoms (such as depression, apathy, anxiety and agitation) and an inability to independently execute basic daily activities [155]. Alzheimer's disease is eventually fatal and reported to be the sixth leading cause of death in US [1]. Besides affecting patients' survival and quality of life, Alzheimer's disease also poses a tremendous public health burden in terms of patient care, healthcare costs (an estimated $226 billion in 2015 [1]), and caregivers' responsibilities.

There are no treatments currently available for mitigating the neuro-

logical effects of Alzheimer's disease. The routine clinical care of Alzheimer's disease patients involves symptomatic therapies (such as acetylcholinesterase inhibitors), which temporarily improve symptoms of Alzheimer's disease. However, effective strategies for the prevention and treatment of Alzheimer's disease are still lacking. All clinical trials of disease-modifying treatments, which showed promise in slowing down neurodegeneration in animal models, have failed in human trials [38]. There are two primary reasons believed to be behind the failure of all clinical trials till date. First, the primary outcome measure used in clinical trials suffers from low sensitivity in detecting treatment effects in clinical trials [27, 131, 55, 147]. The regulatory agencies require the primary end points in clinical trials to be characteristic symptoms of Alzheimer's disease, mainly cognitive and functional impairment [56]. A neuropsychological rating scale called the Alzheimer's Disease Assessment Scale-Cognitive subscale (ADAS-Cog) is currently used as the primary outcome measure in clinical trials to measure cognitive impairment of patients. However, the ADAS-Cog suffers from low sensitivity in measuring progression of cognitive impairment over short durations that are typically considered in clinical trials [27, 131, 75, 74]. This translates to low sensitivity in detecting treatment effects since treatments are evaluated based on their ability to slow down progression of cognitive impairment in Alzheimer's disease patients.

The second reason behind the failure of all trials is considered to be the advanced disease stage, where clinical trials have traditionally focused. Clinical diagnosis of probable Alzheimer's disease using the traditional criteria is

possible only at the dementia phase [107]. As a result, clinical trials have traditionally focused on mild-to-moderate Alzheimer's disease patients for evaluating disease-modifying treatments. However, substantial brain damage has already occurred by the time patients are diagnosed with Alzheimer's disease. Previous studies in animal models have suggested that disease-modifying treatments would be most effective in the early stages of Alzheimer's disease, when patients' brains have not yet undergone severe pathology and neurodegeneration [63, 40]. Alzheimer's disease is preceded by a prodromal stage of mild cognitive impairment (MCI) when the patients experience mild but noticeable changes in cognitive abilities but their ability to independently function in daily life activities is not affected [123]. Since MCI is the earliest stage when patients with Alzheimer's disease can be currently identified, a paradigm shift in the focus of clinical trials towards the MCI stage is underway. However, the population of MCI patients is highly heterogeneous in nature. While all Alzheimer's disease patients go through the prodromal MCI stage, only a small fraction ($\sim$10-15%) of MCI patients progress to Alzheimer's disease annually [123]. In fact, most MCI patients do not progress to dementia even after 10 years of follow-up [111]. Moreover, MCI is not specific to Alzheimer's disease and is also caused by other dementia types such as dementia with Lewy bodies. The currently employed clinical rating scales for MCI diagnosis are unable to rule out non-progressive MCI, let alone other pathologies, such as vascular and non-Alzheimer neurodegeneration [123]. The low specificity in early detection of Alzheimer's disease poses a big challenge for clinical trials focused in

the MCI stage. Besides patient selection, another challenge for clinical trials focused in the MCI stage is the lack of a sensitive outcome measure. The prodromal MCI stage is characterized by the lack of functional impairment in patients [123] and, therefore, cognitive impairment is the only possible primary end point in clinical trials based on the guidelines laid down by the regulatory agencies [56]. However, the sensitivity of the ADAS-Cog is even lower in the MCI stage as compared to the mild-to-moderate Alzheimer's disease stage.

Due to these reasons, clinical trials of disease-modifying treatments suffer from low efficacy both in the mild-to-moderate Alzheimer's disease and the MCI stages. If a treatment effect did exist, large sample sizes and long follow-up durations are required for detecting it using the ADAS-Cog as an outcome measure [55, 125]. This further exacerbates in the MCI stage, where a fraction of MCI patients may never develop Alzheimer's disease and, therefore, may not show any positive treatment effect. This has motivated the development of biomarkers, which can characterize the severity of Alzheimer's disease in patients. Such biomarkers would play an important role both as outcome measures and as inclusion criteria in clinical trials of disease-modifying treatments. Biomarkers would also be crucial in early detection of Alzheimer's disease and facilitating timely intervention using disease-modifying treatments, as they become available. Noting their high significance, academia, industry, and regulatory agencies have prescribed that a valid biomarker should be (a) measured using reliable and validated methods, (b) sensitive and specific as a diagnostic marker, (c) sensitive in measuring effects of treatments on dis-

ease progression, and (d) predictive of clinical outcomes, such as cognitive and functional impairment [59, 60, 69]. Several different classes of biomarkers are under investigation for Alzheimer's disease [82, 81, 83]. In this dissertation, we focus specifically on structural imaging biomarkers, which are highly promising for tracking progression of Alzheimer's disease.

1.2 Cerebral Atrophy due to Alzheimer's Disease

Pathologically, Alzheimer's disease is believed to be caused by progressive deposition of amyloid plaques and neurofibrillary tangles in patients' brains, which eventually leads to neuronal loss and cognitive decline. The current hypothesis of Alzheimer's disease pathological cascade (figure 1.1) is a two-stage process, where deposition of amyloid plaques and neuronal damage happen sequentially rather than simultaneously [82, 81, 83, 113]. Biomarkers that measure concentration of abnormal proteins (CSF Aβ_{42}, CSF tau and amyloid PET in figure 1.1) show abnormality during the asymptomatic stage of Alzheimer's disease, which can be as long as decades before the onset of symptoms [83, 113]. However, these biomarkers have almost plateaued by the time patients start to manifest any symptoms in the MCI stage [81]. Therefore, while protein-based biomarkers may be useful for patient screening, their potential as outcome measures in clinical trials is limited.

On the other hand, biomarkers that measure the extent of neuronal damage in patients' brains (MRI and FDG-PET in figure 1.1) are sensitive to

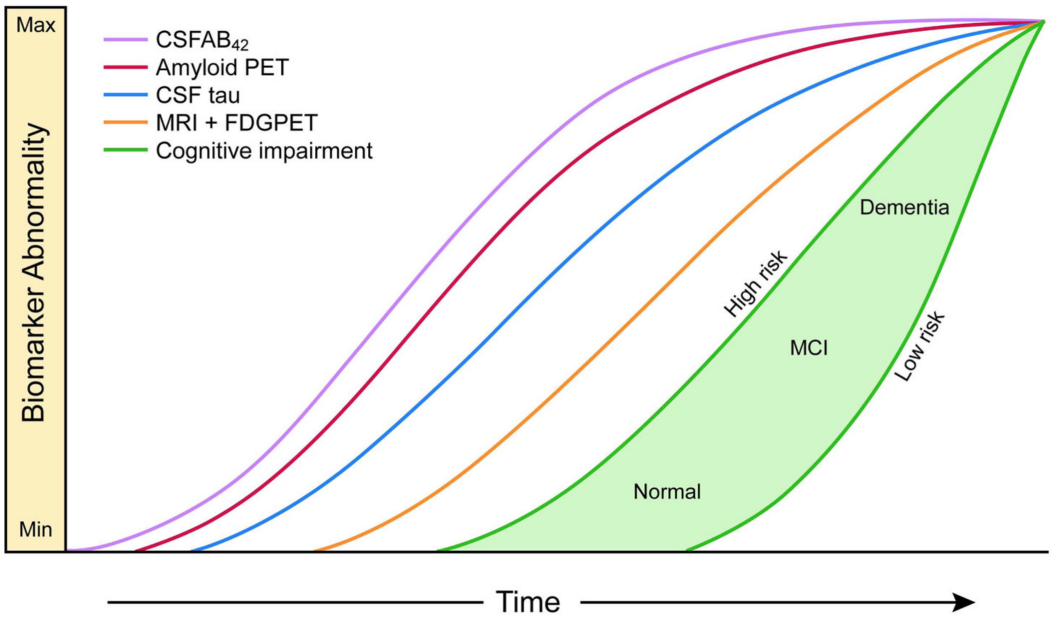
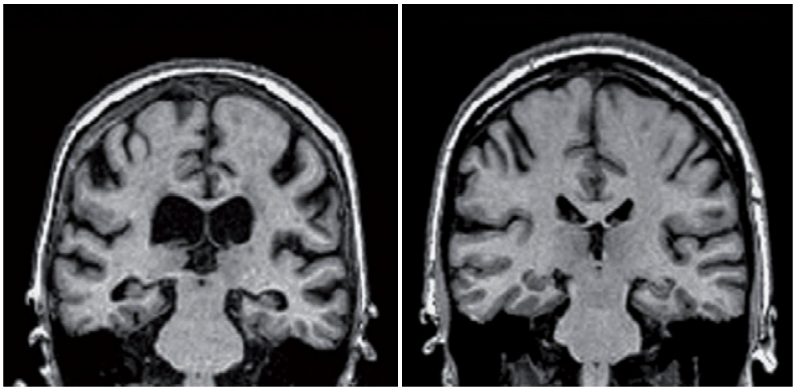


Figure 1.1: Hypothesized pathological cascade of Alzheimer's disease [81]: Illustration showing temporal ordering of the following biomarkers: amyloid plaques measured in CSF ($CSFAB_{42}$), amyloid deposition observed on positron emission tomography (Amyloid PET), hyperphosphorylated tau levels in CSF (CSF tau), cerebral atrophy on magnetic resonance imaging (MRI), metabolism on PET (FDGPET), and cognitive impairment in patients. Since patients show significant variability in progression rates of cognitive impairment, cognitive impairment is depicted as a band with the two edges representing low-risk and high-risk patients.

disease progression in the MCI and the mild-to-moderate Alzheimer's disease stages. Neuronal and synaptic losses in patients' brains manifest macroscopically as cerebral atrophy [18], which can be measured on structural magnetic resonance (MR) imaging volumes of patients (figure 1.2). Structural MR imaging is routinely employed in clinical trials of disease-modifying treatments for excluding patients with other dementia types and evaluating safety of inves-



(a) (b)

Figure 1.2: Cerebral atrophy due to Alzheimer's disease: Matched MR slices showing significantly reduced brain tissue and enlarged ventricles in (a) an Alzheimer's disease patient as compared to (b) an age matched normal control.

tigative treatments. The high sensitivity and ease of implementation makes cerebral atrophy a highly promising biomarker for Alzheimer's disease in clinical trials. The validity of cerebral atrophy as a biomarker is further justified by the fact that cerebral atrophy maps accurately upstream to the Braak stages of pathology deposition at autopsy [192, 176] and downstream to cognitive impairment in patients [42, 167, 169, 178].

1.3 Dissertation Contributions

The promise of cerebral atrophy as a biomarker for Alzheimer's disease has led to significant efforts in two primary research areas. The first research area is concerned with the development of automatic algorithms for analyzing brain MR volumes and quantifying cerebral atrophy. This includes

a wide variety of image analysis tasks such as skull extraction [160], artifact removal [179, 132, 157], tissue segmentation [183, 182, 39, 50], brain cortical parcellation [53, 44, 54], and cortical thickness measurement [90, 145, 70, 134], which are typically involved in any image analysis pipeline for measurement of cerebral atrophy in Alzheimer's disease patients. The second research area involves the development of strategies for improving efficiency of clinical trials of Alzheimer's disease-modifying treatments. This has primarily involved improving the currently used tools and integrating them with cerebral atrophy measurements for better efficiency of clinical trials.

The contributions of this dissertation fall under both these areas of research. As part of the first research area, the contribution of this dissertation is a knowledge-driven decision theory (KDT) approach for segmentation of brain MR volumes (chapter 3). Automatic segmentation of brain MR volumes into tissue classes of white matter, gray matter, and cerebrospinal fluid is a prerequisite step for measuring cerebral atrophy in patients. While easy for humans, automatic tissue segmentation is a complicated task due to significant amounts of image corruptions that are typically present in MR volumes. The proposed KDT approach is motivated from an observation that relative extents of intensity overlap between tissue class pairs stay roughly consistent across MR volumes. We investigated whether the incorporation of prior knowledge on intensity overlaps can improve classification of voxels residing in the intensity overlap spectrum. The segmentation performance of the proposed KDT approach was evaluated on two standardized MR segmentation data sets and

compared against the performance of existing segmentation approaches.

The other contributions of this dissertation fall under the second research area of improving the efficiency of clinical trials of Alzheimer's disease-modifying treatments. The first contribution is an improved ADAS-Cog scoring methodology for measuring cognitive impairment in Alzheimer's disease patients (chapter 5). As discussed earlier, the ADAS-Cog is the standard primary outcome measured used in clinical trials involving mild-to-moderate Alzheimer's disease patients. However, the ADAS-Cog suffers from low sensitivity in measuring progression of cognitive impairment. We identified that a major reason behind the low sensitivity of the ADAS-Cog is its sub-optimal scoring methodology, which is used to score cognitive impairment in patients. We developed a new scoring methodology for the ADAS-Cog based on a comprehensive psychometric analysis using item response theory (ADAS-CogIRT), which addresses several major limitations associated with the current scoring methodology. The sensitivity of the ADAS-CogIRT methodology was evaluated and compared against the current scoring methodology using simulated clinical trials and a real clinical trial, which had shown an evidence of a treatment effect in the original negative trial. The ADAS-CogIRT scoring methodology significantly improves the sensitivity of the ADAS-Cog in detecting treatment effects in clinical trials focused in the mild-to-moderate Alzheimer's disease stage. It also allows separate evaluation of treatment effects in the cognitive domains of memory, language, and praxis, which is not possible using the current scoring methodology.

While the ADAS-CogIRT scoring methodology improves the sensitivity of the ADAS-Cog, it is unable to overcome the inherent limitations of the ADAS-Cog items. As a result, the ADAS-CogIRT suffers from low sensitivity in the MCI stage. The final contribution of this dissertation is a biomarker that combines the ADAS-Cog with cerebral atrophy for even more accurate tracking of progression in clinical trials (chapter 7). Despite its promise, cerebral atrophy is not approved as a valid biomarker of Alzheimer's disease due to the lack of an understanding of the relationship between cerebral atrophy and cognitive impairment. As part of this contribution, we investigated this relationship and found that the spatio-temporal pattern of cerebral atrophy is closely related with progression of cognitive impairment in Alzheimer's disease patients. We developed a biomarker that combines the ADAS-Cog responses of patients with cerebral atrophy on MR imaging (ADAS-CogMRI) for tracking cognitive impairment in clinical trials. We evaluated the ADAS-CogMRI biomarker and compared it against the sole use of the ADAS-Cog and cerebral atrophy in simulated clinical trials focused in the mild-to-moderate Alzheimer's disease and the MCI stages. We validated the simulation results in a real world problem posed as a clinical trial of a disease-modifying treatment, which is hypothesized to prevent conversion of MCI patients to Alzheimer's disease. The ADAS-CogMRI biomarker significantly improves the efficacy of clinical trials focused in the MCI stage and shows good accuracy in predicting MCI patients that will convert to Alzheimer's disease in future.

1.4 Dissertation Outline

The remainder of this dissertation is organized as follows. Chapter 2 provides a brief background on brain MR tissue segmentation and reviews existing segmentation approaches. In chapter 3, we describe the knowledge-driven decision theory segmentation approach, evaluate its performance, and compare against the existing segmentation approaches. Chapter 4 briefly describes the ADAS-Cog assessment and discusses the limitations associated with its current scoring methodology in clinical trials. Chapter 5 presents the new ADAS-CogIRT scoring methodology for the ADAS-Cog, evaluates its performance in clinical trials, and compares against the currently used scoring methodology. Chapter 6 discusses the relationship between cerebral atrophy and cognitive impairment, and the promise of combining them into a biomarker of Alzheimer's disease. In chapter 7, we investigate the relationship between brain-wide cerebral atrophy and cognitive impairment, develop a biomarker based on this relationship, evaluate the performance of the biomarker, and compare it against the sole use of the ADAS-Cog and cerebral atrophy in clinical trials. In chapter 8, we conclude this dissertation discussing the significance of its contributions and some pointers for future research work.

Chapter 2

Volumetric Brain MR Segmentation

2.1 Introduction

Magnetic resonance (MR) imaging is routinely used to obtain detailed anatomical information about patients' brains. Structural changes observed on MR imaging are clinically significant for diagnostic and treatment planning purposes for several neurological diseases [23, 62, 64]. However, due to the large amount of data collected in MR imaging, manual structural measurements (such as cortical thickness) are tedious and time intensive. This has motivated the development of computer-based tools to quantify structural changes on MR volumes that are caused by neurological disorders such as Alzheimer's disease.

MR tissue segmentation is an important, and often prerequisite, component of any comprehensive MR image analysis. This involves classifying brain MR voxels into four classes: white matter (WM), gray matter (GM), cerebrospinal fluid (CSF), and background (BG) as shown in figure 2.1. However, automatic tissue segmentation in brain MR volumes is difficult, due to the presence of image corruptions such as partial volume effects and intensity inhomogeneities (or bias field). Accurate segmentation of MR volumes requires

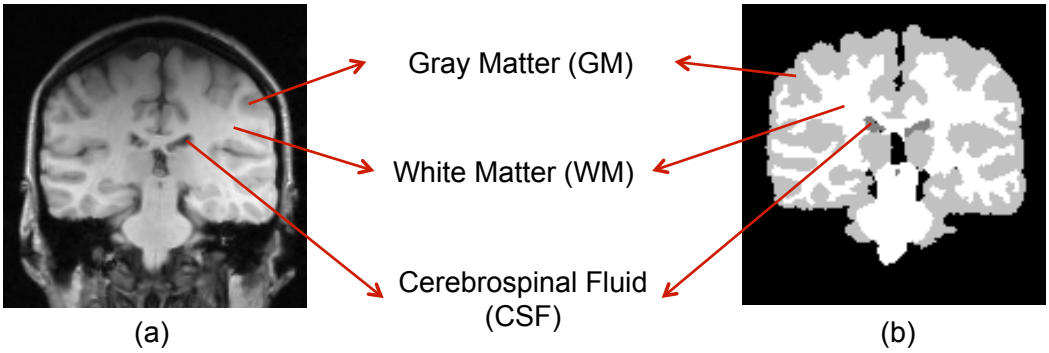


Figure 2.1: Brain MR tissue segmentation: Figure showing a (a) sample slice from a MR volume with intensity inhomogeneities, and (b) labeled tissue classes of white matter, gray matter, and cerebrospinal fluid in the MR slice. The intensity inhomogeneity in the MR slice is seen as a smoothly varying shading artifact such that the upper portion of the slice appears darker than the bottom portion.

incorporating the contributions from such image corruptions while classifying MR voxels into the four classes.

The most common segmentation methods are probabilistic formulations that represent an MR volume with a parametric model such as finite mixture model (FMM) with four Gaussian components [104, 142, 174, 195]. Thereafter, a classification rule attributes class labels to every voxel in the MR volume. However, the presence of image corruptions greatly skews the distribution of voxel intensities in MR volumes. As a result, tissue classes have arbitrarily shaped and variable density functions of intensities in MR volumes, which are difficult to represent using an a priori assumed parametric model. To overcome this, the use of more flexible parametric models has been suggested [36, 49, 67, 80, 135]. While flexible modeling of intensity density

functions yields improved segmentation performance [36, 67, 98], such methods still suffer from the specification bias of the assumed parametric models [91]. Most of the modeling errors are concentrated along the tails of the intensity distributions, which are the regions of intensity overlap between the tissue classes. Therefore, the specification bias of the assumed parametric models directly translates to errors in voxel classification. Several non-parametric approaches (such as kernel density estimation) have also been employed for modeling tissue intensity distributions in MR volumes [4, 100, 136]. They provide better flexibility in modeling arbitrary intensity distributions and show improved tissue segmentation performance [4, 100, 136].

Besides producing arbitrarily shaped intensity density functions, the presence of intensity inhomogeneities (figure 2.1) also results in significant overlap between the intensity density functions of tissue classes. Most segmentation errors are the result of inaccurate classification of MR voxels that reside in this spectrum of intensity overlap and produce similar likelihoods of membership to multiple tissue classes. To minimize such errors, preprocessing methods are typically employed to reduce the effect of intensity inhomogeneities in MR volumes [132, 157]. However, the performance of subsequent tissue segmentation becomes sensitive to the accuracy of the preprocessing methods used to remove intensity inhomogeneities from MR volumes. Moreover, the computation complexity associated with such methods is generally very high.

2.2 Relevant Work

In this section, we review existing tissue segmentation approaches and their methods for voxel classification, modeling tissue classes, prior specification, and energy minimization. Most existing methods have used a Bayesian maximum a-posteriori (MAP) formulation for tissue segmentation and minimized it using the expectation maximization (EM) algorithm. Wells *et al.* [187] proposed an adaptive MAP method for simultaneous MR tissue segmentation and intensity inhomogeneity estimation. Leemput *et al.* [174] extended this approach by using probabilistic atlases for automatic modeling of tissue classes. Marroquin *et al.* [104] also presented a Bayesian MAP formulation for tissue segmentation along with a variant of the EM algorithm for more efficient energy minimization. Adaptive pixon represented segmentation (APRS) method by Lin *et al.* [100] used a MAP formulation but their formulation involved clusters of connected pixels (pixons) rather than individual pixels. Several other well-known segmentation approaches have also used a MAP formulation for driving tissue segmentation [135, 67, 126].

Most of the MAP formulations of tissue segmentation have assumed a parametric Gaussian distribution of intensities within each tissue class [174, 187, 135, 100]. As discussed earlier and also noted by Prastawa *et al.* [126], intensity distributions of tissue classes show significant overlap and modeling with Gaussian distributions results in degenerate decision boundaries. As a result, some segmentation approaches have considered use of alternate para-

metric models for tissue intensities. Marroquin *et al.* [104] assumed a parametric model of spline models with a Gibbsian prior for modeling tissue classes. The constrained Gaussian mixture model (CGMM) framework by Greenspan et al. [67] utilized a mixture of large number of Gaussian components to represent individual tissue classes. However, the intensity parameters of all Gaussian components representing each tissue class were constrained to be equal, which limits the ability of CGMM method to model arbitrary intensity distributions of tissue classes. Kernel density estimation (KDE) or parzen-window estimation has also been previously used in segmentation approaches to model arbitrary intensity distributions inside tissue classes [11, 126, 105, 86]. Awate *et al.*[11] developed an unsupervised tissue segmentation method that adaptively learns image-neighborhood Markov statistics and entails estimation of intensity distributions using parzen-window estimation. KDE has also been utilized in two mean shift inspired approaches of the adaptive mean-shift (AMS) method by Mayer *et al.* [105] and the mean shift method with edge confidence maps (MSECM) by Jimenez-Alaniz *et al.* [86].

MRF based contextual priors and probabilistic tissue atlases are often used for defining prior anatomical information and guide tissue segmentation [24]. Leemput *et al.* [173] proposed an approach (KVL) that combined tissue atlases with MRF priors to define tissue priors and illustrated significant improvement in segmentation performance. A similar approach was also followed by the MPM-MAP [104] and APRS [100] segmentation methods for defining tissue class priors. Rivera *et al.* [135] used a modified MRF methodology

involving quadratic potentials, which allowed for computation of probability estimates for voxels belonging to all tissue classes. The segmentation approach by Awate *et al.* [11] used tissue atlases only for initialization purposes. However, since their segmentation framework implicitly incorporated MRF based smoothness constraints, their approach also utilizes atlases and MRF contextual priors for guiding tissue segmentation. The sub-volume probabilistic atlas segmentation (SVPASEG) method by Tohka *et al.* [170] also utilized a MRF based framework with tissue atlases used for dividing MR volumes into different domains. KDT uses a similar approach as the ones utilized in KVL, MPM-MAP, and APRS for defining tissue priors.

While level set-based approach for energy minimization has been extensively used for segmentation of natural scene images [35, 28], its application in MR tissue segmentation has been relatively scarce. Level set-based approach is highly flexible and enables representation of energy functions containing wide varieties of energy terms (such as local region, smoothness and area terms). The ease of implementation also makes it an attractive framework for representing brain MRI tissue segmentation models. Some level set-based methods have been developed for brain tissue segmentation that illustrated impressive results [184, 96, 98, 95]. However, the relative value of level sets in comparison to alternate energy minimization strategies is difficult to appreciate due to poor documentation of level set-based methods on well-established segmentation datasets.

Chapter 3

Knowledge-driven Decision Theory for Volumetric Brain MR Segmentation

3.1 Introduction

Accurate MR tissue segmentation requires precise modeling of tissue classes and a classification rule that takes into account the effects from image corruptions. In this chapter, we present a new 3D knowledge-driven decision theory (KDT) approach towards handling the intensity overlap across tissue classes[a,b]. The approach is motivated by an observation that tissue class pairs have different relative extents of intensity overlap in MR volumes. In the presence of image corruptions (such as bias field), the intensity overlap between tissue classes increases; however, the relative proportions stay approximately the same across different MR volumes. The incorporation of intensity overlap knowledge in the segmentation model enables more accurate classification of

[a]N. Verma, G. S. Muralidhar, A. C. Bovik, M. C. Cowperthwaite, M. G. Burnett, M. K. Markey, "Three-dimensional brain magnetic resonance imaging segmentation via knowledge-driven decision theory", *Journal of Medical Imaging*, 1(3):034001-034015, 2014.

[b]N. Verma, G. S. Muralidhar, A. C. Bovik, M. C. Cowperthwaite, M. K. Markey, "Model-driven probabilistic level set based segmentation of magnetic resonance images of the brain", In *Proc. Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, p. 2821-2824, 2011.

voxels residing in the intensity overlap spectrum. In KDT, a decision theory based objective function is minimized using a variational level set-based approach.

Variational segmentation methods have gained popularity for brain MR segmentation [98, 136]; however, their performance on well-known datasets is poorly documented. This makes it difficult to establish their potential in comparison to other energy minimization strategies (such as graph cuts). We evaluate our approach using two well-established datasets from the Internet Brain Segmentation Repository and compare against segmentation methods that used different energy minimization techniques. The segmentation approach described in this chapter was published in the SPIE Journal of Medical Imaging in 2014 [182][a]. A preliminary version of this study was also presented and published in the proceedings of the 33^{rd} Annual International Conference of the IEEE Engineering in Medicine & Biology Society in 2011 [183][b]. In both these works, N. Verma developed the methods, performed the analysis, and prepared the manuscripts. G. S. Muralidhar contributed towards the development of the methods, study designs, and preparation of the manuscripts. The rest of the co-authors helped with the study designs and preparation of the manuscripts.

The chapter is organized as follows: Section 3.2 describes the proposed KDT algorithm for tissue segmentation and provides details on its numerical implementation. Section 3.5 evaluates the segmentation performance of KDT,

compares performance with existing methods, illustrates the significance of KDT's individual components, and performs computational complexity analysis. Finally, Section 3.6 summarizes the technical contributions of this study and discusses the advantages and limitations of KDT.

3.2 Mathematical Notations

We define some notations that are frequently used in this chapter. Given an MR volume V defined as a function $V : \Omega \rightarrow \mathbb{R}$ on a continuous 3D domain Ω, the goal of tissue segmentation is to partition Ω into four disjoint classes $C \in \{WM, GM, CSF, BG\}$. Any MR voxel is hence defined by its spatial location (or coordinates) $x \in \Omega$ and associated MR signal/intensity value $V(x)$. Besides the spatial image domain, we also interpret KDT in the intensity range domain. The intensity range domain for a given MR volume V is defined by the space of all possible voxel intensities $\mathcal{I} \in \mathbb{I}$, where $\mathbb{I} = \big[\min_{x \in \Omega}[V(x)], \max_{x \in \Omega}[V(x)]\big]$.

3.3 Motivation

The motivation behind our approach is the observation that the relative extents of intensity overlap between different tissue class pairs are not equal and follow a consistent trend across MR volumes. We illustrate this fact by calculating the intensity overlap areas between all tissue class pairs $k, j \in$

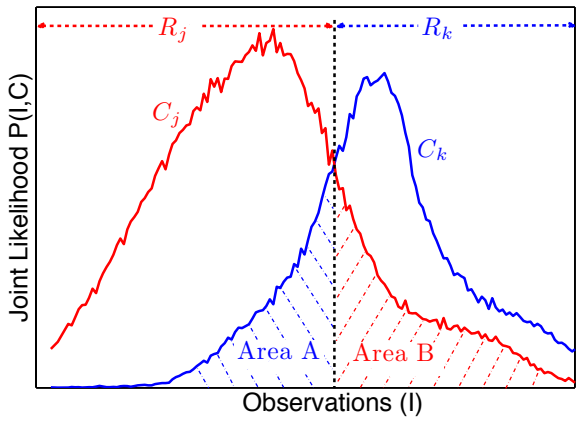


Figure 3.1: Relative intensity overlap between the tissue classes: Schematic illustration showing the partial intensity overlap areas Area A $= \int_{R_j} P(\mathcal{I}, C_k)d\mathcal{I}$ and Area B $= \int_{R_k} P(\mathcal{I}, C_j)d\mathcal{I}$.

$\{WM, GM, CSF\}$ using the expert ground truth segmentations. Intensity overlap area $Overlap(C_k, C_j)$ between tissue classes C_k and C_j is defined as

$$Overlap(C_k, C_j) = \int_{R_k} P(\mathcal{I}, C_j)d\mathcal{I} + \int_{R_j} P(\mathcal{I}, C_k)d\mathcal{I} \tag{3.1}$$

where $\mathcal{I}$ denotes the voxel intensities in the MR volume; $P(\mathcal{I}, C)$ denotes the likelihood of voxel intensity $\mathcal{I}$ belonging to class C; and R_k and R_j represent the intensity ranges defined as $R_k = \{\mathcal{I} : P(\mathcal{I}, C_k) > P(\mathcal{I}, C_j), \mathcal{I} \in \mathbb{I}\}$ and $R_j = \{\mathcal{I} : P(\mathcal{I}, C_j) > P(\mathcal{I}, C_k), \mathcal{I} \in \mathbb{I}\}$, respectively (as illustrated in figure 3.1). Figure 3.2a shows the overlap areas between tissue pairs WM & GM and GM & CSF relative to the overlap areas between WM & CSF. The scatterplot is generated using expert ground truth segmentations of 18 real MR volumes from the Internet Brain Segmentation Repository.

The consistent pattern across MR volumes suggests that the extents of

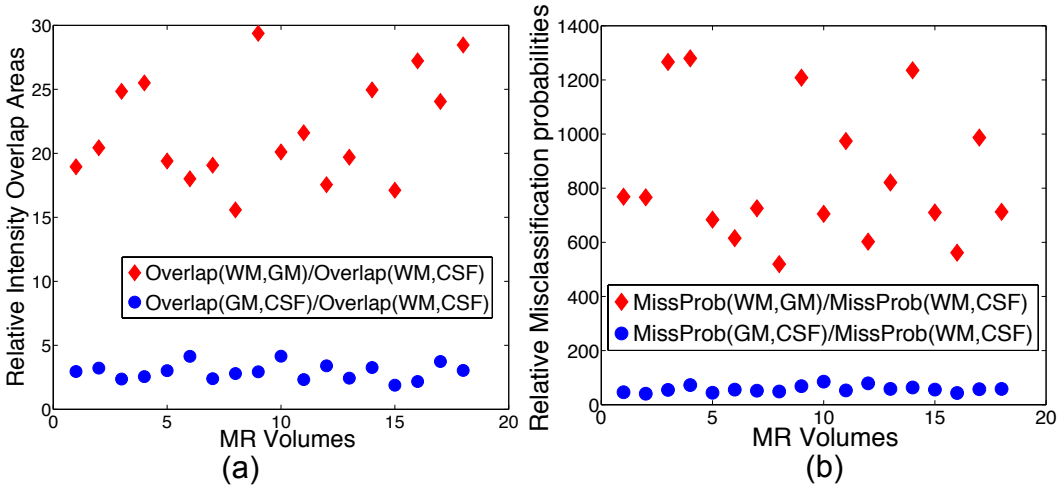


Figure 3.2: Relative intensity overlap between tissue classes: (a) scatterplot of overlap areas $Overlap(WM, GM)$ and $Overlap(GM, CSF)$ across MR volumes relative to $Overlap(WM, CSF)$, and (b) scatterplot comparing the aggregate misclassification probabilities between WM & GM and GM & CSF across MR volumes (same as in (a)) relative to the total misclassification probabilities between WM & CSF. The aggregate misclassification probability between two classes C_j and C_k is defined as: $MissProb(C_j, C_k) = \int_{x \in C_j} P(x, C_k)dx + \int_{x \in C_k} P(x, C_j)dx$.

intensity overlap are different among tissue class pairs. In terms of magnitude, the overlap area between WM & GM is higher than the overlap area between GM & CSF and between WM & CSF. Some of the MR volumes in figure 3.2a contain high levels of intensity inhomogeneities, which show increased overlap areas between the tissue class pairs (such as MR volumes 3 and 10). While we have simply combined the partial overlap areas between tissue classes (areas A and B in figure 3.1) for illustrating that the intensity overlap areas are not equal, asymmetry may exist between the partial overlap areas and has been considered for investigation in our experiments. The relative magnitudes of

overlap areas in MR volumes is a combined effect of several factors such as the lengths of boundaries between tissue types, extent of intensity inhomogeneities, partial volume effects and contrast between the tissue types.

3.4 Knowledge-driven Decision Theory (KDT)

Noting this observation, we now formally present the knowledge-driven decision theory (KDT) algorithm for MR tissue segmentation. We use a Bayesian decision theory framework for integrating knowledge of the relative extents of intensity overlap between tissue class pairs. A loss matrix L is defined, where each element $L_{k,j}$ represents the loss incurred if a voxel from tissue class C_k is classified as belonging to class C_j. Therefore, the total expected loss $\mathbb{E}$ due to classification of voxels $x \in \Omega$ can be defined as

$$\mathbb{E} = \sum_{k} \sum_{j} \int_{x \in C_j} L_{k,j} \times P(x, C_k) dx \tag{3.2}$$

where $P(x, C_k)$ denotes the joint likelihood of voxel x belonging to class C_k. Decision theory has been traditionally used to determine optimum decision boundaries incurring the least expected loss in the class likelihood space based on the loss values ($L_{k,j}$ and $L_{j,k}$) and the overlap between the distributions $P(x, C_k)$ and $P(x, C_j)$ [16]. Since the class distributions $P(x, C)$ for MR volumes are unknown a priori, decision theory has been rarely applied for MR tissue segmentation [127].

We utilize the expected loss $\mathbb{E}$ to iteratively influence the decision

boundaries such that the final voxel classification produces an intensity overlap similar to figure 3.2a. The energy function (3.2) can be interpreted as a weighted sum of the intensity overlap areas between the tissue class pairs. To understand this, it is important to note the relationship between the two terms: (i) $\int_{x \in C_j} P(x, C_k)dx$ in (3.2) measuring the aggregate probability of misclassification of voxels belonging to class C_k into class C_j , and (ii) $\int_{\mathcal{I} \in R_j} P(\mathcal{I}, C_k)d\mathcal{I}$ in (3.1) measuring the partial overlap area (in likelihood space) between classes C_k and C_j in the intensity range of C_j. While the aggregate probability term is calculated over the image region C_j and the partial overlap term is calculated over the intensity range R_j, they both intrinsically measure the same underlying effect. The aggregate probability term is simply the value of partial intensity overlap area scaled with the number of voxels belonging in the overlap area. This relationship can be observed in the scatterplot (figure 3.2b) that shows the aggregate misclassification probabilities between class pairs WM & GM and GM & CSF relative to the aggregate misclassification probabilities between WM & CSF for the same 18 MR volumes. A comparison with the relative intensity overlaps (in figure 3.2a) shows that the relative misclassification probabilities follow a very similar trend across all MR volumes.

MR tissue segmentation based solely on intensity overlaps is sensitive to the presence of image corruptions (such as MR noise). Therefore, in KDT, we define the joint voxel likelihoods $P(x, C_k)$ using a intensity term $P(V(x)|C_k)$

and a spatial prior term $P_{C_k}(x)$,

$$\mathbb{E} = \sum_k \sum_j \int_{x \in C_j} L_{k,j} \times P(V(x)|C_k) \times P_{C_k}(x) dx \tag{3.3}$$

where $V(x)$ denotes the intensity value of the MR volume at voxel location $x \in \Omega$; $P(V(x)|C_k)$ is the likelihood of MR intensity value $V(x)$ belonging to class C_k; and $P_{C_k}(x)$ denotes the prior probability of class C_k at a location x in the MR volume. In (3.3), tissue segmentation is primarily driven by the intensity term $P(V(x)|C_k)$ that controls the relative extents of intensity overlap between tissue class pairs. The spatial priors help identify the tissue types and reduce KDT's sensitivity to image corruptions. The following sections provide detailed descriptions of the likelihood $P(V(x)|C_k)$ and the prior $P_{C_k}(x)$ terms.

3.4.1 Modeling Arbitrary Tissue Intensity Distributions

$P(V(x)|C)$ is estimated by modeling the arbitrarily shaped density functions of intensities inside the classes $C \in \{WM, GM, CSF, BG\}$. Assuming parametric models for intensities results in inaccurate modeling of the tissue classes. The estimation errors are mostly concentrated along the tails of the intensity density functions, which are the major regions of intensity overlaps between the classes. Therefore, accurate modeling of the arbitrarily intensity density functions inside tissue classes is essential for KDT. We use a nonparametric method of adaptive kernel density estimation (KDE) based on linear diffusion processes [20] to model the intensity distributions inside the tissue classes. Adaptive KDE is specifically selected over other KDE methods

because adaptive KDE has better local adaptivity, lower sensitivity to outliers, lower boundary bias, and can handle data that are not normally distributed [103, 120, 166, 20]. These properties become significant in MR volumes due to the non-negative nature of intensity data, presence of outliers (such as noise and artifacts), and intensity distributions that are not normally distributed.

3.4.2 Adaptive Tissue Class Priors

A combination of probabilistic atlas maps and Markov random field (MRF) based contextual priors is used for defining tissue class priors $P_C(x)$ in KDT. Such tissue class priors are commonly employed to guide MR tissue segmentation and reduce sensitivity to image corruptions [174, 104]. In KDT, we use adaptive class priors that are initialized with atlas maps and iteratively superimposed with MRF contextual priors:

$$P_C(x, n+1) = (1 - w) \times P_C(x, n) + w \times P_C^{MRF}(x, n) \tag{3.4}$$

where $P_C^{MRF}(x, n)$ denote the MRF contextual priors computed at iteration n and w is an adaptive weight that controls the contribution of $P_C^{MRF}(x, n)$ in class priors $P_C(x)$ at iteration $n + 1$. The MRF contextual priors are characterized using Potts model [10, 17, 21]:

$$P_C^{MRF}(x) = exp(-\sum_{p \in P} \delta(C, C(p)))/Z \tag{3.5}$$

where Z is a normalizing constant; $p \in \mathcal{P}$ represent all possible cliques (set of voxels) of size two in a six-neighborhood system (in 3D) around voxel location

x; $\delta(.)$ represents the Dirac delta function; and $C(p)$ denote the classes of voxels contained in clique p. The adaptive class priors are initialized with tissue atlas maps $P_C(x, n = 0) = P_C^{Atlas}(x)$ spatially aligned with the MR volume using affine registration.

The expected loss function $\mathbb{E}$ in (3.3) is minimized iteratively by drawing decision boundaries in the likelihood space based on the loss matrix values and the tissue class distributions $P(x, C)$. Any perturbations in the decision boundaries change the voxel classification, which in turn, change the tissue class distributions. The loss matrix is determined such that the final segmentation produces an intensity overlap profile as observed in figure 3.2b. We can relate the energy function in (3.3) with the maximum a posteriori (MAP) classification, which is often used for MR segmentation. In MAP, the objective function is minimized by choosing the tissue classes with maximum posterior probabilities for MR voxels. This decision rule is equivalent to minimizing $\mathbb{E}$ in (3.3) when same loss values are considered for all tissue misclassifications: $L_{i,j} = k$ (constant) $\forall i, j \neq i$ and $L_{i,i} = 0$. The equal misclassification loss values imply that all overlap areas are penalized equally, which would result in equal overlap areas between all tissue class pairs. In Section 3.5.6.2, we quantitatively evaluate the effect of unequal loss values by comparing against MAP for voxel classification.

3.4.3 Energy Minimization using Level Sets Framework

The energy function in (3.2) is difficult to minimize in terms of the evolving image regions $C_j \in \{WM, GM, CSF, BG\}$. A level set formulation enables representation of the regions C_j in terms of higher dimensional level set functions $\Phi : \Omega \to \mathbb{R}$. Each level set Φ partitions the image domain Ω into two disjoint sub-domains $\Omega_1 = \{x \in \Omega : \Phi(x) > 0\}$ and $\Omega_2 = \{x \in \Omega : \Phi(x) < 0\}$. Therefore, two level sets Φ_1, Φ_2 can be simultaneously used to represent the four classes:

$$\Omega = \begin{cases} C_1(WM) & \Phi_1 > 0, \Phi_2 > 0 \\ C_2(GM) & \Phi_1 < 0, \Phi_2 > 0 \\ C_3(CSF) & \Phi_1 > 0, \Phi_2 < 0 \\ C_4(BG) & \Phi_1 < 0, \Phi_2 < 0. \end{cases}$$

The energy function 3.2 for the 4 classes can be written as

$$\mathbb{E} = \sum_{j=1}^{4} \int_{x \in C_j} \Big(\sum_{k=1}^{4} L_{k,j} \times P(V(x)|C_k) \times P_{C_k}(x) \Big) dx \tag{3.6}$$

For notational simplicity, we represent the total expected loss due to classification of voxels into class C_j by $E_j = \sum_{k=1}^{4} L_{k,j} \times P(V(x)|C_k) \times P_{C_k}(x)$. Using the Heaviside function $H(\Phi)$ and the Dirac delta function $\delta(\Phi)$,

$$H(\Phi) = \begin{cases} 1 & \text{if } \Phi \geq 0 \\ 0 & \text{if } \Phi < 0 \end{cases}, \qquad \delta(\Phi) = \frac{d}{d\Phi} H(\Phi)$$

the energy function in (3.6) can be represented as:

$$\begin{aligned} \mathbb{E}(\Phi_1, \Phi_2) = \int_{x \in \Omega} \Big[& E_1 H(\Phi_1) H(\Phi_2) + E_2 H(-\Phi_1) H(\Phi_2) \\ & + E_3 H(\Phi_1) H(-\Phi_2) + E_4 H(-\Phi_1) H(-\Phi_2) \Big] dx \end{aligned}$$

$\mathbb{E}(\Phi_1, \Phi_2)$ is used as the data term in the level set energy functional $\mathbb{F}(\Phi_1, \Phi_2) = \mathbb{E}(\Phi_1, \Phi_2) + \mu \times \mathbb{R}(\Phi_1, \Phi_2)$, where $\mathbb{R}(\Phi_1, \Phi_2)$ is a regularization term with weight μ on the evolving level set functions $\Phi_1(x)$ and $\Phi_2(x)$. $\mathbb{R}(\Phi_1, \Phi_2)$ ensures smoothness of the level set functions Φ_1 and Φ_2 by penalizing the arc length of their zero level contours (tissue boundaries):

$$\mathbb{R}(\Phi_1, \Phi_2) = \int\limits_{x \in \Omega} \delta(\Phi_1)|\nabla \Phi_1| dx + \int\limits_{x \in \Omega} \delta(\Phi_2)|\nabla \Phi_2| dx$$

The energy functional $\mathbb{F}(\Phi_1, \Phi_2)$ can hence be represented

$$\begin{aligned} \mathbb{F}(\Phi_1, \Phi_2) = \int\limits_{x \in \Omega} \Big[& E_1 H(\Phi_1) H(\Phi_2) + E_2 H(-\Phi_1) H(\Phi_2) \\ & + E_3 H(\Phi_1) H(-\Phi_2) + E_4 H(-\Phi_1) H(-\Phi_2) \\ & + \mu \times \delta(\Phi_1)|\nabla \Phi_1| + \mu \times \delta(\Phi_2)|\nabla \Phi_2| \Big] dx \end{aligned} \tag{3.7}$$

In MR volumes, the image domain Ω is a 3D Cartesian grid where any location $x \in \Omega$ is defined by coordinates $x = \{x_1, x_2, x_3\}$ along the three orthogonal axes. The energy functional (3.7) can be rewritten as

$$\mathbb{F}(\Phi_1, \Phi_2) = \int\limits_{x \in \Omega} G(\Phi_1, \Phi_2, \Phi_{1,1}, \Phi_{2,1}, \Phi_{1,2}, \Phi_{2,2}, \Phi_{1,3}, \Phi_{2,3}) dx_1 dx_2 dx_3 \tag{3.8}$$

where G is a real-valued function of level sets Φ_1, Φ_2 and their derivatives. The partial derivative of Φ_i with respect to x_j is denoted as $\Phi_{i,j}$ in the above equation.

$$\begin{aligned} G = & E_1 H(\Phi_1) H(\Phi_2) + E_2 H(-\Phi_1) H(\Phi_2) + E_3 H(\Phi_1) H(-\Phi_2) + \\ & E_4 H(-\Phi_1) H(-\Phi_2) + \mu \times \delta(\Phi_1)|\nabla \Phi_1| + \mu \times \delta(\Phi_2)|\nabla \Phi_2| \end{aligned}$$

The energy functional $\mathbb{F}$ is minimized using a gradient descent method with t as an artificial time parameter:

$$\frac{\partial \Phi_i}{\partial t} = -\nabla_{\Phi_i}\mathbb{F}(\Phi_1, \Phi_2), \qquad i = 1, 2 \tag{3.9}$$

The partial derivatives of energy functional $\mathbb{F}(\Phi_1, \Phi_2)$ with respect to level sets Ω_1, Ω_2 are obtained by writing the Euler-Lagrange equations of (3.8):

$$\begin{aligned}
\nabla_{\Phi_1}\mathbb{F}(\Phi_1, \Phi_2) &= \frac{\partial G}{\partial \Phi_1} - \frac{\partial}{\partial x_1}\left(\frac{\partial G}{\partial \Phi_{1,1}}\right) - \frac{\partial}{\partial x_2}\left(\frac{\partial G}{\partial \Phi_{1,2}}\right) - \frac{\partial}{\partial x_3}\left(\frac{\partial G}{\partial \Phi_{1,3}}\right) \\
&= \delta(\Phi_1) \times [E_1 H(\Phi_2) - E_2 H(\Phi_2) + E_3 H(-\Phi_2) - E_4 H(-\Phi_2)] \\
&\quad - \left(\frac{\partial}{\partial x_1}\left[\frac{\Phi_{1,1}}{\left(\Phi_{1,1}^2 + \Phi_{1,2}^2 + \Phi_{1,3}^2\right)^{1/2}}\right] \right. \\
&\quad + \frac{\partial}{\partial x_2}\left[\frac{\Phi_{1,2}}{\left(\Phi_{1,1}^2 + \Phi_{1,2}^2 + \Phi_{1,3}^2\right)^{1/2}}\right] \\
&\quad \left. + \frac{\partial}{\partial x_3}\left[\frac{\Phi_{1,1}}{\left(\Phi_{1,1}^2 + \Phi_{1,2}^2 + \Phi_{1,3}^2\right)^{1/2}}\right] \right) \times \mu \times \delta(\Phi_1) \\
&= \delta(\Phi_1) \times \left[(E_1 - E_2) \times H(\Phi_2) + (E_3 - E_4) \times H(-\Phi_2) \right. \\
&\quad \left. - \mu \times div\left(\frac{\nabla \Phi_1}{|\nabla \Phi_1|}\right) \right]
\end{aligned}$$

Similarly for the level set function Φ_2,

$$\begin{aligned}
\nabla_{\Phi_2}\mathbb{F}(\Phi_1, \Phi_2) &= \frac{\partial G}{\partial \Phi_2} - \frac{\partial}{\partial x_1}\left(\frac{\partial G}{\partial \Phi_{2,1}}\right) - \frac{\partial}{\partial x_2}\left(\frac{\partial G}{\partial \Phi_{2,2}}\right) - \frac{\partial}{\partial x_3}\left(\frac{\partial G}{\partial \Phi_{2,3}}\right) \\
&= \delta(\Phi_2) \times \left[(E_1 - E_3) \times H(\Phi_1) + (E_2 - E_4) \times H(-\Phi_1) \right. \\
&\quad \left. - \mu \times div\left(\frac{\nabla \Phi_2}{|\nabla \Phi_2|}\right) \right]
\end{aligned}$$

Using these partial derivatives in (3.9) gives the following update equations of level set functions $\Phi_1(x)$, $\Phi_2(x)$ in the steepest gradient descent direction:

$$\frac{\partial \Phi_1}{\partial t} = \delta(\Phi_1)\Big[\mu \times div\Big(\frac{\nabla \Phi_1}{|\nabla \Phi_1|}\Big) + (E_2 - E_1) \times H(\Phi_2) + (E_4 - E_3) \times H(-\Phi_2)\Big] \tag{3.10}$$

$$\frac{\partial \Phi_2}{\partial t} = \delta(\Phi_2)\Big[\mu \times div\Big(\frac{\nabla \Phi_2}{|\nabla \Phi_2|}\Big) + (E_3 - E_1) \times H(\Phi_1) + (E_4 - E_2) \times H(-\Phi_1)\Big] \tag{3.11}$$

where ∇ and div are the gradient and divergent operators, respectively.

To summarize, the energy function (3.3) is minimized iteratively. At every iteration, the arbitrary intensity density functions of tissue classes are modeled using KDE and tissue priors are updated by the MRF spatial contextual prior calculated on the previous iteration's segmentation. The flowchart in figure 3.3 summarizes these main steps of KDT for energy minimization. Additionally, an example of updating intensity density functions of tissue classes is shown in figure 3.4 at different iterations to ultimately produce an overlap profile similar to the one observed in figure 3.2a.

3.4.4 Numerical Implementation

For the numerical implementation of level sets, we use $C^{\infty}(\bar{\Omega})$ regularized versions of the Heaviside function and the Dirac delta function, denoted H_ϵ and δ_ϵ, respectively [28]: $H_\epsilon(\Phi) = 1/_2 + 1/_\pi tan^{-1}(\Phi/\epsilon)$, $\delta_\epsilon(\Phi) =$

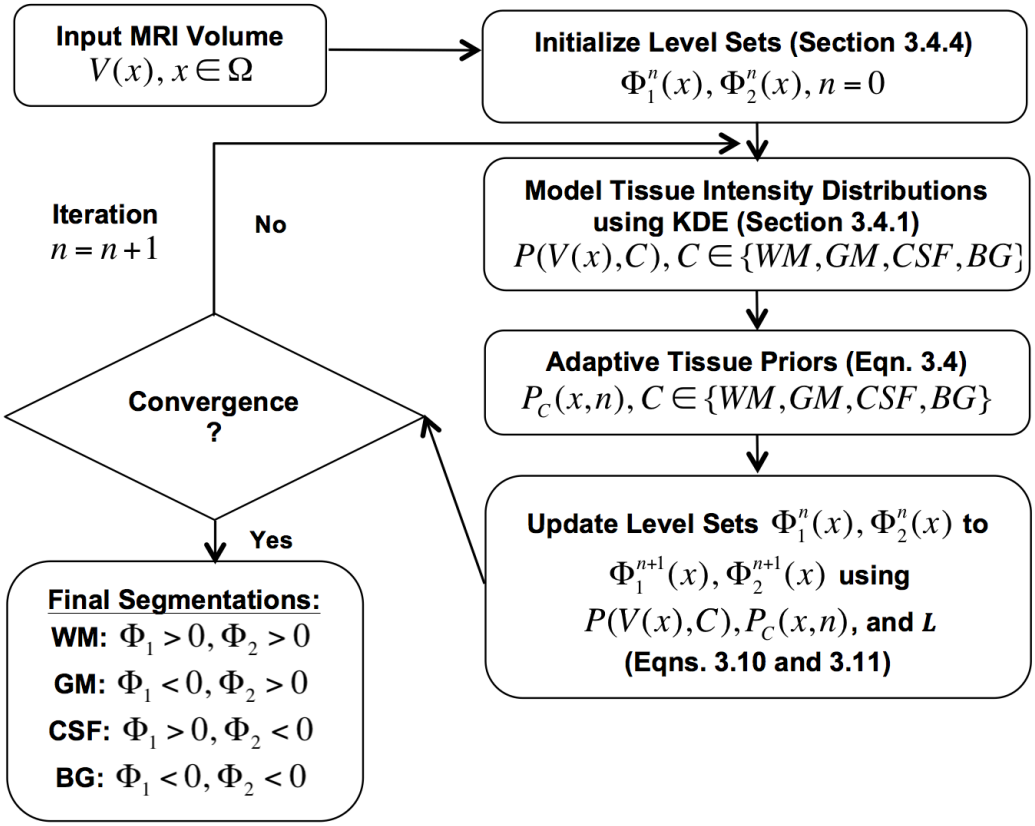


Figure 3.3: KDT segmentation summary: Flowchart summarizing the main steps of KDT tissue segmentation algorithm.

$\partial H_\epsilon(\Phi)/\partial\Phi = \epsilon/\pi(\epsilon^2 + \Phi^2)$. This regularization has the tendency to compute a global minimizer without being affected by the initialization of level sets [28]. An implicit finite difference scheme is used to discretize and linearize the update equations (3.10) and (3.11)[28, 97]. As frequently recommended in previous level set implementations [28, 175], the space step in the finite difference scheme is chosen as $h = 1$ and $\epsilon = 1$ is used to obtain the regularized functions H_ϵ and δ_ϵ. Similar to previous level set implementations,

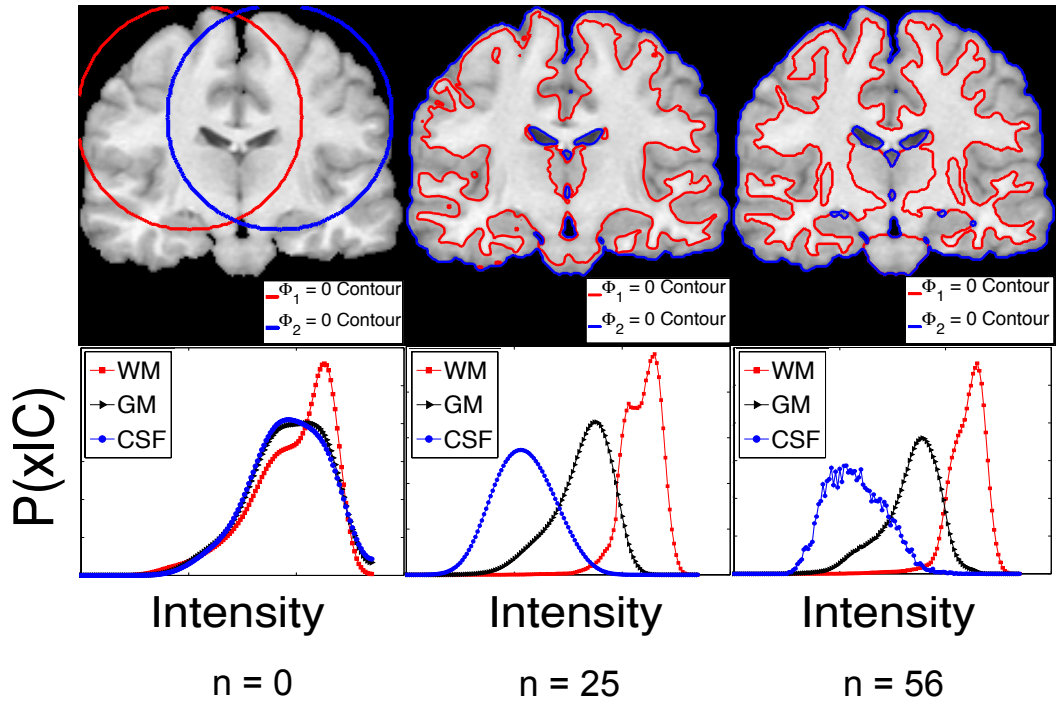


Figure 3.4: KDT segmentation summary: An illustration on the update of class intensity density functions (2^{nd} row) and corresponding tissue segmentations (1^{st} row) at iterations $n = 0$, $n = 25$, and $n = 56$ (convergence). The red and blue outlines show the zero contours of the level set functions Φ_1 and Φ_2, respectively (as described in Section 3.4.3).

the regularization weight μ is set to the standard value of 0.1×255^2 [28, 97]. Since adaptive tissue priors iteratively superimpose MRF contextual prior on the atlas maps, the performance of adaptive priors is expected to be partially dependent on the time step Δt used in equations (3.10) and (3.11). Therefore, besides optimizing MRF weight w, we also consider optimization of the time step Δt to obtain the optimum adaptive prior performance (Section 3.5.4). The initial level set functions $\Phi_1^0(x), \Phi_2^0(x)$ were defined as the signed distance

transforms of two intersecting spherical surfaces randomly selected on the image domain Ω. The diameter of the spherical surfaces was defined to be 1/8th of the smallest dimension in the image domain Ω. As recommended [29], the criteria for convergence of level set evolution is set as $|\Delta C(x)|/\Delta n < \tau$, where $|\Delta C(x)|$ denotes the number of voxels where the class labels change during a span of Δn iterations. In our implementation, we use the threshold $\tau = 1$ and iteration span $\Delta n = 15$.

3.5 Experiments and Results

3.5.1 Data

We consider two real brain MR datasets obtained from the Internet Brain Segmentation Repository (IBSR) to evaluate the segmentation performance of KDT. The first dataset (IBSR-20) contains MR volumes from 20 normal subjects along with expert ground truth tissue segmentations. The data were collected using 1.5 Tesla T1-weighted spoiled gradient echo MR scans on two different imaging systems with a slice thickness of 3.1mm. The second dataset (IBSR-18) contains MR volumes from 18 normal subjects under IBSR V2.0. The data have higher spatial resolution in comparison to IBSR-20 and were collected using 3 Tesla T1-weighted MR scans with a slice thickness of 1.5mm. These datasets are established references for brain segmentation algorithm evaluation because they contain images with varying levels of difficulty (such as low contrast and high intensity inhomogeneity) to comprehensively

evaluate automatic segmentation methods. We only consider real datasets in this study because simulated datasets often implicitly assume a normal distribution of tissue intensities and, therefore, exclude any analysis on intensity overlaps due to arbitrary intensity density functions.

As a preprocessing step, all MR volumes in IBSR-20 and IBSR-18 datasets underwent automatic skull stripping using the brain extraction tool (BET) [160]. The outputs from skull stripping were visually inspected and any skull stripping errors were manually corrected before tissue segmentation using KDT. For adaptive tissue class priors, we used the International Consortium for Brain Mapping atlas maps provided by the Laboratory of Neuroimaging, University of California at Los Angeles [106]. The spatial alignment of atlas maps with subject MR data was performed using the linear registration tool (FLIRT) [85].

3.5.2 Evaluation Metrics

We quantify the segmentation accuracy of KDT by comparing against the expert ground truth segmentations. The Dice similarity coefficient $D(A, B)$ and the Jaccard index $J(A, B)$ are the two most commonly reported metrics in the literature for calculating the overlap between an obtained segmentation and the ground-truth of each class. However, these metrics are inter-related as $J = D/(2-D)$. Therefore, we only use the Jaccard index $J(A, B)$ to assess performance in this study as it is more intuitive for both quantitative evaluation

and comparison purposes. The indices are given by $J(A,B) = |A \cap B|/|A \cup B|$ and $D(A,B) = 2|A \cap B|/(|A| + |B|)$, where A and B are the sets of voxels labeled as tissue class in KDT and the ground truth segmentations, respectively. $|.|$ represents the cardinality of the voxel sets. The results from studies reporting only Dice coefficients were converted to their equivalent Jaccard indices to equitably compare the performances of the methods. We use a second-order Taylor expansion to approximate the mean and standard deviation of $J(A,B)$,

$$\mu_{J(A,B)} \approx \frac{\mu_{D(A,B)}}{2 - \mu_{D(A,B)}} + \frac{2\sigma^2_{D(A,B)}}{(2 - \mu_{D(A,B)})^3}$$

$$\sigma_{J(A,B)} \approx \frac{2\sigma_{D(A,B)}}{(2 - \mu_{D(A,B)})^2}$$

3.5.3 Statistical Comparisons

We statistically compare the segmentation performance of KDT with other competitive methods that reported accuracies on IBSR-20 and IBSR-18 datasets. Due to the paired nature of segmentation accuracies, we perform a two-sided Wilcoxon signed rank test with methods that reported the subject-wise segmentation accuracies. While such a comparison would be ideal, most of the studies did not report the subject-wise accuracies and only reported the summary statistics (mean and standard deviation) of the overlap metric, which excludes any paired statistical comparisons.

3.5.4 Parameter Optimization

The parameters that need to be optimized are the loss matrix L, the adaptive class prior weight w and the time step Δt in the level set implementation. We randomly select a set of 3 MR volumes each from IBSR-18 and IBSR-20 datasets to find the optimum parameter values. We consider a range of possible values for each parameter and select the values that produce the best segmentation performance. The following ranges for the parameters are considered: {0-10} for every $L_{i,j}$ in the loss matrix L (with a step size of 1), {0, 0.05, 0.1, 0.2, 0.4, 0.5, 0.7, 1} for w and {0-1} for time step Δt (with a step size of 0.1). To measure the segmentation performance across all tissue classes simultaneously, we use voxel misclassification rate (VMR) defined as $VMR = \sum_i \sum_{j,j\neq i} |G_i \cap S_j|/|G_i|$ where G_i denotes the set of ground truth voxels for the i^{th} class, S_j denotes the set of voxels classified by KDT as belonging to the j^{th} class, $|.|$ denotes the cardinality of the set, and $i, j \in \{WM, GM, CSF\}$.

The effects of prior weight w and time step Δt values on the segmentation performance are expected to be interdependent while the loss matrix is expected to be independent from the other two variables. The independence assumption is justified because the loss matrix values in theory should be solely determined by the relative extents of intensity overlap between tissue classes. While the inclusion of spatial information using adaptive priors helps improve segmentation accuracy, its omission should not affect the loss matrix values.

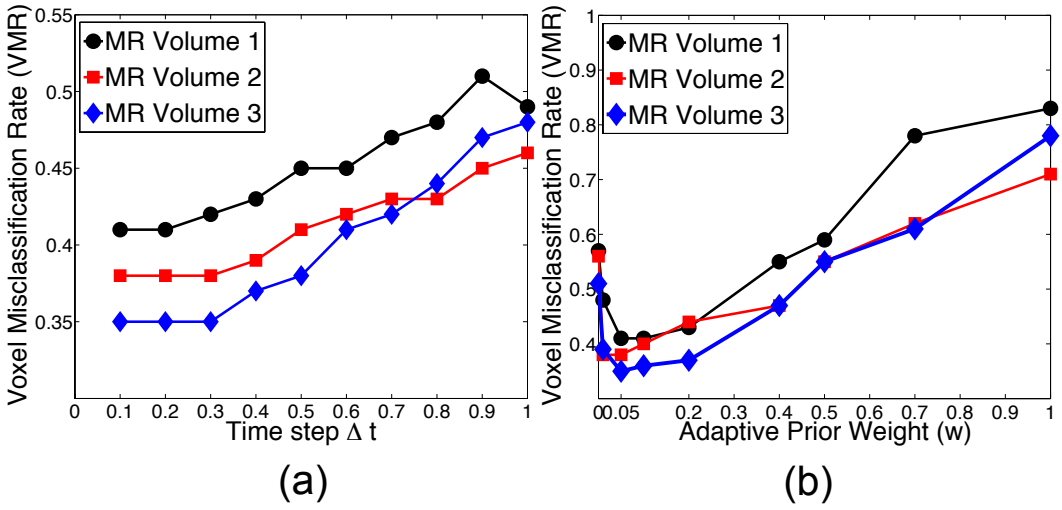


Figure 3.5: Parameter optimization: Plots showing the dynamics of segmentation performance against different values of (a) time step Δt (using $w = 0.05$), and (b) adaptive tissue prior weighting w (using $\Delta t = 0.2$). For clarity, we have only shown the results on 3 out of the 6 MR volumes considered for optimization.

The loss matrix elements are optimized first using a small time step $\Delta t = 0.1$ and $w = 0.2$. The cost of classification into the correct class is considered as zero ($L_{i,i} = 0$) and the cost of misclassification into background (BG) is set to 20 (very high loss since the background has already been removed using BET). While figure 3.2b simply combined the partial overlap areas between any two tissue classes into a single overlap area value, we consider asymmetric loss matrix to investigate any differences in partial intensity overlap areas. For optimizing the loss matrix, one of the elements ($L_{WM,GM}$) is set to 1 and others are estimated relative to this value. We further assume that $L_{WM,CSF}, L_{CSF,WM} > L_{GM,CSF}, L_{CSF,GM} > L_{GM,WM}$ based on the pattern of intensity overlap areas observed in figure 3.2b. We find that the following

asymmetric loss matrix produces the best segmentation performance:

$$L = \begin{array}{c} \\ WM \\ GM \\ CSF \end{array} \begin{array}{c} \begin{array}{ccc} WM & GM & CSF \end{array} \\ \begin{pmatrix} 0 & 1 & 10 \\ 1 & 0 & 6 \\ 10 & 9 & 0 \end{pmatrix} \end{array}$$

While GM and CSF show differences between their partial overlap areas ($GM \rightarrow CSF$ overlap $> CSF \rightarrow GM$ overlap), the overlap distributions corresponding to $WM - GM$ and $WM - CSF$ tissue pairs are symmetric.

The prior weight w and time step Δt are optimized simultaneously by considering all possible combinations. We find that a combination of $w = 0.05$ and $\Delta t = 0.2$ produces the best segmentation performance across all tissue classes. Figure 3.5 shows the dynamics of VMR for different values of Δt and w around the optimum combination of $w = 0.05$ and $\Delta t = 0.2$. To test the validity of independence assumption, a subset of the loss matrix elements are again optimized using $w = 0.05$ and $\Delta t = 0.2$. No changes in the optimum loss matrix elements are observed which confirms that the loss matrix optimization is independent of other parameters in KDT.

3.5.5 Segmentation Performance on IBSR Datasets

3.5.5.1 IBSR-20 Dataset

Table 3.1 compares the overlap metrics between KDT and other methods that have reported segmentation results using IBSR-20 dataset. Some

Table 3.1: Segmentation performance on the IBSR-20 dataset: Table comparing tissue segmentation accuracy (in terms of Jaccard index) of KDT with other existing approaches using MR volumes of the IBSR-20 dataset.

Table 3.1-A: Segmentation into WM, GM, CSF

Method	J^{WM}	J^{GM}	J^{CSF}
KDT	**76.98±3.44**	**83.68±2.58**	**72.94±2.17**
APRS [100]	74.10±2.92	82.60±2.53	70.80±5.65
SPM [9][a]	71.50±3.75	79.80±4.10	70.50±4.32
DMC-EM [188]	69.00±12.00	71.00±8.00	71.00±7.00
Rueda *et al.* [143]	70.10±4.20	70.80±4.50	-
Zheng *et al.* [196]	70.79	65.02	5.10
Dual-Front [98]	67.00	73.90	-
MPM-MAP [104]	68.30	66.20	22.70
Akselrod-Ballin *et al.* [5]	66.85±5.56	75.65±6.16	28.13±9.74
AMS [105]	69.10±4.20	68.30±3.50	-
SVPASEG [170]	68.50	69.80	-
CGMM [67]	66.00±6.00	68.00±4.00	-
MSECM [86][b]	62.80	59.40	21.0

[a] Reported by Lin *et al.* [100]

[b] Mean shift with edge confidence map method by Jimenez-Alaniz *et al.* [86]

Table 3.1-B: Segmentation into WM, $GM + CSF$

Method	J^{WM}	J^{GM+CSF}
KDT	**76.98±3.44**	**85.86±2.12**
Rivera *et al.* [135]	74.20±3.90	81.90±2.80
Ibrahim *et al.* [80]	66.83	77.43

studies combined GM and CSF into a single class $GM + CSF$ and evaluated their segmentation algorithms using overlap metrics of WM and $GM + CSF$ classes. For a fair comparison, we compare KDT with methods belonging to both the categories: segmentation into WM, GM, CSF (Table 3.1-A) and segmentation into WM, $GM + CSF$ (Table 3.1-B). Besides Rivera *et al.*

[135], none of the other studies reported the subject-wise accuracies which excludes any pairwise statistical comparisons. From comparing summary overlap statistics among existing methods, Rivera *et al.* [135] produces better WM segmentation accuracy (74.20 ± 3.90) than all other existing methods [5, 9, 67, 80, 98, 100, 104, 143, 188, 196, 86, 170, 105]. KDT produces statistically significant improvements in WM and $GM + CSF$ segmentation accuracies over Rivera *et al.* [135] ($p = 2.19 \times 10^{-4}$ for WM and $p = 8.9 \times 10^{-5}$ for $GM + CSF$) and, therefore, has better WM accuracy than other competitive methods as well. KDT also produces significantly better GM segmentation accuracy than most of the existing methods [9, 67, 98, 104, 143, 188, 196, 86, 170, 105]. The methods by Akselrod-Ballin *et al.* [5] and Lin *et al.* [100] produce similar GM segmentation accuracy; however, KDT produces significantly better WM and CSF segmentation accuracies. When compared for CSF segmentation, KDT performs significantly better than all existing methods. If the 3 MR volumes selected for parameter optimization are excluded, the WM, GM and CSF segmentation accuracies on the remaining 17 MR volumes are 76.75±3.44, 83.66±2.78, and 72.54±1.14, respectively. The negligible differences in the summary segmentation accuracies after removal of the 3 MR volumes suggest that the parameter values are not biased towards the volumes suggested for parameter optimization.

Table 3.2: Segmentation performance on the IBSR-18 dataset: Table comparing tissue segmentation accuracy (in terms of Jaccard index) of KDT with other existing approaches using MR volumes of the IBSR-18 dataset.

Table 3.2-A: Segmentation into WM, GM, CSF

Method	J^{WM}	J^{GM}	J^{CSF}
KDT	**79.93±2.58**	**88.62±1.32**	**74.55±5.86**
Akselrod *et al.* [4]	76.99	75.44	70.94
DCM-EM [188]	77.00±6.00	73.00±13.00	62.00±11.00
Local-Linear [136]	79.53	84.84	20.77
RCM++[136]	77.62	81.82	17.37
KVPASEG [170]	80.31±2.14	71.92±3.15	-
Awate *et al.* [11]	79.71±2.89	67.91±5.99	-
CGMM [67]	73.71±6.62	65.56±8.18	12.65±6.31
KVL [174][c]	75.04±3.21	65.02±6.79	9.05±3.56

[c] Reported by Greenspan *et al.* [67]

Table 3.2-B: Segmentation into WM, GM+CSF

Method	J^{WM}	J^{GM+CSF}
KDT	**79.93±2.58**	**89.71±1.74**
Rivera *et al.* [135]	78.82±2.83	86.17±2.30
FAST [195][d]	76.77±1.64	86.43±1.89
RiCE [142]	76.28±2.62	88.09±1.36
SURFER-FCM [39][d]	76.40±2.35	87.63±1.34
SPM [9][d]	74.90±4.32	84.08±3.67

[d] Reported by Roy *et al.* [142]

3.5.5.2 IBSR-18 Dataset

Table 3.2 shows the segmentation performance of KDT on the IBSR-18 dataset. Similar to IBSR-20, we report segmentation results for both the cases when brain tissue is segmented into all three tissue types WM, GM, CSF (Table 3.2-A) and when CSF and GM classes are combined into one class $GM + CSF$ (Table 3.2-B). KDT produces better GM and CSF seg-

mentation accuracies than all other methods in Table 3.2-A. Besides three methods (Local-Linear, KVPASEG, and Awate *et al.* [11]), Rivera *et al.* [135] produces better WM segmentation accuracy (78.82 ± 2.83) than all other existing methods [4, 9, 39, 67, 142, 174, 188, 195]. KDT produces statistically significant improvements in the segmentation accuracies of WM and $GM + CSF$ over Rivera *et al.* [135] ($p = 0.02$ for WM and $p = 2 \times 10^{-4}$ for $GM + CSF$) and, therefore, has better WM segmentation performance than the rest of the methods. While the Local-Linear [136], KVPASEG [170] and Awate *et al.* [11] methods produce similar WM segmentation, KDT produces much better average GM and CSF segmentation accuracies. When compared to methods that combined GM and CSF (Table 3.2-B), KDT produces better average $GM + CSF$ segmentation accuracy than all other methods [9, 39, 135, 142, 195]. The summary WM, GM and CSF segmentation accuracies are 79.72±2.78, 88.59±1.23, and 74.39±6.43 when the 3 MR volumes used for parameter optimization are removed, which suggests that there is no significant bias of the selected parameter values on the segmentation results.

3.5.5.3 Performance Comparison between IBSR Datasets

In comparison to IBSR-20, KDT produces better segmentation accuracies for IBSR-18 MR volumes. This is due to the higher resolution of MR volumes in IBSR-18 (less slice thickness and higher magnetic field strength) with less partial volume effects than in IBSR-20. Figure 3.6 shows variations in J^{WM}, J^{GM}, and J^{CSF} across subjects in IBSR-20 and IBSR-18 datasets. Some

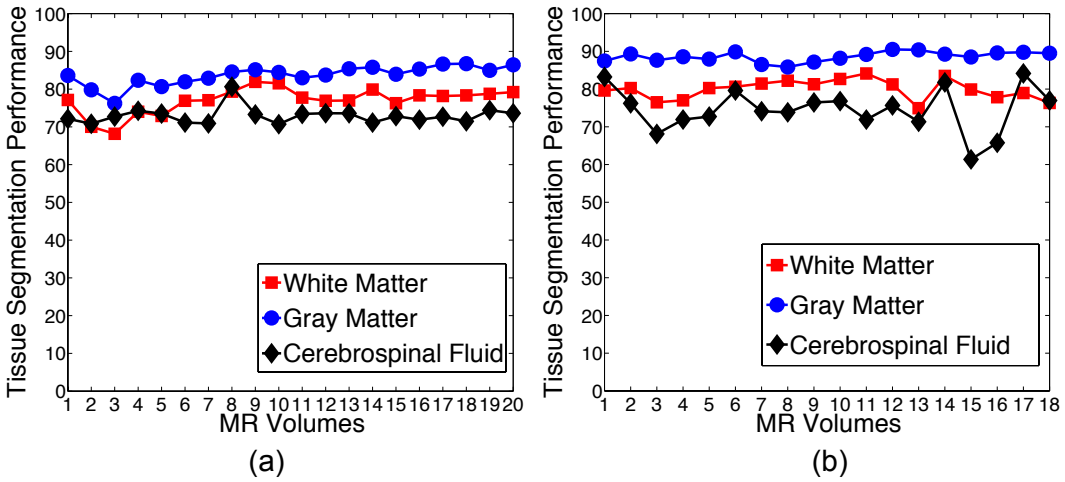


Figure 3.6: Variation in KDT's segmentation performance: Plots showing the variations in J^{WM}, J^{GM}, and J^{CSF} across subjects in (a) IBSR-20, and (b) IBSR-18 datasets.

MR volumes in IBSR-18 dataset (such as subjects 15 and 16) have significantly fewer CSF voxels (smaller ventricles), which results in relatively lower J^{CSF} for those volumes (same number of misclassified voxels produce much higher reduction in Jaccard overlap values). As a result, we observe higher variability of J^{CSF} in IBSR-18 dataset as compared to IBSR-20 dataset. While KDT produces consistent segmentations in both datasets with small variations, the segmentation performance slightly declines in MR volumes that contain high levels of intensity inhomogeneities (such as subjects 2, 3 in IBSR-20 and subjects 11, 13 in IBSR-18). This suggests that while KDT may have better ability in handling intensity overlaps between tissue classes, it is still sensitive to the presence of high levels of intensity inhomogeneities.

3.5.6 Significance of Individual Components

In this section, we evaluate the significance of individual components in KDT by comparing against the most commonly used alternatives. Besides the component being evaluated, all other aspects of KDT are kept exactly the same to ensure that the results truly reflect the significance of that particular component.

3.5.6.1 Significance of Modeling Arbitrarily Shaped Intensity Distributions

We illustrate the significance of modeling arbitrarily shaped intensity density functions by comparing with the case when the most common parametric assumption of normal distribution of intensities is assumed inside each class. Table 3.3 and Table 3.4 show comparisons between the overlap scores, sensitivity, and specificity when modeling arbitrary distributions (1^{st} row) and assuming normal distribution (2^{nd} row). The tissue segmentation accuracies

Table 3.3: Significance of individual components: Table comparing tissue segmentation accuracy (using Jaccard index) of KDT against its variants, where individual components are replaced with their most commonly used alternatives.

Method	J^{WM}	J^{GM}	J^{CSF}
KDT	**79.93±2.58**	**88.62±1.32**	**74.55±5.86**
Normal Dist.	75.11±5.19	82.67±2.64	69.06±7.31
MAP	79.70±4.08	86.88±1.82	59.34±7.74
Atlas	74.48±5.02	83.71±2.43	59.43±8.55
MRF	77.72±4.80	86.82±2.74	72.34±7.15

Table 3.4: Significance of individual components: Table comparing sensitivity (SN) and specificity (SC) in tissue segmentation of KDT against its variants, where individual components are replaced with their most commonly used alternatives.

Table 3.4-A: Sensitivity

Method	SN^{WM}	SN^{GM}	SN^{CSF}
KDT	**89.6±2.3**	**92.0±2.5**	**72.5±5.9**
Normal Dist.	95.0±1.4	83.5±2.8	68.5±5.5
MAP	89.7±2.4	91.6±2.7	62.2±5.5
Atlas	89.6±2.3	86.2±3.4	64.3±4.8
MRF	89.7±3.5	89.9±3.9	77.7±7.3

Table 3.4-B: Specificity

Method	SC^{WM}	SC^{GM}	SC^{CSF}
KDT	**92.2±2.5**	**89.3±2.3**	**99.98±0.0**
Normal Dist.	83.9±2.7	94.2±1.3	99.90±0.1
MAP	91.9±2.6	88.7±2.5	99.84±0.1
Atlas	86.6±3.3	89.1±2.4	99.79±0.2
MRF	90.1±3.9	89.9±3.4	99.84±0.1

are significantly higher when arbitrary distributions inside tissue classes are modeled ($p = 1.53\times10^{-5}$ for WM, $p = 1.53\times10^{-5}$ for GM, and $p < 7.63\times10^{-6}$ for CSF). In normal distribution case, we observe an improvement in SN^{WM}; however, both SC^{WM} and J^{WM} decrease. On the other hand, SN^{GM} and J^{GM} decrease while SC^{GM} improves. This indicates over-classification of voxels as WM and, therefore, an improvement in sensitivity is produced although the overlap score and specificity suffers. Over-classification of voxels as WM is due to the inaccurate estimation of intensity distributions of the tissue classes, which is crucial for the analysis of overlap areas. Figure 3.7 visually illustrates over-classification of voxels as WM using normal distribution (4^{th} column)

when compared to segmentation results from KDT (3^{rd} column) and ground truth segmentations (2^{nd} column).

3.5.6.2 Significance of Incorporating Intensity Overlap Knowledge

We illustrate the significance of incorporating knowledge of the relative extents of intensity overlap between tissue class pairs (loss matrix) by comparing with the case when equal loss values are considered for all tissue misclassifications (equivalent to the MAP model). Table 3.3 compares the performance between using optimum loss matrix (1^{st} row) and the MAP model (3^{rd} row). While similar J^{WM} is observed, J^{GM} and J^{CSF} of MAP are significantly lower than KDT ($p = 3 \times 10^{-3}$ for GM, and $p < 7.63 \times 10^{-6}$ for CSF). MAP penalizes the overlap areas between tissue class pairs with equal costs and results in voxel misclassification between GM and CSF. Figure 3.7 visually illustrates the $GM - CSF$ voxel misclassification in MAP (5^{th} column) in comparison to segmentations produced by using optimum loss matrix (3^{rd} column). This illustrates the importance of incorporating knowledge regarding the relative extents of intensity overlap between tissue class pairs.

3.5.6.3 Significance of Adaptive Class Priors

Adaptive class priors combine atlas maps with MRF contextual information to incorporate spatial information in voxel classification. First, we illustrate the significance of MRF contextual information in class priors by

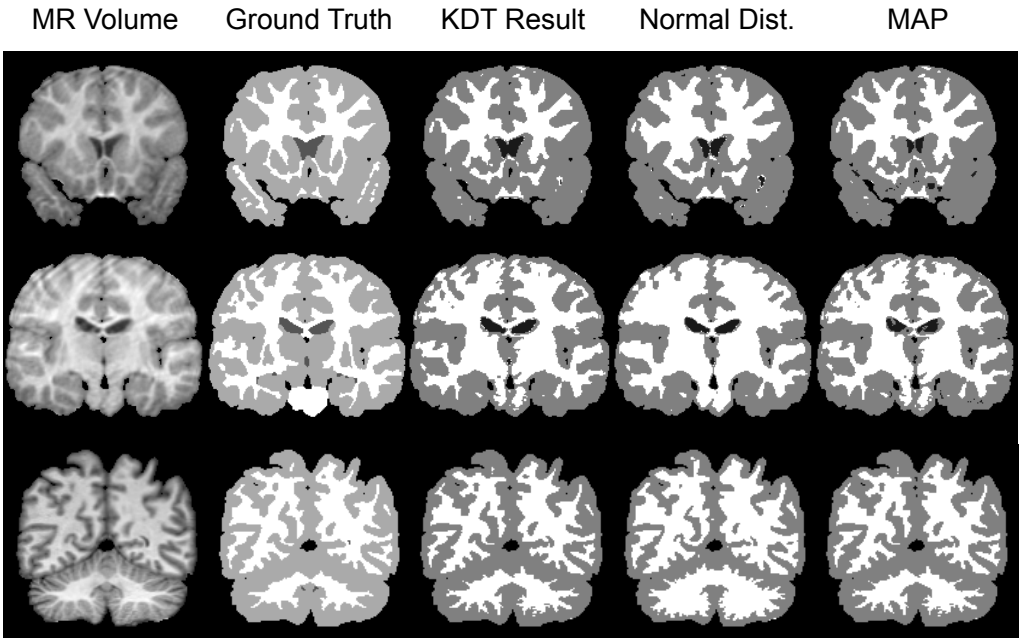


Figure 3.7: Significance of individual components: Visual comparisons between ground truths (2^{nd} column), KDT segmentations (3^{rd} column), segmentations obtained from assuming normal distribution for tissue classes (4^{th} column), and maximum a-posteriori classification (5^{th} column).

comparing with the case when atlas maps are solely utilized as class priors (4^{th} row). The atlas maps are aligned to the MR volumes using the 3D non-rigid demon registration method [121, 168]. The sole use of atlas maps results in significantly lower WM, GM, and CSF segmentation performance ($p < 7.63 \times 10^{-6}$ for WM, GM, and CSF) than using adaptive class priors. This reduction in segmentation performance is due to errors in alignment, which directly translate to segmentation errors. The inclusion of MRF contextual information helps reduce the impact of errors made during the alignment of atlas maps with the MR volumes. In adaptive class priors, the atlas

maps are aligned with the MR volumes using simple linear registration [85]. The application of non-rigid registration methods did not produce any statistically significant differences in the final segmentation performance. This further illustrates the significance of MRF contextual information in the class priors.

In KDT, the methodology for combining MRF contextual priors with atlas maps is slightly different from the traditional way of defining MRF class priors [104, 174]. Traditional methodology combined atlas maps and MRF contextual priors with fixed weightings throughout the segmentation process. As a result, accurate alignment of atlas maps with the MR volumes is essential for obtaining good tissue segmentation performance. Any alignment errors between atlas maps and MR volumes directly translate to errors in tissue segmentation. On the other hand, the methodology used in KDT initializes tissue priors with atlas maps and keeps superimposing MRF contextual priors at every iteration on the tissue priors. As a result, the contribution of atlas maps reduces over the course of the segmentation iterations. Therefore, while tissue atlases still provide important prior anatomical information in the early stages of segmentation, any alignment errors do not result in final segmentation errors. In proposed decision theory framework, we found that the modified methodology is more efficient in incorporating spatial information and produces better segmentation results. We illustrate this by comparing the segmentation performance obtained using adaptive class priors with the traditional MRF class priors (5^{th} row in Tables 3.3 and 3.4). The atlas maps

in the case of traditional MRF class priors were spatially aligned with the MR volumes using 3D non-rigid demon registration method [121, 168]. The use of adaptive tissue class priors produces significantly better WM and GM segmentation performance than the traditional MRF class priors ($p = 1.39 \times 10^{-2}$ for WM and $p = 2.68 \times 10^{-2}$ for GM). These differences in WM and GM segmentation performances are due to errors made during the alignment of atlas maps with the MR volumes, which persist throughout the segmentation. No significant difference in CSF segmentation performance is observed between traditional MRF priors and adaptive class priors. This is because of the significantly higher contrast between CSF and other tissues, which results in good segmentation performance even if the priors are inaccurately defined.

3.5.6.4 Impact of Loss Matrix Elements on Segmentation

In comparison to equal loss values for all tissue pairs (MAP model), we illustrated that the optimum loss matrix produces significantly better segmentation performance across all tissue classes (section 3.5.6.2). Based on the relative loss values assigned to different tissue misclassification types, decisions are taken for voxels that have similar posterior probability of belonging to multiple classes. However, certain applications require higher sensitivity in segmentation of specific tissue types. Here, we take the example of computer-based support systems for diseases such as Multiple Sclerosis, which require high sensitivity in WM segmentation. As shown in figure 3.8, significant improvement in SN^{WM} can be obtained by simply increasing the relative loss

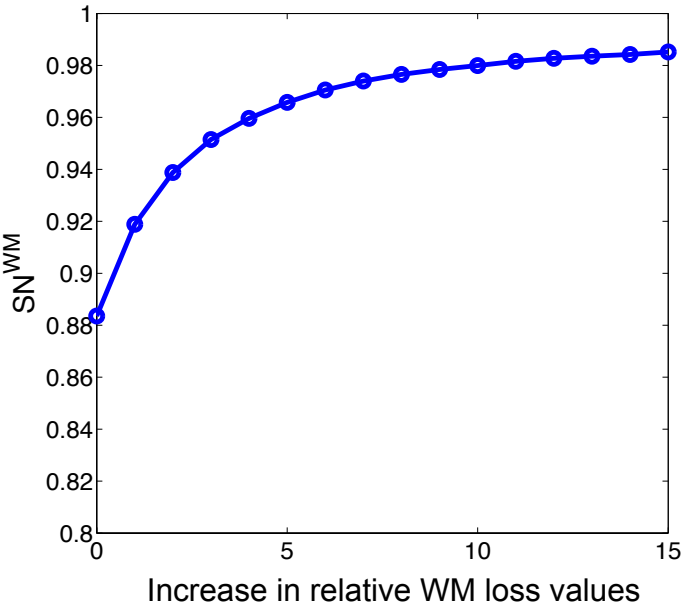


Figure 3.8: Significance of individual components: Plot showing the effect on SN^{WM} when the relative loss values associated with WM misclassification are increased.

values associated with WM misclassification.

3.5.7 Robustness to Initialization of Level Set Functions

The level set framework for energy minimization is robust to the initialization of functions Φ_1, Φ_2. To illustrate this, we evaluated the variation in KDT's segmentation performance across all tissue classes (using VMR) on a MR volume for 20 random initializations (as discussed in Section 3.4.4). We observe the mean and standard deviation of VMR to be 0.3748 and 0.021, respectively. The small variation in misclassification rate ($\sim$0.7% per brain tissue) shows that the level set framework is robust to the initialization of

level set functions Φ_1, Φ_2. Similar results have also been reported by previous level-set based segmentation methods [28, 98, 136].

3.5.8 Computational Complexity

We analyze the computational complexity of KDT and compare it with other segmentation methods. Adaptive kernel density estimation [20] involves use of fast Fourier transform (FFT) for calculation of cosine and inverse cosine transforms on MR voxel intensities. Therefore, modeling class distribution has a complexity of O(Klog$K$) where K denotes the number of MR voxels in the tissue class. In practice, $K < N/2$, where N are the total number of voxels in a MR volume. MRF calculation using the standard belief propagation for a clique size of 2 has a complexity of O(NL^2), where L is the number of class labels (L=4). The level set evolution using equations (7), (8) has a linear complexity O(N). Therefore, the overall complexity of KDT is O(Nlog$N$). The number of iterations required for convergence in our numerical implementation on the IBSR data are typically around $T \sim 50 - 60$. The average physical run time for segmenting a MR volume from IBSR-20 dataset (typical size $256 \times 256 \times 60$) on a Intel Core 2.7Ghz desktop machine was 4.72 minutes. On IBSR-18, the average running time per volume increased to 9.27 minutes due to the higher resolution MR data (typical size $256 \times 256 \times 120$).

Among other segmentation methods that performed complexity analysis, KDT is one of the least computationally intensive. Local-Linear method

[136] has a complexity of $O(nM^2S)$ for 2D segmentation framework, where $M \sim 71 - 91$ is the window size, $n \sim 256 \times 256$ the number of voxels per MR slice and $S \sim 60 - 120$ is the number of slices in a MR volume. This results in physical run times of 30 minutes per MR volume in IBSR-18 using a Intel Core 2 Ghz machine. Ibrahim *et al.* [80] also reported a complexity of $O(N^2GL)$, where G is the number of Gaussian components and L is the sequence length. Rivera *et al.* [135] did not perform complexity analysis but reported physical running times of 3.2 hours per MR volume for 2D framework and 4.1 hours per MR volume for their 3D framework on a 3Ghz machine. We also compared the physical run times of KDT with the state of the art segmentation method FAST [195] included in the FMRIB Software Library (FSL). The physical run times of FAST were 7.11 minutes and 12.38 minutes for IBSR-20 and IBSR-18 volumes respectively, using the same desktop machine used for all experiments in this study. Therefore, both KDT and FAST have comparable physical run times; however, KDT produces significantly better segmentation results (Table 3.2-B).

3.6 Conclusion

MR tissue segmentation is a difficult task due to significant overlaps in the intensity distributions of the tissue classes. Most of the voxel classification errors occur in these regions of intensity overlap where voxels have similar likelihoods of belonging to multiple tissue classes. To address this,

the most common approach has been to correct for image corruptions that reduce the intensity overlap between tissue classes prior to tissue segmentation [132, 157]. In this study, we proposed a new strategy to better deal with intensity overlaps between tissue classes without separately accounting for image corruptions. We illustrated that such a strategy produce more accurate classification of voxels belonging to intensity overlap regions in comparison to existing methods, several of which employed methods for correction of image corruptions.

There are four main technical contributions of this study. First, we demonstrated that the relative extents of intensity overlap between tissue classes are different. The incorporation of this knowledge of the relative intensity overlaps significantly improves the tissue segmentation performance. We illustrated this (Section 3.5.6.2) by comparing the tissue segmentation performance of KDT (with optimal loss matrix) against the segmentation performance obtained using MAP, which is a specific case of KDT when all intensity overlaps are penalized with the same cost. Second, we presented a Bayesian decision theory framework (KDT) to incorporate the knowledge on relative intensity overlaps between tissue classes in tissue segmentation. Decision theory has been traditionally utilized to make decisions on new observations, once the class likelihood distributions are known. Since tissue distributions are unknown prior to segmentation, we utilize the Bayesian decision theory in a different manner. We exploit its ability to draw decision boundaries iteratively such that the final location of decision boundaries produces class distributions

that conform to the overlap profile as observed in figure 3.2b.

Third, we presented a modified approach of adaptive MRF class priors for tissue segmentation. The adaptive MRF priors show better adaptivity than the traditional MRF class priors [104, 174]. Adaptive MRF priors also have lower computational complexity because they do not require the use of time consuming non-rigid image registration methods for aligning patient MR volumes with the atlas maps (Section 3.5.6.3). We illustrated these benefits by comparing the tissue segmentation performances obtained using adaptive MRF priors and traditional MRF class priors, while keeping all other components of the segmentation framework the same. While adaptive class priors show significant improvements in WM and GM segmentation performances, these improvements might be specific only for the proposed decision theory framework. Therefore, further investigation of adaptive class priors incorporated in different segmentation frameworks is required to establish their significance in MR tissue segmentation.

Fourth, we illustrated that the level set approach for energy minimization is highy promising for MR segmentation. While level set-based methods have become popular in computer vision, their application in MR tissue segmentation still remains to be validated owing to lack of evaluation on standardized datasets. We evaluated the performance of KDT on two very popular datasets of real MR volumes, which have been extensively utilized for evaluating tissue segmentation methods. In comparison to methods that employed

other energy minimization techniques (such as expectation maximization and graph cuts), our method using a level-set framework produced significantly better segmentation results. This demonstrates that the level set-based framework is quite promising as a tool for minimizing complicated energy functions.

KDT performs better than most existing segmentation methods for simultaneously segmenting brain MR images into WM, GM, and CSF [4, 9, 39, 67, 80, 104, 135, 142, 174, 195, 196, 170, 11]. Some methods report similar segmentation performance on certain tissue types; however, they fail to perform as well on other brain tissue types [5, 100, 136, 170, 11]. KDT also performs better than the popular segmentation method FAST, which is widely used by the neuroimaging community [195]. Several of these methods involve minimizing image corruptions as part of their segmentation framework. Therefore, KDT illustrates better ability in handling intensity overlaps between tissue classes without the use of any pre-processing method to reduce MR corruptions. Besides improved segmentation performance, KDT also has one of the best computational complexities $\mathrm{O}(N\log N)$ in comparison to other segmentation methods. For applications that require higher sensitivity in segmentation of specific tissue types, KDT also provides a very convenient framework for adapting the segmentation method by simply increasing the relative values of loss matrix elements.

Besides illustrating advantages, KDT also suffers from certain limitations. While KDT better handles intensity overlaps between the tissue classes,

it is still affected by the presence of high levels of intensity inhomogeneities. This can be observed in figure 3.5, where the segmentation performance of KDT declines in MR volumes that suffer from high levels of intensity inhomogeneities. Another associated limitation of KDT is its sensitivity to the presence of partial volume effects in MR volumes. This is the reason behind the lower segmentation performance of KDT on MR volumes of IBSR-20 as compared to MR volumes of IBSR-18. The use of pre-processing steps for reducing the effects of intensity inhomogeneities and partial volume effects can help improve the segmentation performance in MR volumes that contain high levels of MR artifacts. However, inclusion of pre-processing steps will increase the overall computational complexity of tissue segmentation task. Moreover, KDT's segmentation performance will become highly sensitive to the performance of pre-processing steps. The need for skull and background extraction in MR volumes prior to segmentation is another limitation of KDT. The skull and other background structures often present with very similar intensity distributions as the brain tissues, which results in erroneous segmentations using KDT.

This study is limited by its strategy for comparing segmentation accuracy of KDT against the segmentation accuracies of existing segmentation methods. Since IBSR datasets were developed to contain MR volumes with varying level of difficulties, a paired statistical test is ideal for comparing performance between segmentation methods. However, most studies only reported the summary statistics of overlap metrics, which makes it impossible

Table 3.5: IBSR-20 and IBSR-18 subject-wise tissue segmentation accuracy: Table showing subject-wise tissue segmentation accuracies (in terms of Jaccard index) for the MR volumes in the IBSR-20 and the IBSR-18 datasets.

IBSR-20 Volumes	J^{WM}	J^{GM}	J^{CSF}	IBSR-18 Volumes	J^{WM}	J^{GM}	J^{CSF}
5_8	77.12	83.61	72.13	01	79.62	87.44	83.21
4_8	70.02	79.81	70.83	02	80.26	89.30	76.21
2_4	68.17	76.23	72.71	03	76.49	87.64	68.11
6_10	74.01	82.34	74.26	04	77.01	88.56	71.92
15_3	72.89	80.67	73.53	05	80.27	87.94	72.71
16_3	76.92	81.93	71.13	06	80.62	89.86	79.62
17_3	77.05	82.89	70.91	07	81.46	86.51	74.12
8_4	79.34	84.55	80.63	08	82.22	85.85	73.83
7_8	81.92	85.13	73.29	09	81.28	87.16	76.45
110_3	81.54	84.43	70.71	10	82.68	88.17	76.81
111_2	77.74	83.01	73.43	11	84.13	89.18	71.91
112_2	76.91	83.72	73.61	12	81.23	90.50	75.67
100_23	76.98	85.39	73.58	13	74.94	90.39	71.34
202_3	79.89	85.79	71.13	14	83.49	89.25	81.77
191_3	76.27	83.93	72.84	15	79.89	88.51	61.36
12_3	78.35	85.31	71.91	16	77.85	89.59	65.76
13_3	78.18	86.64	72.68	17	78.96	89.76	84.15
1_24	78.34	86.74	71.48	18	76.32	89.50	76.96
205_3	78.76	84.98	74.41				
11_3	79.21	86.44	73.67				

to perform statistical comparisons with existing methods. Noting this limitation, we provide the subject-wise segmentation accuracies for IBSR-20 and IBSR-18 datasets in Table 3.5 to facilitate paired statistical comparisons in future studies.

3.7 Summary

In this chapter, we presented a new knowledge-driven decision theory (KDT) approach for MR tissue segmentation, which embeds prior knowledge on relative extents of intensity overlaps between the tissue classes in the segmentation framework. KDT illustrates good segmentation performance and outperforms other segmentation approaches evaluated on two standardized datasets. In the future, KDT can be incorporated in the established MR analysis pipelines, which are routinely used by the neuroimaging research community. While this chapter presented the contribution of this dissertation in the area of brain MR image analysis, the next several chapters (chapters 4-7) focus specifically on the problem of improving the efficiency of clinical trials of Alzheimer's disease-modifying treatments. In the next chapter, we introduce the currently used outcome measure and discuss limitations associated with its application in clinical trials. Henceforth, in chapter 5, we develop a new methodology for its application, which significantly improves the efficiency of clinical trials focused in the mild-to-moderate Alzheimer's disease stage.

Chapter 4

Clinical Trials of Disease-Modifying Treatments

4.1 Alzheimer's Disease Assessment Scale-Cognitive subscale

The Alzheimer's Disease Assessment Scale's cognitive subscale (ADAS-Cog) is the standard primary cognitive outcome measure for evaluating treatments in clinical trials of mild-to-moderate Alzheimer's disease. In patients, the ADAS-Cog measures impairment across several cognitive domains that are considered to be affected early and characteristically in Alzheimer's disease [140]. However, several concerns have been raised recently regarding its sensitivity in measuring progression of cognitive impairment in clinical trials [27, 131, 75, 74]. The low sensitivity of the ADAS-Cog has been suggested as a possible reason behind the failure of all clinical trials to date of Alzheimer's disease treatments [27, 131, 55, 147].

The low sensitivity of the ADAS-Cog is primarily due to most of its items suffering from either floor or ceiling effects in different stages of Alzheimer's disease [27, 75, 74, 2]. As a result, the ADAS-Cog is limited in measuring pro-

gression of cognitive impairment over the course of disease progression. Noting this limitation, research efforts are underway towards modifying the ADAS-Cog and developing new cognitive assessments with better sensitivity [158, 71]. While the importance of developing better assessments cannot be overstated, their in-depth evaluation and eventual utilization in clinical trials is expected to take a significant amount of time. This opens up a parallel research avenue towards improving the application of the ADAS-Cog in clinical trials, which could help make trials more efficient until a better tool is available.

4.2 Limitations of the Current Scoring Methodology

Another major reason behind the low sensitivity of the ADAS-Cog is its suboptimal scoring methodology, which lacks precision in measuring cognitive impairment. Currently, cognitive impairment is estimated by simply summing scores across the ADAS-Cog items. This methodology suffers from several limitations. Firstly, psychometric analysis of the ADAS-Cog indicates that its items measure impairment in multiple cognitive domains [164, 117, 89]. The current scoring methodology is equivalent to a weighted summation of impairment in the cognitive domains measured by the ADAS-Cog. In studies of treatments that improve only a subset of cognitive domains, such as improvement in memory but not language or praxis, the current methodology obscures the detection of treatment effects [186].

Secondly, the current scoring methodology loses precision in measuring

cognitive impairment by ignoring the pattern of item-wise scores [12]. The difficulty levels of the ADAS-Cog items are not uniform [27, 131, 75] and, therefore, most of the total ADAS-Cog scores can be achieved by different patterns of scores across the ADAS-Cog items [12]. Moreover, since the ADAS-Cog items vary in their ability to measure the underlying cognitive domains [27, 131, 75, 164, 117, 89, 186], an item-level analysis is expected to yield better precision in measuring cognitive impairment. An item-level analysis is also significant for addressing psychometric problems of the ADAS-cog (such as measurement bias), which were not detected at the time of its design [34]. A similar bias concern is raised in clinical trials involving patients undergoing symptomatic therapy using acetylcholinesterase inhibitor (AChEI) drugs, which provide short term improvements on the memory-related ADAS-Cog items [122]. The current scoring methodology does not allow adjustments for such item-level biases, which lead to unaccounted inter-patient variability and further complicates the detection of treatment effects in clinical trials.

Thirdly, the current scoring methodology violates core assumptions of the statistical methods typically employed in clinical trials. The primary efficacy analysis of treatments typically involves linear modeling of serial determinations of the total ADAS-Cog scores of patients using an analysis-of-covariance (ANCOVA) framework [130, 129, 165, 3, 150]. It is reasonable to assume that a patient's true underlying cognitive impairment progresses linearly over short trial durations. However, when cognitive impairment is estimated using the total ADAS-Cog scores, linear modeling using ANCOVA

results in correlated errors due to the categorical nature of the ADAS-Cog items [149, 148]. ANCOVA assumes errors to be independent and normally distributed, which is violated when the total ADAS-Cog scores are used and results in biased efficacy analysis in trials.

Fourthly, the current scoring methodology lacks a proper definition for the measurement scale, which makes comparison and interpretation of cognitive impairment across patients challenging when different variants of the ADAS-Cog are used. In theory, the administration of additional items should only improve measurement precision. However, the current scoring methodology also changes the scale of measurement, with a wider range of scores possible when additional items are administered. The current scoring methodology is also sensitive to missing item responses, scoring errors and variability in the administration of the ADAS-Cog, which are common in clinical trials [151, 33].

In combination, these limitations associated with the current scoring methodology result in low sensitivity of the ADAS-Cog in measuring progression of cognitive impairment in clinical trials. In the next chapter, we develop a new scoring methodology for the ADAS-Cog and investigate the hypothesis that addressing the limitations associated with the current scoring methodology would improve the sensitivity of the ADAS-Cog in clinical trials.

Chapter 5

Improved Scoring Methodology for the ADAS-Cog in Clinical Trials

5.1 Introduction

In this chapter, we present a new scoring methodology for the ADAS-Cog based on psychometric modeling using item response theory (ADAS-CogIRT). Some prior studies have investigated the potential of item response theory for scoring the ADAS-Cog and reported very promising preliminary results [12, 172]. The ADAS-CogIRT methodology is based on extending their prior work, addressing their limitations, and developing a clinically meaningful scale to measure cognitive impairment. We evaluated the sensitivity of the ADAS-CogIRT methodology and compared it with the current scoring methodology for detecting treatment effects in clinical trials using simulation experiments and data from a real negative clinical trial [130]. A preliminary version of this study was presented at the 36^{th} Annual International Conference of the IEEE Engineering in Medicine & Biology Society [180][c] and the

[c]N. Verma, M. K. Markey, "Item response analysis of Alzheimer's disease assessment scale", In *Proc. Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, p. 2476-2479, 2014.

Annual Meeting of the Biomedical Engineering Society in 2014 [181]. This work is currently under review for publication in a peer-reviewed journal. In these works, N. Verma developed the methods, performed the analysis, and prepared the manuscripts. M. Markey helped with the study designs and manuscript revisions.

The chapter is organized as follows. Section 5.2 provides details on data description (section 5.2.1), data preprocessing (section 5.2.2), psychometric analysis of the ADAS-cog (section 5.2.3), the proposed ADAS-CogIRT scoring methodology (section 5.2.4), and application of the ADAS-CogIRT methodology in clinical trials (section 5.2.5). Section 5.3 presents the results of psychometric analysis of the ADAS-Cog (section 5.3.1), evaluation of the ADAS-CogIRT scoring methodology in measuring cognitive impairment (section 5.3.3), and evaluation of the ADAS-CogIRT methodology in clinical trials (section 5.3.4). Finally, section 5.4 discusses the significance of this study, the advantages, and the limitations of the ADAS-CogIRT scoring methodology.

5.2 Materials & Methods

5.2.1 Data

The data for this study were assembled from three public cohorts to ensure that the developed scoring methodology is robust against heterogeneity in patients and study designs. The three cohorts are the Alzheimer's Disease Neuroimaging Initiative (ADNI), the Coalition Against Major Diseases

(CAMD), and the Alzheimer's Disease Cooperative Study (ADCS). A brief description of the ADNI, CAMD, and ADCS cohorts is as follows:

1. **ADNI**: The ADNI was launched in 2003 as a collaboration between several private and public institutions including the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), and the Food and Drug Administration (FDA). The primary goal of ADNI has been to test whether medical imaging, biological markers, clinical and neuropsychological assessments can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). The subjects in ADNI have been recruited from over 50 sites across the U.S. and Canada.

2. **CAMD**: The Critical Path Institute, in collaboration with the Engelberg Center for Health Care Reform at the Brookings Institution, formed the Coalition Against Major Diseases (CAMD) in 2008. The Coalition brings together patient groups, biopharmaceutical companies, and scientists from academia, FDA, the European Medicines Agency (EMA), the National Institute of Neurological Disorders and Stroke (NINDS), and NIA. The data available in the CAMD database were volunteered by CAMD member companies and non-member organizations. CAMD database contains de-identified control arm data on AD patients from 24 clinical trials of disease-modifying treatments.

3. **ADCS**: The ADCS is a major initiative for Alzheimer's disease clinical studies, developed as a cooperative agreement between the NIA and the University of California, San Diego in 1991. The goal of ADCS is to facilitate discovery, development and testing of new treatments for Alzheimer's disease. Since 1991, ADCS has initiated 30 research studies (23 drug studies and 7 instrumental development protocols) over 20 Alzheimer's disease research centers.

We obtained data from 1275 Alzheimer's patients in ADNI, 1828 patients in the placebo arms of 6 clinical trials in CAMD, and 2496 patients in the placebo and treatment arms of 6 clinical trials in ADCS. The data consist of longitudinal ADAS-Cog responses over the duration of trial, basic demographics, Apolipoprotein-E (APOE) genotype, and status of concomitant AChEI therapy of patients. The most common version of the ADAS-Cog, which contains a 'delayed word recall' item in addition to the original 11 items, was used in this study [140, 112]. Table 5.1 summarizes the data from ADNI and the 12 clinical trials of the CAMD and ADCS databases. The data were divided into two subsets. The first subset was used for psychometric analysis of the ADAS-Cog and contained data from ADNI and the placebo arms of all clinical trials except the trial of huperzine A [130]. For psychometric analysis, data from a single visit of every patient was randomly selected to avoid correlated ADAS-Cog responses. The second subset was used to evaluate the scoring methodology we describe in this chapter and contained data from the

Table 5.1: Data description: Summary of patient characteristics from ADNI and clinical trials of CAMD and ADCS databases.

Study	Sample Size	Gender (% Females)	APOE (% $\varepsilon 4$)	ADAS-Cog ($\mu \pm \sigma$)†	Study duration
ADNI	1275	41.7	58.7	14.2±8.5	8 years
CAMD-1105	325	51.0	-	25.2±12.2	20 months
CAMD-1131	57	59.6	-	20.5±3.6	24 weeks
CAMD-1132	412	43.4	38.0	19.1±3.1	51 weeks
CAMD-1140	137	42.3	-	19.1±3.4	24 weeks
CAMD-1141	492	55.3	-	9.9±6.0	23 months
CAMD-1142	405	56.0	64.1	25.3±10.4	18 months
ADCS-HU [130]	210	64.4	65.2	27.1±10.8	24 months
ADCS-DHA [129]	402	52.5	57.7	23.9±9.0	18 months
ADCS-VN [165]	300	63.1	71.3	30.1±9.8	24 months
ADCS-HC [3]	409	53.9	70.0	22.6±8.6	18 months
ADCS-LL [150]	406	59.9	55.3	23.9±10.5	18 months
ADCS-MCI [124]	769	47.0	53.0	11.03±4.2	26 months

† μ: mean score, σ: standard deviation of scores

treatment arms of 11 clinical trials. In addition, the clinical trial of huperzine A, which detected a marginally significant treatment effect [130], was used separately to evaluate the sensitivity of the new scoring methodology in a real clinical trial scenario.

5.2.2 ADAS-Cog Summary & Preprocessing

Out of the twelve items, five items ('Naming objects and fingers', 'Commands', 'Constructional praxis', 'Ideational praxis', and 'Orientation') contain

several subitems such as 'Draw a cube'. Instead of combining the subitem scores, we analyzed these five items at their subitem-level as dichotomous items. The remaining ADAS-Cog items have ordinal responses and were considered as polytomous items for item response theory modeling.

Several items in the ADAS-Cog suffer from severe floor and ceiling effects, which are difficult to model using item response theory. Therefore, as part of the preprocessing step, items with <5% incorrect response rate from mild-to-moderate Alzheimer's patients were either combined with other similar items or removed from the analysis. In the 'Naming objects and fingers' item, all the high frequency objects ('Flower', 'Bed', 'Whistle', and 'Pencil') were combined into a single subitem called the 'High frequency objects'. While the 'Wallet' object is listed as a low frequency object, its incorrect response rate matched with the rates of the medium frequency objects. Therefore, the objects 'Scissors', 'Comb', and 'Wallet' were combined into a single subitem called the 'Medium frequency objects'. The subitem requiring patients to name the finger 'Thumb' was removed from the analysis due to very low incorrect response rate.

In the 'Commands' item, the subitems 'Point to the ceiling, then to the floor' and 'Put the pencil on top of the card, then put it back' were combined into a 'Easy commands' subitem. Similarly, in the 'Constructional praxis' item, the subitems requiring patients to draw a 'Circle' and 'Two overlapping rectangles' were combined into a 'Easy constructional praxis' subitem. The

subitems 'Fold a letter', 'Put letter in envelope', and 'Seal envelope' have low incorrect response rates and, therefore, were combined into a single 'Easy ideational praxis' subitem. The subitem asking patients to recall their 'Full name' has very low incorrect response rate and, therefore, was removed from the analysis. For the ordinal items 'Language', 'Comprehension of spoken language', 'Word finding difficulty', and 'Remembering test instructions', the 'Severe' response category was merged with the 'Moderately severe' response category.

5.2.3 Psychometric Analysis of the ADAS-Cog

Patients' responses to the ADAS-Cog items were probabilistically modeled by defining ADAS-Cog item characteristic functions, which specify relationships between the characteristics of the ADAS-Cog items (slope and intercept) and characteristics of the patients (cognitive impairment). For the sake of understanding the underlying motivation behind defining item characteristic functions, lets consider the case of a dichotomous item with possible responses as either a correct response or an incorrect response. The level of cognitive impairment in a patient can be considered as a continuous measure such that as the underlying cognitive impairment progresses, the patient's probability to answer an item incorrectly increases. The rate of increase in probability of an incorrect response with progression in cognitive impairment can be considered as a characteristic (slope) of the item. Another characteristic of the item can be a threshold value of cognitive impairment (or difficulty) such that patients

with more pronounced cognitive impairment have higher chances of answering the item incorrectly than answering it correctly and vice versa. By using these two characteristics of the item, a relationship (item characteristic function) can be defined between a patient's cognitive impairment and probability of an incorrect response to the item.

Mathematically, for a dichotomous ADAS-Cog item j with response categories as $x_{\cdot j} \in \{0, 1\}$, the item characteristic function relating the probability of an incorrect response $x_{ij} = 1$ by patient i with cognitive impairment $\boldsymbol{\theta_i}$ was defined as:

$$P(x_{ij} = 1|\boldsymbol{\theta_i}, \boldsymbol{\alpha_j}, d_j, g_j) = g_j + \frac{(1 - g_j)}{1 + \exp[-(\boldsymbol{\alpha_j^T \theta_i} + d_j)]} \tag{5.1}$$

where, $\boldsymbol{\theta_i} = (\theta_{i1}, \ldots, \theta_{im})$ denotes a vector of impairment in the m cognitive domains that are assessed by the ADAS-Cog, $\boldsymbol{\alpha_j} = (\alpha_{j1}, \ldots, \alpha_{jm})$ are the item slope components associated with impairment in the m cognitive domains, and d_j is the item intercept. The item intercept d_j represents the relative difficulty level of the item j in comparison to rest of the ADAS-Cog items. The lower asymptotes g_j were included to account for really difficult items, which are answered incorrectly even by cognitively normal individuals.

The definition of item characteristic function for the dichotomous ADAS-Cog items in equation (5.1) was extended to the polytomous ADAS-Cog items with $C_j \geq 2$ response categories $x_{\cdot j} \in \{0, \ldots, C_j - 1\}$ by modeling the bound-

aries between the response categories as

$$
\begin{aligned}
P(x_{ij} \geq 0|\boldsymbol{\theta_i}, \boldsymbol{\alpha_j}, \boldsymbol{d_j}) &= 1, \\
P(x_{ij} \geq 1|\boldsymbol{\theta_i}, \boldsymbol{\alpha_j}, \boldsymbol{d_j}) &= \frac{1}{1+\exp[-(\boldsymbol{\alpha_j^T}\boldsymbol{\theta_i} + d_{j1})]}, \\
P(x_{ij} \geq 2|\boldsymbol{\theta_i}, \boldsymbol{\alpha_j}, \boldsymbol{d_j}) &= \frac{1}{1+\exp[-(\boldsymbol{\alpha_j^T}\boldsymbol{\theta_i} + d_{j2})]}, \\
&\vdots \\
P(x_{ij} \geq C_j|\boldsymbol{\theta_i}, \boldsymbol{\alpha_j}, \boldsymbol{d_j}) &= 0
\end{aligned}
$$

where, $\boldsymbol{d_j} = (d_{j1}, \ldots, d_{j(C_j-1)})$ are the intercepts corresponding to the boundaries between the response categories of item j. The item characteristic functions for individual response categories $x_{ij} = k$ of the ADAS-Cog polytomous items were obtained as

$$
\begin{aligned}
P(x_{ij} = k|\boldsymbol{\theta_i}, \boldsymbol{\alpha_j}, \boldsymbol{d_j}) = P(x_{ij} \geq k|\boldsymbol{\theta_i}, \boldsymbol{\alpha_j}, \boldsymbol{d_j}) - \\
P(x_{ij} \geq k+1|\boldsymbol{\theta_i}, \boldsymbol{\alpha_j}, \boldsymbol{d_j})
\end{aligned} \tag{5.2}
$$

IRT assumes a multivariate normal distribution $g(\boldsymbol{\theta})$ over the latent traits $\boldsymbol{\theta_i} = (\theta_{i1}, \ldots, \theta_{im})$ and integrates them out of the likelihood function. Therefore, the marginal likelihood of the observed response data $\boldsymbol{X} = (\boldsymbol{x_1}, \ldots, \boldsymbol{x_N})$ becomes

$$
L(\boldsymbol{X}|\boldsymbol{\Psi}) = \prod_{i=1}^{N} \left[\int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} P(\boldsymbol{x_i}|\boldsymbol{\Psi}, \boldsymbol{\theta}) g(\boldsymbol{\theta}) d\boldsymbol{\theta} \right] \tag{5.3}
$$

where $\boldsymbol{\Psi}$ is the set of all ADAS-Cog item parameters and $\boldsymbol{x_i} = (x_{i1}, \ldots, x_{in})$ represents the ADAS-Cog item responses by i^{th} patient. Metropolis-Hastings

Robbins-Monro (MHRM) algorithm [25] was used for estimating the ADAS-Cog item parameters $\boldsymbol{\Psi} = \{\boldsymbol{\alpha_j}, \boldsymbol{d_j}; j = 1, \ldots, n\}$ as it is more computationally efficient than the traditional expectation maximization algorithm [19] for estimating multidimensional item response theory models.

5.2.3.1 Cognitive Domains Assessed by the ADAS-Cog

The evaluation of the cognitive domains assessed by the ADAS-Cog in Alzheimer's patients is important not only for its associated clinical significance but also for ensuring the validity of IRT analysis. The estimation of parameters $\boldsymbol{\Psi}$ in IRT assumes local item independence, i.e., patients' responses to the ADAS-Cog items are determined only by their underlying cognitive impairment. The use of an inappropriate set of latent traits $\boldsymbol{\theta_i} = (\theta_{i1}, \ldots, \theta_{im})$ violates this key assumption, which severely compromises the validity of inferences and estimates of cognitive impairment from IRT analysis [193]. This was the primary reason behind the use of a single visit data from every patient to avoid correlated item responses in estimation of IRT parameters.

We performed a parallel analysis on pair-wise polychoric correlations between the ADAS-Cog item responses [77, 73] to determine the number of cognitive domains assessed by the ADAS-Cog. However, parallel analysis typically overestimates the number of latent traits and, therefore, the estimate from parallel analysis was only used as an upper limit on the number of latent traits to be considered for a more in-depth evaluation. Exploratory IRT mod-

els were developed for all possible latent trait structures and were compared using the following criteria:

1. *Model fit*: The latent trait structure should have good global and item-level fits to the ADAS-Cog responses. Global fit was assessed using the two standard statistics of root mean squared error of approximation (RMSEA) [31] and Tucker Lewis index (TLI) [171]. The criteria of RMSEA ≤ 0.05 and TLI ≥ 0.95 are required for a good global fit [78]. Item-level fit was assessed using the recommended S-X^2 statistic, which effectively controls type-I error rates for dichotomous and polytomous items [118, 119, 194, 88].

2. *Local item independence*: The local item independence assumption was tested using the recommended G^2 statistic, which has high sensitivity in detecting local item dependence [30].

3. *Clinical relevance*: The individual latent traits should be clinically meaningful constructs.

During the exploratory analysis, no restrictions on item-trait loadings were imposed and latent traits were allowed to be correlated with each other. After determining the most appropriate latent trait structure based on the above criteria, a confirmatory IRT model was estimated and used for subsequent psychometric analysis of the ADAS-Cog. Cross-loading of items on multiple latent traits were allowed if it significantly improved item-level fit and

reduced local item dependence with other items. The global and item-level fits of the confirmatory IRT model were evaluated using the RMSEA, TLI, and S-X^2 statistics.

5.2.3.2 Measurement Invariance of the ADAS-Cog Items

The ADAS-Cog items should show measurement invariance across patients, despite their characteristics. We performed differential item functioning (DIF) [76] analyses to investigate measurement bias in the ADAS-Cog items due to patient-level factors of gender (men/women), education level (less/greater than 13 years), APOE genotype (presence/absence of an $\varepsilon 4$ allele), and status of concomitant AChEI therapy (yes/no). The ADNI, CAMD, and ADCS cohorts contain predominantly non-Hispanic Caucasian patients, which did not allow DIF analysis due to racial and ethnic factors. All patients undergoing any of the AChEI medications (donepezil, rivastigmine, and galantamine) were labeled as positive for concomitant AChEI therapy. For every DIF factor, ADAS-Cog item characteristic functions were estimated separately inside each patient group and parameter estimates $\mathbf{\Psi}$ were compared using the Wald chi-square test with false discovery rate correction [101]. Before comparison, parameter estimates of patients groups were linearly transformed to a common scale by equating the means and variances of item difficulties across all the groups. If parameter estimates of certain ADAS-Cog items were found to be significantly different between patient groups, those items were flagged as potentially suffering from measurement bias. The ADAS-Cog items that did

not show any significant differences in parameter estimates were anchored by constraining their estimates to be equal across the patient groups. After item anchoring, parameters were re-estimated for all the ADAS-Cog items flagged as potentially suffering from measurement bias to validate if significant differences in parameters still exist between the patient groups. For DIF analysis, the sample size was kept similar across patient groups by randomly selecting patients from bigger patient groups.

Longitudinal invariance of item characteristic functions across different disease stages was investigated by comparing item parameters estimated using baseline responses of patients versus using their responses at 24-months visit, when the disease has significantly progressed. We additionally investigated the extent of sample bias and variance in the ADAS-Cog item characteristic functions due to different patient samples considered for estimation. Sample bias was assessed as the goodness-of-fit of item characteristic functions to response data from the treatment arms of ADCS studies, which were not used for parameter estimation. Sample variance of the ADAS-Cog item characteristic functions was estimated by conducting 1000 bootstrap replications of estimation of item parameters $\mathbf{\Psi}$ with sample replacement. The large sample of patients considered in this study with diverse demographic and clinical characteristics provides a good representation of the overall variability in mild-to-moderate Alzheimer's patient population. Therefore, bootstrapping provides a rough estimate on the expected variability in the ADAS-Cog item characteristic functions if different samples of Alzheimer's patients are considered for

IRT model estimation.

5.2.4 Measurement of Cognitive Impairment

5.2.4.1 ADAS-Cog Scoring Methodology based on IRT Modeling (ADAS-CogIRT)

We propose a new ADAS-Cog scoring methodology based on psychometric modeling using IRT (ADAS-CogIRT) for more accurate measurement of cognitive impairment. The ADAS-CogIRT scoring methodology uses the ADAS-Cog item characteristic functions to measure cognitive impairment in patients based on their ADAS-Cog item response patterns. Given a patient's responses to the ADAS-Cog items $\boldsymbol{x_i} = (x_{i1}, \ldots, x_{in})$, cognitive impairment is measured as the values of the latent traits $\boldsymbol{\theta_i}$ that have the maximum likelihood of observing the ADAS-Cog item responses $\boldsymbol{x_i} = (x_{i1}, \ldots, x_{in})$:

$$L(\boldsymbol{x_i}|\boldsymbol{\theta_i}) = \sum_{j=1}^{n} \log(P(x_{ij}|\boldsymbol{\theta_i}, \boldsymbol{\Psi}))$$
$$\hat{\boldsymbol{\theta}}_{\boldsymbol{i}} = \arg\max_{\boldsymbol{\theta}} L(\boldsymbol{\theta_i}) \tag{5.4}$$

where, $L(\boldsymbol{x_i}|\boldsymbol{\theta_i})$ denotes the log-likelihood of observing the ADAS-Cog item responses $\boldsymbol{x_i}$ in a patient with cognitive impairment $\boldsymbol{\theta_i}$. $\boldsymbol{\Psi}$ denotes the parameters of the ADAS-Cog item characteristic functions after adjusting for measurement bias of the ADAS-Cog items due to patient-level factors. After DIF analysis, the updated ADAS-Cog item characteristic functions with

adjustments for patient-level factors were defined as:

$$P(x_{ij} = 1|\boldsymbol{\theta_i}, \boldsymbol{\Psi}) = g_j + \frac{(1 - g_j)}{1 + \exp[-(\boldsymbol{\alpha_j^T \theta_i} + \boldsymbol{W_i^T \tau_j^T \theta_i} + d_j + \boldsymbol{Z_i \delta_j})]} \quad (5.5)$$

where the fixed effects $\boldsymbol{\tau_j}$ and $\boldsymbol{\delta_j}$ denote adjustments in the ADAS-Cog item slopes and intercepts to account for measurement bias due to patient-level factors with $\boldsymbol{W_i}$ and $\boldsymbol{Z_i}$ as the associated design matrices, respectively. In order to define an appropriate measurement scale for cognitive impairment, we considered the criteria that the scores of cognitive impairment in mild-to-moderate Alzheimer's patients should be non-negative and can be rounded off to the nearest integers without loss of precision.

5.2.4.2 Accuracy of the ADAS-CogIRT Scoring Methodology

Since the ground truth cognitive impairment is unknown, the accuracy of the ADAS-CogIRT methodology for measuring cognitive impairment cannot be directly evaluated. Therefore, we indirectly evaluated the ADAS-CogIRT methodology by assessing its accuracy to predict future ADAS-Cog responses of patients based on their responses in a few initial visits. The ADAS-CogIRT methodology in (5.4) was used to separately estimate cognitive impairment in patients at the baseline, 6 months, and 12 months visits using their ADAS-Cog responses. Assuming linear progression, cognitive impairment at the 24-months visit was estimated for every patient by fitting a linear regression line to the estimates of cognitive impairment from the earlier visits. The estimated cognitive impairment at the 24-months visit was used to predict the ADAS-

Cog item responses of patients as:

$$\hat{x_{ij}} = E\left[x_{ij}\right] = \sum_{k=0}^{C_j-1} k \times P(x_{ij} = k|\hat{\boldsymbol{\theta}}_{\boldsymbol{i}}, \boldsymbol{\Psi}) \tag{5.6}$$

where $k = \{0, \ldots, C_j - 1\}$ are the response categories of the ADAS-Cog item j and $\hat{\boldsymbol{\theta}}_{\boldsymbol{i}}$ represents the estimated cognitive impairment in patient i at the 24-months visit. $P(x_{ij} = k|\hat{\boldsymbol{\theta}}_{\boldsymbol{i}}, \boldsymbol{\Psi})$ was calculated for every patient using the ADAS-Cog item characteristic functions in equation (5.2).

The current scoring methodology measures cognitive impairment in patients by adding scores across the ADAS-Cog items. Therefore, using the current scoring methodology, the total ADAS-Cog scores at the 24-months visit can be predicted by simply fitting a linear regression line to the total ADAS-Cog scores of patients from the earlier visits. The prediction accuracy of the ADAS-CogIRT methodology was calculated using the root mean squared error (RMSE_{ADAS}) between the observed total ADAS-Cog scores $\sum_j x_{ij}$ and the predicted total ADAS-Cog scores $\sum_j \hat{x_{ij}}$ at the 24-months visit:

$$\text{RMSE}_{ADAS} = \sqrt{\frac{\sum_i (\sum_j x_{ij} - \sum_j \hat{x_{ij}})^2}{N_T}} \tag{5.7}$$

where N_T represents the number of patients belonging to the treatment arms of the five ADCS clinical trials. The RMSE_{ADAS} of the ADAS-CogIRT methodology was compared to the RMSE_{ADAS} achieved by using the total ADAS-Cog scores as estimates of cognitive impairment in the initial visits.

5.2.4.3 Precision of the ADAS-CogIRT Scoring Methodology

The precision of the ADAS-CogIRT methodology is dependent on the ability of the ADAS-Cog items to measure different levels of cognitive impairment. The precision of the ADAS-CogIRT methodology was evaluated by calculating the item information functions of the ADAS-Cog items [41]:

$$I_j(\theta) = \sum_{k=0}^{C_j-1} \frac{1}{P_{jk}(\theta)} \left(\frac{dP_{jk}(\theta)}{d\theta} \right)^2 \tag{5.8}$$

where $P_{jk}(\theta)$ represents the probability of k^{th} response by a patient with cognitive impairment θ to the j^{th} ADAS-Cog item, as defined in equation (5.5). A high value for the item information at a given level of cognitive impairment implies that the item measures that level of cognitive impairment with high precision and vice-versa. The cumulative information across all the ADAS-Cog items was used to estimate the expected standard error of measurement of different levels of cognitive impairment using the ADAS-CogIRT methodology.

5.2.5 Improving the Sensitivity of the ADAS-Cog

5.2.5.1 Application of the ADAS-CogIRT Methodology in Clinical Trials

We propose a generalized mixed-effects approach for using the ADAS-CogIRT methodology in clinical trials. Besides estimating baseline cognitive impairment, this approach also estimates the rates of progression in cognitive

impairment based on patients' longitudinal ADAS-Cog responses. In longitudinal settings, the ADAS-Cog item characteristic functions are represented as

$$P(x_{ij}^t = 1|\boldsymbol{\theta_i^t}, \boldsymbol{\Psi}) = g_j + \frac{(1-g_j)}{1+\exp[-(\boldsymbol{\alpha_j^T}\boldsymbol{\theta_i^t} + \boldsymbol{W_i^T}\boldsymbol{\tau_j^T}\boldsymbol{\theta_i^t} + d_j + \boldsymbol{Z_i}\boldsymbol{\delta_j})]} \quad (5.9)$$

where x_{ij}^t and $\boldsymbol{\theta_i^t}$ represent the ADAS-Cog item responses and cognitive impairment of patients at time t. We assumed linear progression of cognitive impairment in patients because the duration of clinical trials are typically too short (∼2-3 years) to observe any complex patterns of disease progression.

$$\boldsymbol{\theta_i^t} = \boldsymbol{\theta_i^0} + \boldsymbol{r_i} \times t \quad (5.10)$$

where $\boldsymbol{\theta_i^0}$ and $\boldsymbol{r_i}$ represent baseline cognitive impairment and progression rates in patients. Significant inter-patient variability in baseline cognitive impairment and progression rates is typically observed in clinical trials. While some variability is systematic due to patient-level factors (such as APOE genotype) and treatment effects, random variability across patients is also substantial. Therefore, we modeled baseline cognitive impairment $\boldsymbol{\theta_i^0}$ and progression rates $\boldsymbol{r_i}$ as mixed-effects in the model to ensure validity of the key assumptions of efficacy analysis.

$$\begin{aligned}
&\boldsymbol{\theta_i^0} = \boldsymbol{\mu_\theta} + \boldsymbol{\beta_{Arm}} \times (\mathrm{Arm}_i) + \boldsymbol{\beta_{Patient}} \times (\boldsymbol{P_i}) + \boldsymbol{\varepsilon_{i,\theta}} \\
&\boldsymbol{r_i} = \boldsymbol{\mu_r} + \boldsymbol{\gamma_{Arm}} \times (\mathrm{Arm}_i) + \boldsymbol{\gamma_{Patient}} \times (\boldsymbol{P_i}) + \boldsymbol{\varepsilon_{i,r}} \\
&\begin{pmatrix} \boldsymbol{\varepsilon_{i,\theta}} \\ \boldsymbol{\varepsilon_{i,r}} \end{pmatrix} \sim N\big(\mathbf{0}, \begin{bmatrix} \boldsymbol{\Sigma_{\theta,\theta}} & \boldsymbol{\Sigma_{\theta,r}} \\ \boldsymbol{\Sigma_{\theta,r}} & \boldsymbol{\Sigma_{r,r}} \end{bmatrix}\big)
\end{aligned} \quad (5.11)$$

where, $\boldsymbol{\mu_\theta}$ and $\boldsymbol{\mu_r}$ represent the average levels of baseline cognitive impairment and progression rates across patients in the placebo arm. The trial arm information of patients is included in the form of a categorical covariate Arm_i such that $\text{Arm}_i = \begin{cases} 0 & \text{if placebo arm} \\ 1 & \text{if treatment arm} \end{cases}$. The fixed effects $\boldsymbol{\beta_{Arm}}$ and $\boldsymbol{\gamma_{Arm}}$ measure differences in the average levels of baseline cognitive impairment and progression rates of patients between the placebo and treatment arms. Patient-level covariates $\boldsymbol{P_i}$ are included to model systematic variability in baseline cognitive impairment and progression rates with $\boldsymbol{\beta_{Patient}}$ and $\boldsymbol{\gamma_{Patient}}$ representing the associated fixed effects. Random effects $\boldsymbol{\varepsilon_{i,\theta}}$ and $\boldsymbol{\varepsilon_{i,r}}$ are included to model random variations in baseline cognitive impairment and progression rates across patients. The cognitive impairment and progression rates in Alzheimer's patients are inter-correlated and, therefore, the random effects $\boldsymbol{\varepsilon_{i,\theta}}$ and $\boldsymbol{\varepsilon_{i,r}}$ are allowed to covary. The parameters of the proposed methodology are estimated using maximum likelihood estimation with adaptive Gauss-Hermite quadrature.

We evaluated the sensitivity of the ADAS-CogIRT methodology for detecting treatment effects in clinical trials using simulation experiments and a real clinical trial, which had been reported as negative but which showed some evidence of a treatment effect [130].

5.2.5.2 Sensitivity Analysis using Simulated Clinical Trials

Clinical trials were simulated to mimic the complexity of real-world clinical trials by including unbalanced patient samples, systematic and random inter-patient variability in cognitive impairment and progression rates, and dropout of patients from clinical trials. The parameters for simulating these characteristics were obtained by analyzing the longitudinal ADAS-Cog data of mild-to-moderate Alzheimer's patients (total ADAS-Cog scores of 25±10) in the placebo arms of ADCS and CAMD trials using a generalized mixed-effects model approach similar to (5.11). Besides the patient-level random effects, nested study-level random effects were also included to model variability in disease stages, where these clinical trials were focused. The parameters estimated for simulating clinical trials were average baseline cognitive impairment and progression rates ($\boldsymbol{\mu_\theta}$, $\boldsymbol{\mu_r}$), random inter-patient variability in baseline cognitive impairment and progression rates ($\boldsymbol{\Sigma_{\theta,\theta}}$, $\boldsymbol{\Sigma_{r,r}}$, $\boldsymbol{\Sigma_{\theta,r}}$), and systematic variability in baseline cognitive impairment and progression rates due to patient-factors ($\boldsymbol{\beta_{Patient}}$, $\boldsymbol{\gamma_{Patient}}$). A Cox proportional hazards model was used for modeling hazard of patient dropout with baseline cognitive impairment, progression rates, and patient-level factors as covariates.

The statistical power of the newly proposed (ADAS-CogIRT) and the standard ADAS-Cog methodologies for detecting treatment effects was evaluated through two simulation experiments. In the first experiment, their power was evaluated for different sample sizes of 200, 400, 600, 800, and 1000 pa-

tients considered in clinical trials of fixed 24 months duration. For the second experiment, the sample size was fixed as 400 patients and the statistical power was evaluated for different durations of 12, 24, 36, and 48 months. These fixed values were selected based on the average characteristics of past clinical trials. Both experiments were repeated for four hypothetical treatment effects of Cohen's $d = 0$ (no effect), 0.2 (mild effect), 0.5 (moderate effect), and 0.8 (large effect) simulated in treatment arms of clinical trials [32]. The case of no treatment effect evaluated the type-I error rates of the proposed scoring methodology. The follow-up visits in both experiments were considered to be biannual during the duration of each trial. In both experiments, 500 clinical trials were simulated for every possible combination of treatment effect, sample size, and trial duration.

For simulating a clinical trial, a large sample of 10000 patients was simulated with normally distributed levels of baseline cognitive impairment and progression rates (using $\boldsymbol{\mu_\theta}$, $\boldsymbol{\mu_r}$, $\boldsymbol{\Sigma_{\theta,\theta}}$, $\boldsymbol{\Sigma_{r,r}}$ and $\boldsymbol{\Sigma_{\theta,r}}$) to represent the population of mild-to-moderate Alzheimer's patients. Based on the sample characteristics of previous trials, 58.5% patients were randomly labeled as APOE-ϵ4 positive, 52.8% were randomly labeled as women, and patient ages were simulated as normally distributed with mean of 74.7 years and standard deviation of 8.54 years. The systematic effects of patient-level factors were simulated in the patient sample using the estimated $\boldsymbol{\beta_{Patient}}$ and $\boldsymbol{\gamma_{Patient}}$ fixed effects. For each simulated clinical trial, S patients were selected at random from this population and randomly distributed between the placebo and

treatment arms. A hypothetical treatment effect of effect size d (in terms of Cohen's $d = \gamma_{Arm}/\sqrt{\Sigma_{r,r}}$) was introduced in the progression rates of impairment in randomly chosen subset of cognitive domains of patients belonging to the treatment arm. The dropout of patients from clinical trials was simulated by using the estimated Cox proportional hazards model. The baseline cognitive impairment and progression rates of patients were used to calculate their longitudinal levels of cognitive impairment at each visit until the duration of the trial. The longitudinal ADAS-Cog responses of patients were simulated using the ADAS-Cog item characteristic functions.

The simulated ADAS-Cog responses were analyzed using the proposed ADAS-CogIRT methodology, single latent trait variant of the ADAS-CogIRT methodology, and the currently employed analysis-of-covariance (ANCOVA) methodologies. The statistical significance of the treatment effect in both methodologies was assessed using z-statistic with correction for multiple comparisons. The statistical power was evaluated as the proportion of clinical trials wherein a statistically significant treatment effect on patients' progression rates was detected.

5.2.5.3 Sensitivity Analysis using Huperzine A Clinical Trial

Besides simulations, we additionally evaluated the sensitivity of the ADAS-CogIRT methodology in a real clinical trial study of huperzine A [130]. In the original negative trial, the higher dose level of 400μg had a marginal

effect (p-value = 0.07) on patients' cognitive functioning after 16 weeks [130]. Given this trend from the original ANCOVA analysis, we were interested in determining whether a more sensitive methodology would change the significance of the treatment effect on progression rates of cognitive impairment. Therefore, we re-analyzed the data from the placebo and 400μg huperzine A arms using the ADAS-CogIRT methodology. The sample size was 141 patients across the two arms in the 16 weeks long trial. Besides statistical significance, we also calculated the size of treatment effects estimated by the ANCOVA and the ADAS-CogIRT methodologies for comparison of sensitivities.

All data analyses in this study were performed using the R software version 3.0.2 environment for statistical computing. The main R scripts for the implementation and evaluation of the ADAS-CogIRT scoring methodology are available at the following repository: `https://github.com/nishant3115/ADAS-CogIRT-Scoring-Methodology`.

5.3 Results

5.3.1 Psychometric Analysis of the ADAS-Cog

5.3.1.1 Cognitive Domains Assessed by the ADAS-Cog

Parallel analysis estimated the number of latent traits as $m = 7$, where only 5 traits were associated with eigenvalues ≥ 1. This suggested that the parallel analysis overestimated the number of latent traits by even accounting

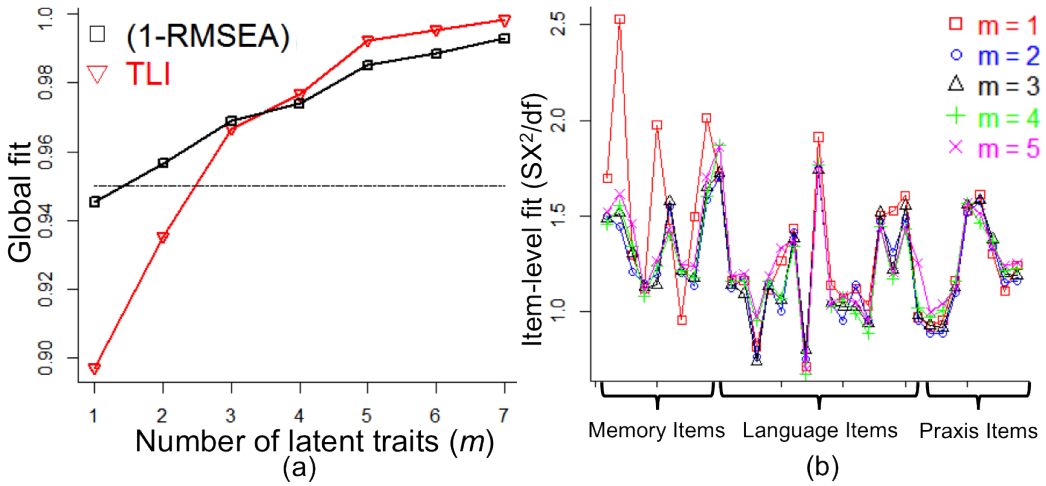


Figure 5.1: Goodness-of-model fit to the ADAS-Cog response data: Figure comparing (a) global fit and (b) item-level fit of the seven latent trait structures to the ADAS-Cog response data. The black dashed line in subfigure (a) represents the typical cut-off of RMSEA = 0.05 and TLI = 0.95 for a good model fit. The item-level fit in subfigure (b) did not improve after $m \geq 3$ latent traits and, therefore, the cases of $m \geq 6$ have not been included for clarity of presentation.

for weak traits measured by small subsets of the ADAS-Cog items. Therefore, the $m = 7$ estimate from the parallel analysis was used only as an upper limit on the number of latent traits to be considered for a more comprehensive psychometric evaluation. Exploratory IRT models were developed with the number of latent traits ranging from $m = 1$ to 7.

By comparing the latent trait structures for up to seven latent traits using the criteria defined in Section 5.2.3.1, the three-dimensional latent trait structure was found to be the most appropriate one. All latent trait structures with the number of latent traits $m \geq 3$ showed good global fit to the ADAS-

Cog response data with RMSEA$\leq$0.05 and TLI$\geq$0.95 (figure 5.1a). While the unidimensional IRT model showed an acceptable value for RMSEA$\sim$0.05, it failed to illustrate an acceptable global fit with TLI$\sim$0.89. This misfit is evident from an item-level assessment of model fit, where the unidimensional structure illustrates poor fit to the response data of all the memory-related ADAS-Cog items (figure 5.1b). While the inclusion of additional latent traits improved the model fit of memory items, the item-level fit did not show any significant improvements after the inclusion of 3 latent traits (figure 5.1b). Local item dependence (LID) between a set of items typically indicates that the item set measures additional latent traits besides the traits already considered in the model. The three-dimensional trait structure had LID only between a few subitems, which belong to the same ADAS-Cog items. Since most ADAS-Cog items have item-specific contexts, such LID is expected. To eliminate all LID, seven traits were required. However, the item parameter estimates of IRT models with three and seven latent traits were very similar, which means that LID in the three-dimensional model is negligible and does not affect item parameter estimates. The three-dimensional trait structure also provides a clinically meaningful interpretation. The pattern of dominant item-trait loadings suggests that the three traits basically represent impairment in the memory, language, and praxis cognitive domains (figure 5.2).

We verified the findings of the exploratory IRT analysis by performing a confirmatory IRT analysis on an independent sample of the ADAS-Cog response data. A three-dimensional confirmatory IRT model was developed

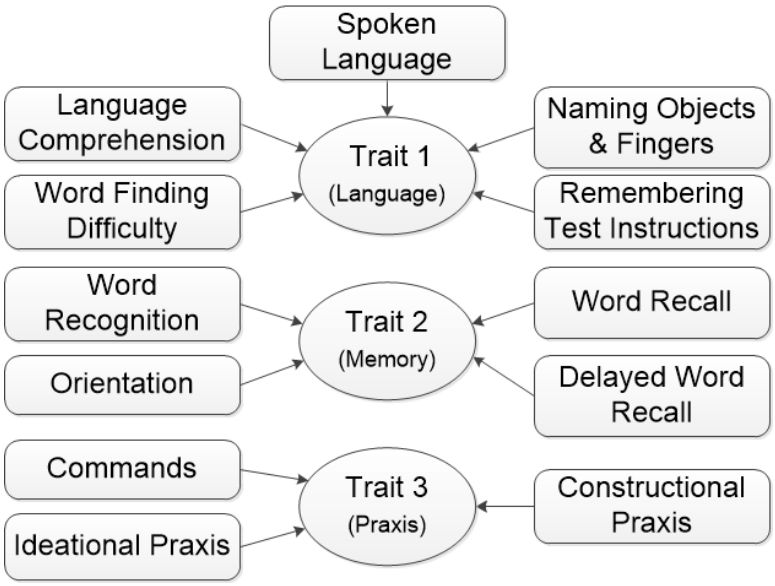


Figure 5.2: Cognitive domains assessed by the ADAS-Cog: Figure showing the item-trait loading structure for the three-dimensional latent trait structure.

using the ADAS-Cog response data from the treatment arms of ADCS clinical trials. The confirmatory IRT model showed good global fit (RMSEA = 0.039 and TLI = 0.95), good item-level fit (S-X^2 insignificant), and low levels of local item dependence between subitems that belong to the same ADAS-Cog items.

5.3.2 Measurement Invariance of the ADAS-Cog Items

Table 5.2 lists the ADAS-Cog items that have measurement bias due to patient-level factors and the directions of those biases. Four ADAS-Cog items have measurement bias due to gender because of different item difficulty for men and women. While naming the object 'rattle' is easier for women, they are less likely to correctly name 'harmonica' and have more difficulty in drawing

Table 5.2: Differential item functioning: Measurement bias of ADAS-Cog items with respect to gender (men/women) and status of concomitant AChEI symptomatic therapy (yes/no)

DIF factor	ADAS-Cog item	Bias type	p-value
Gender	Naming objects & fingers: Rattle	$d_{Men} < d_{Women}$	4.82×10^{-8}
Gender	Naming objects & fingers: Harmonica	$d_{Men} > d_{Women}$	3.76×10^{-5}
Gender	Constructional praxis: Cube	$d_{Men} > d_{Women}$	2.78×10^{-5}
Gender	Remembering test instructions	$d_{Men} > d_{Women}$	3.90×10^{-9}
AChEI	Word recall	$a_{Yes} < a_{No}$	5.40×10^{-10}
AChEI	Word recognition	$a_{Yes} < a_{No}$	1.06×10^{-11}
AChEI	Delayed word recall	$a_{Yes} < a_{No}$	$< 10^{-16}$

* AChEI: acetylcholinesterase inhibitors; d: item intercept/difficulty; a: item slope

a cube. A strong measurement bias due to gender was also observed for the item 'Remembering test instruction', where women are more likely to forget test instructions during administration of the ADAS-Cog. No measurement bias was observed due to education level and APOE genotype. However, the status of AChEI therapy was associated with strong measurement bias for word recall, delayed word recall, and word recognition items. The patients undergoing AChEI therapy had much slower deterioration in their ability to recall and recognize words. However, other memory-related items were not affected by the use of AChEI therapy.

The ADAS-Cog item parameters estimated using baseline data and the 24-months visit data did not show any statistically significant differences, which suggests that the ADAS-Cog item characteristic functions are longitudinally invariant. The item characteristic functions also illustrated little sample bias, with good global (RMSEA = 0.039 and TLI = 0.95) and item-level fit

(S-X^2 was not statistically significant) to response data from the treatment arms of ADCS clinical trials. The item characteristics functions also showed little variance across different patient samples with a tight agreement observed across 1000 bootstrap replicates (figures 5.3, 5.4, and 5.5).

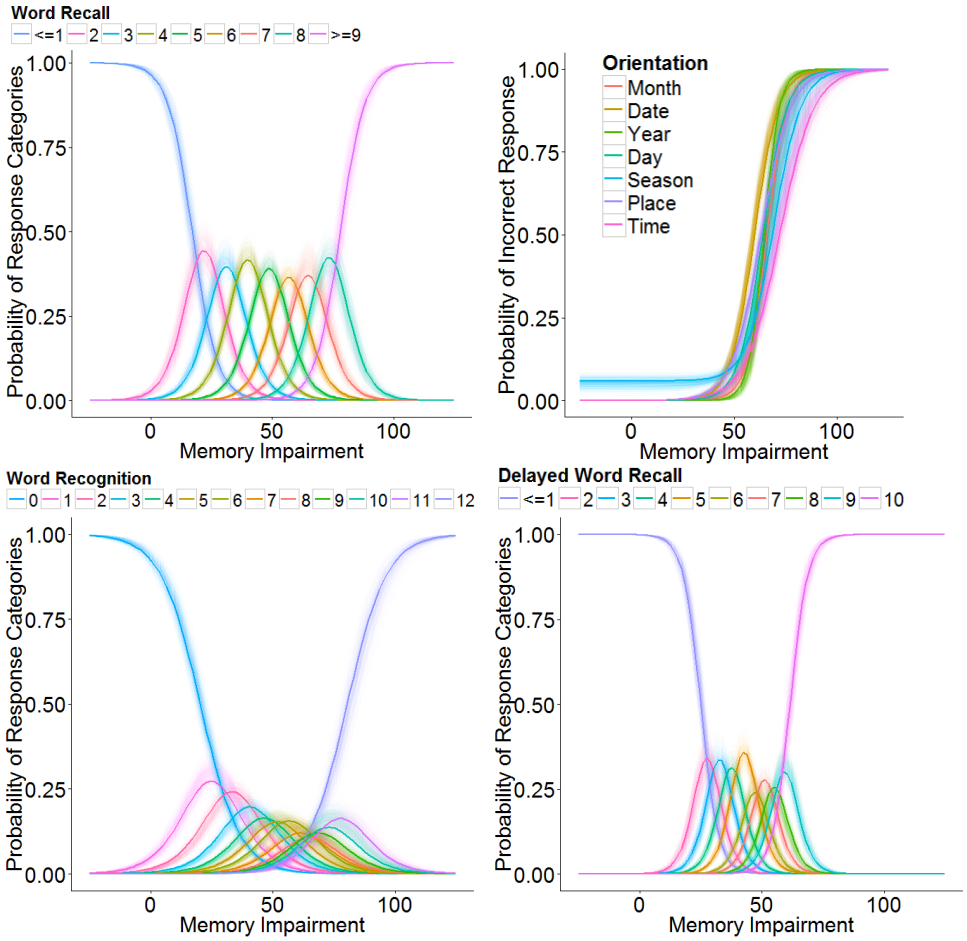


Figure 5.3: Item characteristic functions of memory items: Plots showing item characteristic functions (solid lines) of the ADAS-Cog items that measure memory impairment. The faint lines show variability in the item characteristic functions from 1000 bootstrap replications of parameter estimation with sample replacement.

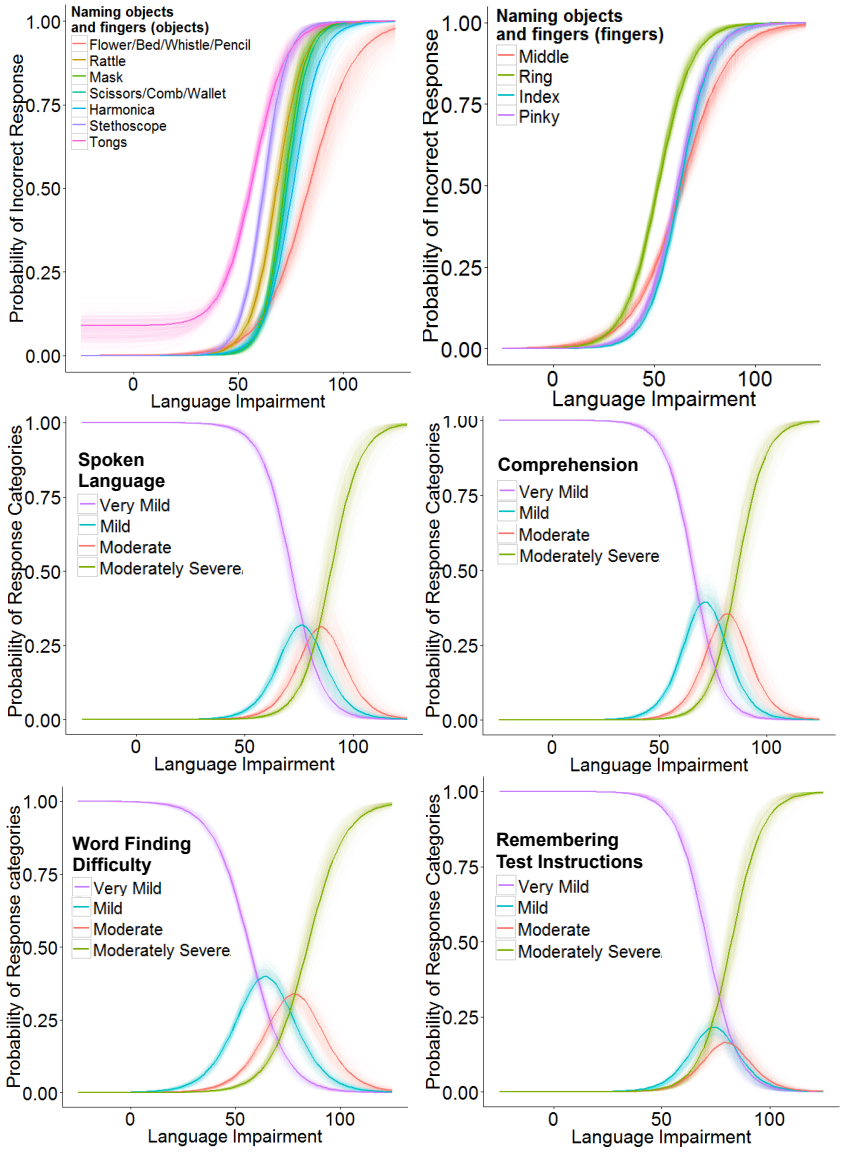


Figure 5.4: Item characteristic functions of language items: Plots showing item characteristic functions (solid lines) of the ADAS-Cog items that measure language impairment. The faint lines show variability in the item characteristic functions from 1000 bootstrap replications of parameter estimation with sample replacement.

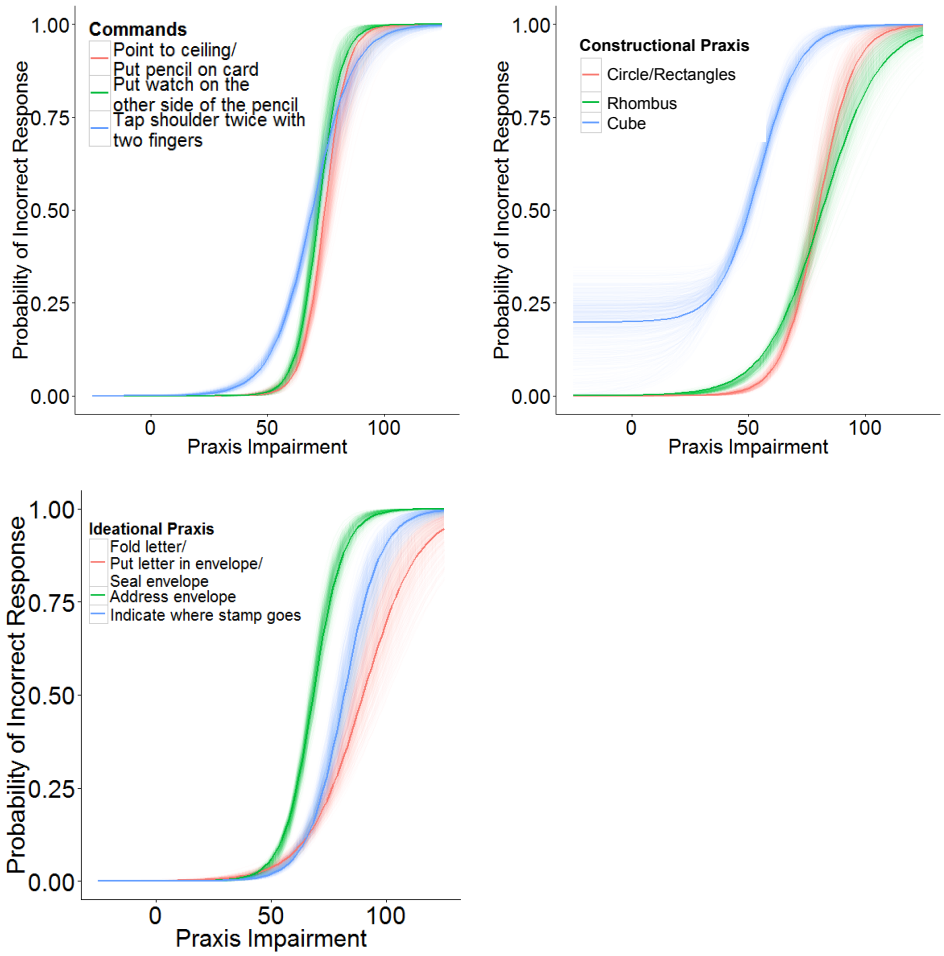


Figure 5.5: Item characteristic functions of praxis items: Plots showing item characteristic functions (solid lines) of the ADAS-Cog items that measure praxis impairment. The faint lines show variability in the item characteristic functions from 1000 bootstrap replications of parameter estimation with sample replacement.

5.3.3 Measurement of Cognitive Impairment

The item parameters were linearly scaled to define measurement scales for memory, language, and praxis impairments such that mild-to-moderate

Alzheimer's patients have non-negative scores in the range of 0 to 100 points and standard errors in estimation of cognitive impairment have magnitudes $\sim$1 point.

5.3.3.1 Accuracy of the ADAS-CogIRT Scoring Methodology

The ADAS-CogIRT methodology illustrated good accuracy in predicting total ADAS-Cog scores at the 24-months visit with $\text{RMSE}_{ADAS} = 1.82$ points. In comparison, the current scoring methodology resulted in an error of $\text{RMSE}_{ADAS} = 6.05$ points, which is similar in magnitude to the annual change of 5-10 points in total ADAS-Cog scores of mild-to-moderate Alzheimer's patients [13, 153]. Figure 5.6 qualitatively compares the predictive accuracies of the ADAS-CogIRT and the current scoring methodologies.

5.3.3.2 Precision of the ADAS-CogIRT Scoring Methodology

While the memory items of the ADAS-Cog are informative over the whole range of memory impairment, language and praxis items hold information only for pronounced levels of language and praxis impairment (figures 5.7a-c). The ADAS-CogIRT methodology shows good precision for almost the whole range of memory impairment. However, due to the inherent limitation of the ADAS-Cog items, the precision of the ADAS-CogIRT methodology in measuring language and praxis impairments is good only when a patients performance is quite poor (figure 5.7d).

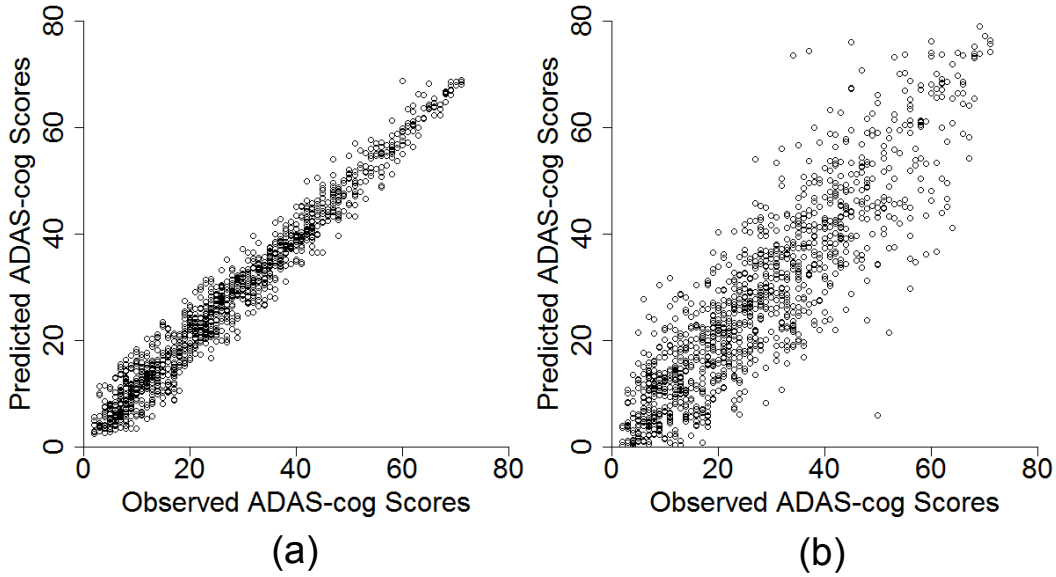


Figure 5.6: Accuracy of the ADAS-CogIRT methodology: Scatterplots showing agreement between the observed total ADAS-Cog scores and the predicted total ADAS-Cog scores at the 24-months visit using (a) the proposed ADAS-CogIRT methodology and (b) the standard scoring methodology.

5.3.4 Improving the Sensitivity of the ADAS-Cog

5.3.4.1 Sensitivity Analysis using Simulated Clinical Trials

The average baseline memory, language, and praxis impairment in mild-to-moderate Alzheimer's patients were estimated as 56.50, 57.83, and 60.27 points. The random inter-patient variability (standard deviation) in baseline memory, language, and praxis impairment were estimated to be 6.31, 7.91, and 8.47 points, respectively. The annual rates of progression in memory, language, and praxis impairment had averages of 2.61, 3.03, and 2.10 points and inter-patient variability of 3.83, 5.56, and 5.04 points, respec-

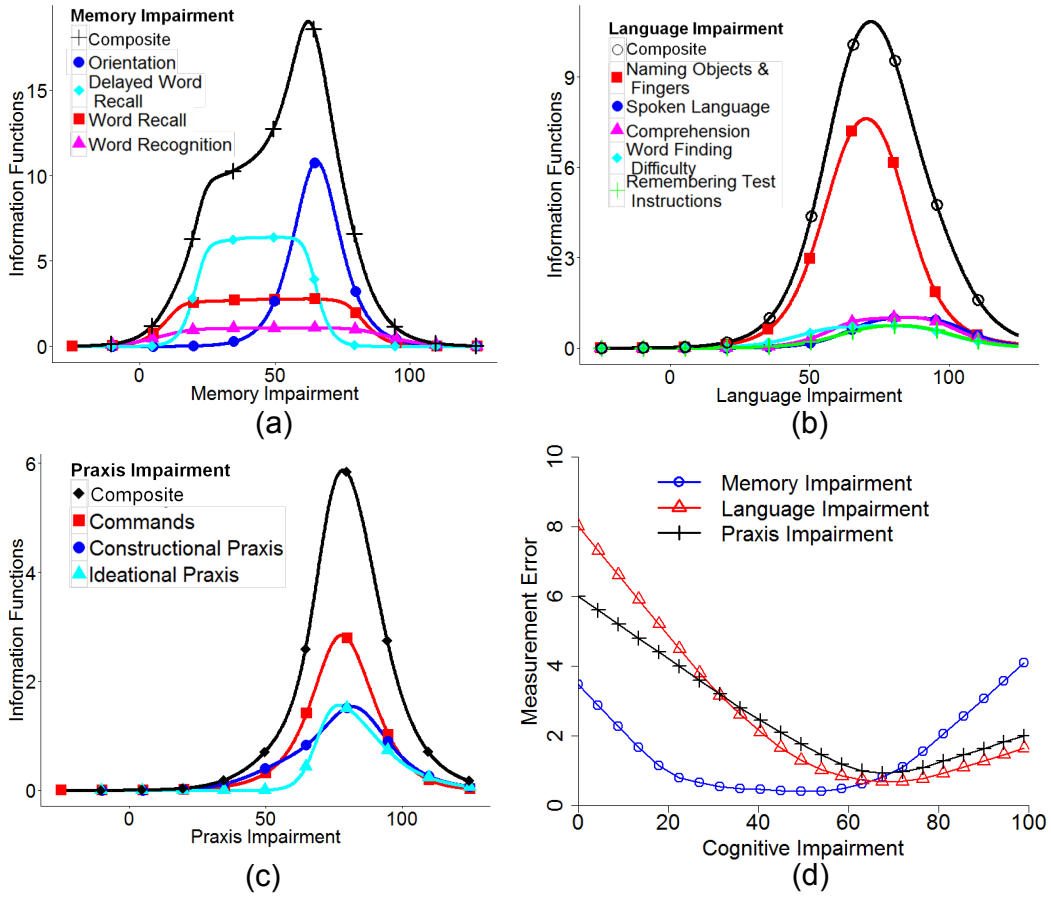


Figure 5.7: Precision of the ADAS-CogIRT methodology: Figure showing item-wise and cumulative Fisher information associated with estimation of (a) memory impairments, (b) language impairments, and (c) praxis impairments. The plot in (d) shows the expected estimation errors associated with different levels of memory, language, and praxis impairments.

tively. Patient age was associated with more pronounced baseline cognitive impairment ($\beta_{Age,Mem} = 0.19$, $\beta_{Age,Lang} = 0.10$); however, the progression rates decreased with patient age ($\gamma_{Age,Mem} = -0.08$, $\gamma_{Age,Lang} = -0.09$, $\gamma_{Age,Prax} = -0.16$). APOE-$\epsilon$4 genotype was associated with higher baseline

memory impairment ($\beta_{APOE,Mem} = 2.83$); however, the progression rates of impairment in all the cognitive domains increased with the presence of an $\epsilon 4$ allele ($\gamma_{APOE,Mem} = 0.97$, $\gamma_{APOE,Lang} = 1.92$, $\gamma_{APOE,Prax} = 1.17$). In the Cox proportional hazards model for patient dropout, the progression rates in various cognitive domains were found to increase the hazard by factors of 2.77 (memory), 1.42 (language), and 2.92 (praxis), while age increased the dropout hazard by a factor of 1.02.

In detecting simulated treatment effects, the ADAS-CogIRT methodology provides significant improvements in statistical power over the currently used ANCOVA methodology (figures 5.8b-d and 5.9b-d). For a mild treatment effect (figures 5.8b and 5.9b), both methodologies have low power and are unable to attain the 80% power cut-off even with large sample sizes and long trial durations. This is due to large inter-patient variability in progression rates within each trial arm, which obscures the presence of a mild treatment effect. However, in comparison to the ANCOVA methodology, the ADAS-CogIRT methodology shows better improvements in statistical power as sample size and trial duration are increased (figures 5.8b and 5.9b). In the case of a moderate treatment effect, the ADAS-CogIRT methodology shows significantly better statistical power than the ANCOVA methodology. The ADAS-CogIRT methodology attains the 80% power threshold in trials with much smaller sample size ($\sim$300 patients) and shorter trial duration ($\sim$18 months) than the ANCOVA methodology, which requires $\sim$1000 patients in a 24 months trial to achieve 80% power (figure 5.8c). With a sample size of 400 patients, the

ANCOVA methodology never achieves 80% power even if the trial duration is increased to over 4 years (figure 5.9c). However, the performance of the ANCOVA methodology improves for a large treatment effect (figures 5.8d and 5.9d). While the ADAS-CogIRT methodology achieves ∼100% power for all sample sizes and trial durations, the ANCOVA methodology also shows good sensitivity reaching 80% power with ∼450 patients in a 24 months trial. The improvement in statistical power of both methodologies with an increase in trial duration was less than that observed with an increase in sample size. Both methodologies have acceptable type-1 error rates of ∼5% for different sample sizes and trial durations (figure 5.8a and 5.9a).

5.3.4.2 Sensitivity Analysis using Huperzine A Clinical Trial

The analysis of the huperzine A trial data using the ADAS-CogIRT methodology revealed that 400μg huperzine A reduces the annual progression rate of praxis impairment by 14.75 points (p-value = 0.0066). The effects of huperzine A on progression rates of memory and language impairment were not statistically significant. The size of the treatment effect detected by the ADAS-CogIRT methodology ($d = 1.97$) was significantly higher than that detected by the ANCOVA methodology ($d = 0.35$). Since praxis items contribute the least to the total ADAS-Cog scores (15/70 points), the ANCOVA methodology detects a much smaller treatment effect in comparison to the ADAS-CogIRT methodology.

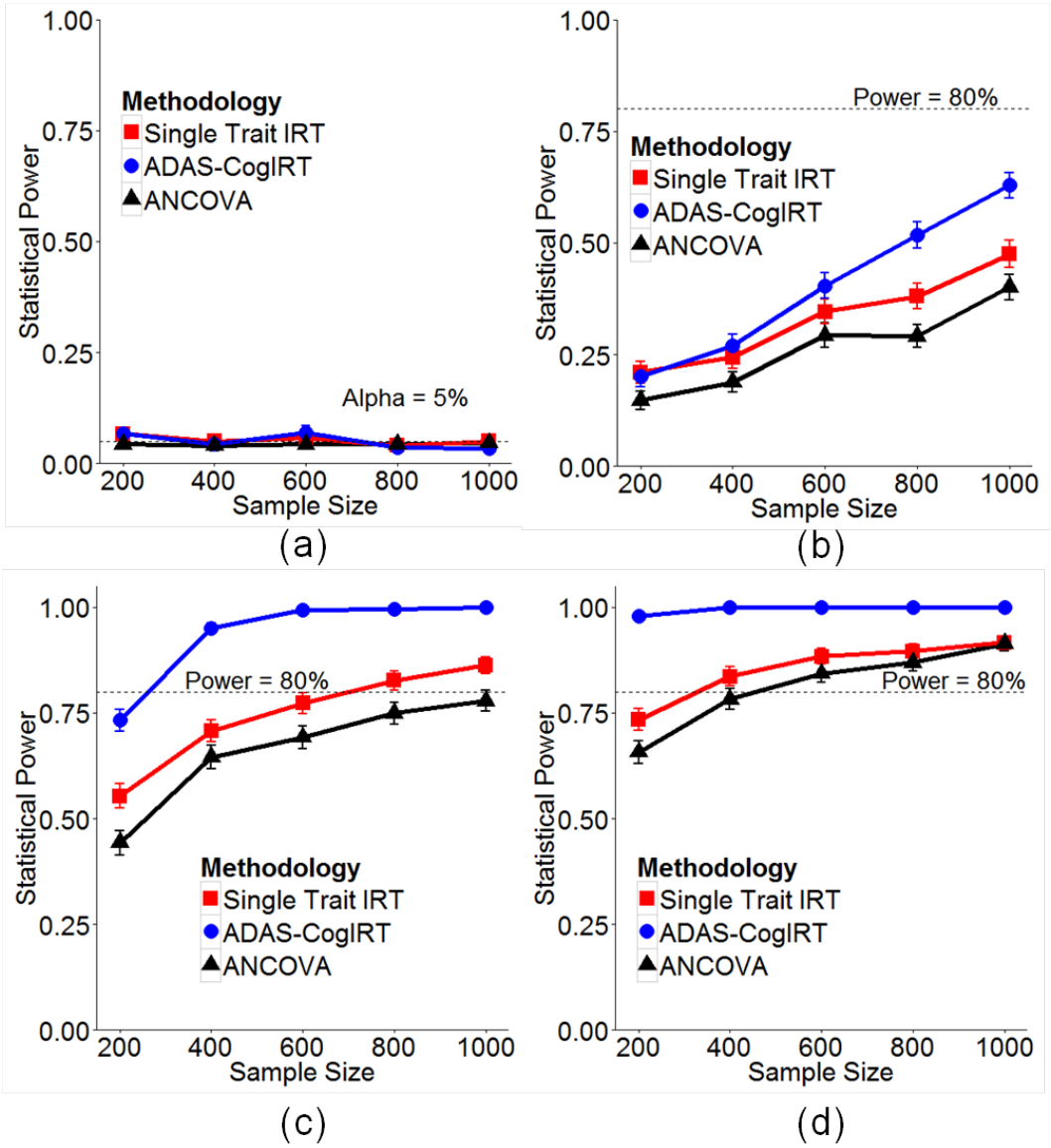


Figure 5.8: Statistical power against sample size: Plots showing the relationship between the statistical power of the ADAS-CogIRT, single latent trait variant of the ADAS-CogIRT and ANCOVA methodologies and sample size for hypothetical treatment levels of (a) $d = 0$, (b) $d = 0.2$, (c) $d = 0.5$, and (d) $d = 0.8$. The trial duration was fixed at 24 months.

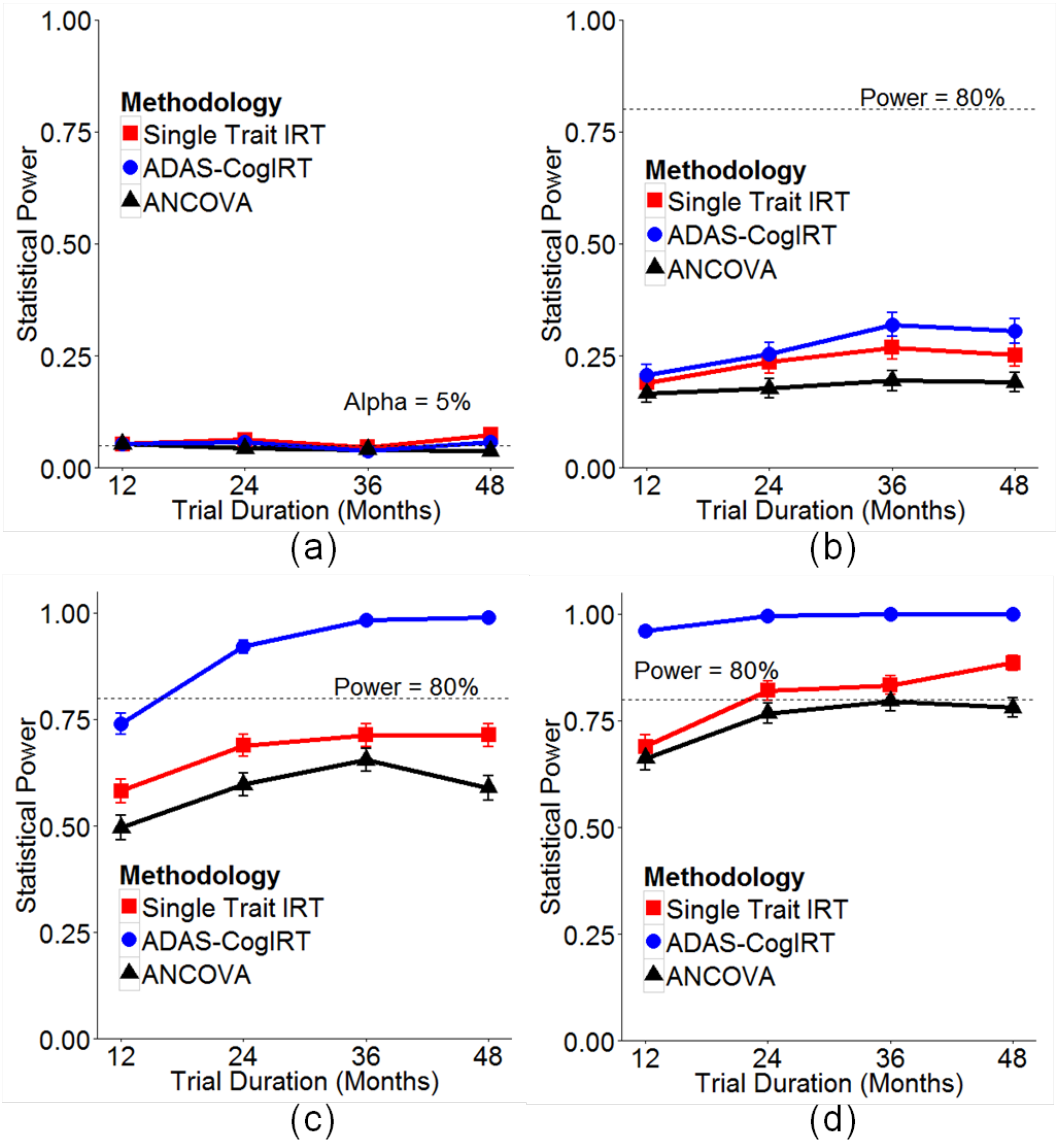


Figure 5.9: Statistical power against trial duration: Plots showing the relationship between the statistical power of the ADAS-CogIRT, single latent trait variant of the ADAS-CogIRT and ANCOVA methodologies and duration of clinical trials for hypothetical treatment levels of (a) $d = 0$, (b) $d = 0.2$, (c) $d = 0.5$, and (d) $d = 0.8$. The sample size was fixed at 400 patients.

5.4 Conclusion

The proposed ADAS-CogIRT scoring methodology addresses several limitations associated with the current scoring methodology. An in-depth psychometric analysis showed that the ADAS-Cog measures impairment in three distinct cognitive domains of memory, language, and praxis in patients. This is in agreement with the design of items in the original ADAS-Cog study [140] and findings of several other factor analysis studies [164, 117, 89, 180]. While memory loss has been long considered characteristic of Alzheimer's disease, its classic neuropathology can also be associated with important language and praxis impairment in patients with predominant posterior perisylvian damage [191]. Similar to AChEI drugs, which specifically target memory mechanisms, and to the effect we detected in the huperzine A trial, investigative treatments in the future may also have non-uniform effects across cognitive domains. The current scoring methodology cannot detect non-uniform effects across cognitive domains. In contrast, the ADAS-CogIRT methodology allows for separate evaluation of treatment effects on the memory, language, and praxis domains.

The ADAS-CogIRT methodology estimates cognitive impairment based on patients response patterns across the ADAS-Cog items. Such an item-level analysis also allows adjustment for measurement bias of the ADAS-Cog items due to gender and status of AChEI therapy. Gender differences in item difficulty are likely due to socio-cultural factors that expose one gender to certain objects and tasks more often than the other gender experiences them. The

status of AChEI therapy strongly affects slopes of word recall and recognition items. Since these items contribute heavily to the total ADAS-Cog scores (32/80 points), this may be the reason behind the slower cognitive deterioration observed in patients undergoing AChEI therapy, as assessed by the current methodology [141, 138].

Inspired by the application of IRT in educational testing, we defined a clinically meaningful scale to measure cognitive impairment. In mild-to-moderate patients, the scale allows estimates to be rounded off to the nearest integers without loss of precision. The scale also facilitates a fractional interpretation of cognitive impairment in study patients, relative to severely impaired patients, who have a cognitive impairment score of 100 points. The parameters of the ADAS-CogIRT methodology are scale independent. Therefore, items can be easily added or removed from the ADAS-CogIRT methodology without having to re-estimate parameters or redefine properties (such as range) of the measurement scale. This is relevant because active research towards improving the ADAS-Cog items is already underway [158]. Since the ADAS-CogIRT methodology pools information across items for estimating cognitive impairment, it is less sensitive to scoring errors in individual items as compared to the current scoring methodology, which is linearly affected. For patients with missing responses to certain items, the ADAS-CogIRT methodology does not require data imputation and estimates cognitive impairment using the set of items answered by the patients. However, measurement precision is lower for patients with missing responses, as would be expected from

psychometric theory.

By addressing limitations of the current scoring methodology, the ADAS-CogIRT methodology measures cognitive impairment more accurately (figure 5.6) and makes clinical trials more efficient by reducing the sample size and follow-up duration required to investigate treatments (figures 5.8 and 5.9). More importantly, it allows for the detection of treatment effects that may have been missed by using the current scoring methodology. This was validated in the huperzine A clinical trial, where the ADAS-CogIRT methodology detected a significant improvement in the praxis domain, that had been overlooked by the traditional ANCOVA methodology. In agreement with our findings, a positive effect of huperzine A on praxis abilities of patients has been found using the activities of daily living scale [185, 99].

Prior work on the application of IRT to the ADAS-Cog mostly focused on evaluating its measurement properties [75, 180, 14]. A few studies additionally investigated IRT for measuring cognitive impairment [12, 172]; however, they assumed that the ADAS-Cog measures a single trait in patients. While a single trait is easy to interpret and model using IRT, it does not adequately fit patient response data (figure 5.1) and severely violates the core IRT assumption of local item independence, which has severe effects on trait estimates [193]. Similar to the total ADAS-Cog scores, the single trait also measures a weighted average of impairment across multiple cognitive domains. Memory items, which have the highest weights, show the poorest fit to the ADAS-Cog

response data (figure 5.1). As a result, measurement of cognitive impairment from a single latent trait IRT model suffers from low precision and reliability. Despite these shortcomings, a single trait IRT model has been demonstrated to significantly improve the sensitivity of the ADAS-Cog in clinical trial simulations [172]. However, those reported results may be overly optimistic because several of the trial characteristics simulated in the analysis [172] are atypical for real clinical trials, such as frequent follow-ups, no patient dropouts, and no heterogeneity due to patient-level factors. Therefore, for a proper comparison, we additionally evaluated the single trait version of the ADAS-CogIRT methodology in more realistic clinical trial simulations and found it to illustrate significantly lower power than the proposed ADAS-CogIRT methodology (figures 5.8 and 5.9). Since prior studies were primarily focused on evaluating the potential of IRT in this application domain, they did not define a measurement scale [12, 172], resulting in counterintuitive negative scores of cognitive impairment in study patients. As also noted by the authors [12, 172], they were additionally limited by ignoring measurement bias and heterogeneity in disease severity of patients.

While this study addressed several limitations of the current scoring methodology, it also suffers from certain limitations. Firstly, we could not investigate measurement invariance of the proposed scoring methodology across all patient-level factors (such as race and ethnicity) due to a lack of heterogeneity in the data. This limitation should be noted in future work, in order to avoid biased estimates of cognitive impairment using the ADAS-CogIRT

methodology with patient groups not included in this study. Secondly, when compared to the current scoring methodology, the ADAS-CogIRT methodology is a bit more cumbersome and requires the use of a computer or a handheld device for measuring cognitive impairment in patients. However, this limitation is less relevant for clinical trials than for routine practice because computing is already required for efficacy analysis of investigative treatments. For routine practice, a specialized application could be developed for making the use of the ADAS-CogIRT methodology straightforward. Thirdly, the precision of the ADAS-CogIRT methodology for measuring language and praxis impairment gets affected due to the inherent limitations of the ADAS-Cog items (figure 5.7). As a result, the improvement in sensitivity afforded by the ADAS-CogIRT methodology will decrease for clinical trials focusing on milder stages of Alzheimer's disease. In those disease stages, it may be better to use this tool only for investigating treatment effects on memory impairment. However, this approach would not be applicable to mild Alzheimer's disease patients who have predominant involvement of the parietal lobe [191]. The inclusion of more difficult items probing subtle levels of language and praxis impairment would improve its measurement precision in milder stages of Alzheimer's disease.

Despite these limitations, the ADAS-CogIRT methodology holds great significance for clinical trials of Alzheimer's treatments. A significant proportion of clinical trials still focus on mild-to-moderate disease stages due to the inability to early detect Alzheimer's disease with high specificity. The

proposed scoring methodology significantly improves the efficiency of clinical trials focused in the mild-to-moderate stages of Alzheimer's disease. Such an improvement in efficiency of clinical trials is highly desirable for rapid testing of future treatments in the critical quest for a disease-modifying treatment. The ADAS-CogIRT methodology also allows separate evaluation of treatment effects in the memory, language, and praxis domains, which can potentially provide additional information on the pharmacological properties of treatments and facilitate development of improved therapies. Future clinical trials of Alzheimer's treatments should consider the proposed ADAS-CogIRT scoring methodology as part of their secondary efficacy analysis to further evaluate and establish the significance of the proposed methodology in comparison to the current scoring methodology.

5.5 Summary

The sensitivity of the Alzheimer's Disease Assessment Scale-Cognitive subscale (ADAS-Cog) in its current form can be significantly improved by addressing limitations associated with its scoring methodology. In this chapter, we described a new scoring methodology for the ADAS-Cog, calling it as the ADAS-CogIRT scoring methodology. The ADAS-CogIRT methodology addresses several major limitations of the current scoring methodology and significantly improves the sensitivity of the ADAS-Cog in clinical trials focused in the mild-to-moderate Alzheimer's disease stage. However, the precision of

the ADAS-CogIRT scoring methodology in measuring language and praxis impairment is poor in the mild cognitive impairment (MCI) stage of Alzheimer's disease, where a significant proportion of clinical trials have started to focus. Cerebral atrophy is closely related with cognitive impairment and, therefore, can potentially be used as a surrogate measure of cognitive impairment in clinical trials. We build upon this concept in the next two chapters (chapters 6-7) with an attempt towards improving the efficiency of clinical trials in the MCI stage. In chapter 6, we review the relationship between cerebral atrophy and cognitive impairment, and the promise of combining them into a biomarker of Alzheimer's disease. In chapter 7, we employ a latent variable modeling framework similar to the one used in this chapter to investigate the relationship between brain-wide cerebral atrophy and cognitive impairment. Based on the relationship, we develop a biomarker and evaluate its performance in clinical trials focused in the MCI stage.

Chapter 6

Cerebral Atrophy and Cognitive Impairment in Alzheimer's Disease

6.1 Introduction

A comprehensive psychometric analysis of the ADAS-Cog in chapter 5 showed that several items in the ADAS-Cog, especially the items that assess language and praxis impairment, suffer from floor effects. While the improved ADAS-CogIRT scoring methodology addresses several limitations associated with the current scoring methodology, it is not able to address the inherent limitations of the ADAS-Cog items. The floor effects of the ADAS-Cog items are even more severe in the mild cognitive impairment (MCI) stage, where an increasing number of clinical trials have started to concentrate. This transition in disease stage for clinical trials is motivated from an understanding that disease-modifying treatments would be more effective in the MCI stage as compared to the mild-to-moderate AD stage [63, 40]. However, the lack of a sensitive outcome measure and an inability to specifically select MCI patients that will convert to AD against other dementia types severely affects the efficacy of clinical trials conducted in the MCI stage. As a result, all clinical trials

of disease-modifying treatments in the MCI stage have failed to show a significant treatment effect, including treatments that show significant effects even in the mild-to-moderate AD stage [146, 102, 48]. Besides cognitive impairment, the regulatory agencies also allow functional impairment as a possible primary end-point; however, AD patients in the MCI stage do not show any deficits in executive functioning [83]. Therefore, there is a critical need for an outcome measure that can track progression of cognitive impairment with good sensitivity in the MCI stage.

While the underlying pathology of AD is believed to be amyloid plaques and neurofibrillary tangles, their deposition is not directly related to cognitive impairment in patients [83, 113, 66]. The current hypothesis of the AD pathological cascade considers amyloidosis, tau pathology, and neuronal injury as sequential rather than simultaneous processes [83, 82]. In fact, cognitive impairment is more closely related to the extent of neuronal and synaptic loss than any other pathological process [42, 167]. Since cerebral atrophy is a manifestation of regional neuronal loss at a macroscopic scale, atrophy is also closely related with cognitive impairment [62, 58]. Advances in medical image analysis have enabled measurement of cerebral atrophy on structural magnetic resonance (MR) images with good accuracy and reliability across scanner manufactures and field strengths [65, 47, 139, 90, 145, 70, 134]. Therefore, if the relationship between cerebral atrophy and cognitive impairment can be accurately established, cerebral atrophy can potentially serve as a surrogate outcome measure or be used in conjunction with the ADAS-Cog for more sensitive

tracking of progression of cognitive impairment in clinical trials.

The promise of cerebral atrophy as a sensitive AD biomarker has led to several small and large-scale investigations. The medial temporal lobe, which is affected characteristically in AD, has been the most widely studied brain region [62, 177]. Previous studies have measured cerebral atrophy on MR volumes using a variety of approaches including visual assessment, measurement based on manual tracing of brain regions, and the use of semi/fully automatic techniques, and voxel-based methods [62, 177]. These prior studies have reported significant improvements in clinical trials using cerebral atrophy as an outcome measure [115, 79]. However, cerebral atrophy is still not approved as a valid biomarker due to a limited understanding of the relationship between cerebral atrophy and clinically relevant outcomes such as cognitive and functional impairment.

In the next chapter, the relationship between brain-wide cerebral atrophy measured on MR volumes and cognitive impairment assessed using the ADAS-Cog is investigated. A biomarker is developed based on the relationship between cerebral atrophy and cognitive impairment, which uses cerebral atrophy as a surrogate marker of cognitive impairment in the MCI stage, where the ADAS-Cog shows low sensitivity in tracking cognitive impairment of patients.

Chapter 7

Biomarker for Tracking Alzheimer's Disease Progression in Clinical Trials

7.1 Introduction

Since both cerebral atrophy and patients' responses to the ADAS-Cog items are related to underlying cognitive impairment, we performed a combined latent variable analysis to investigate this relationship. We extended the latent variable modeling framework to develop a biomarker that combines the ADAS-Cog responses of patients with cerebral atrophy on MR imaging (ADAS-CogMRI) for more accurate measurement of progression of cognitive impairment in clinical trials. The sensitivity of the proposed ADAS-CogMRI biomarker was evaluated and compared with the ADAS-Cog using simulated clinical trials. We additionally evaluated the ADAS-CogMRI and the ADAS-Cog in a real world problem, posed as a clinical trial of a treatment hypothesized to prevent disease progression to dementia. Currently, the primary efficacy analysis of treatments in clinical trials typically involves linear modeling of serial determinations of the total ADAS-Cog scores of patients using an analysis-of-covariance (ANCOVA) framework [146, 102, 48]. In chapter 5,

we developed an improved scoring methodology for the ADAS-Cog using item response theory (ADAS-CogIRT), which addresses several major limitations of the current scoring methodology and significantly improves the sensitivity of the ADAS-Cog in clinical trials. Therefore, we considered the use of both the current and the ADAS-CogIRT scoring methodologies in evaluating the sensitivity of the ADAS-Cog. A manuscript on the work presented in this chapter is currently under preparation for a peer-reviewed journal.

7.2 Materials & Methods

7.2.1 Data

The data used in this study were collected from the Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort (adni.loni.usc.edu). We collected structural magnetic resonance (MR) imaging data, clinical, demographic and ADAS-Cog response data from 437 patients diagnosed with amnestic MCI based on the revised MCI criteria [123] and 122 patients diagnosed with probable Alzheimer's disease at baseline. These patients formed only a subset of the ADNI data that were used in chapter 5, which had structural MR volumes available. The patients underwent follow-up visits roughly every 6 months until an average duration of 1.92 years. During follow-up, 172 MCI patients progressed to meet the clinical criteria for probable Alzheimer's disease while 265 MCI patients stayed stable or reverted back to normal cognitive functioning. For the rest of this chapter, the MCI patients that progressed to Alzheimer's

disease will be referred to as MCI-Converters (MCI-C) and patients that did not progress to Alzheimer's disease as MCI-Nonconverters (MCI-NC). A summary of characteristics of the Alzheimer's disease, MCI-Converters, and MCI-Nonconverters patients is provided in table 7.1.

Table 7.1: Patient summary: Summary of the characteristics of Alzheimer's disease (AD), MCI-Converters (MCI-C), and MCI-Nonconverters (MCI-NC) patients considered in this study.

	AD	MCI-C	MCI-NC
Sample Size	122	172	265
Age	86.9 (7.8)	86.7 (7.4)	82.8 (8.8)
Gender (% Female)	46.3%	44.8%	40.0%
APOE (% $\varepsilon 4$ positive)	67.5%	66.3%	38.6%
Total ADAS-Cog	15.3 (5.1)	11.1 (4.4)	10.8 (3.7)

The data were divided into training and validation sets. The training set comprised of data from randomly selected 61 Alzheimer's disease, 132 MCI-NC, and 86 MCI-C patients. The training set was used for exploratory latent variable analysis and obtaining parameters for design of clinical trial simulations in the MCI stage. Due to the small number of Alzheimer's patients in the training set, data from all the 122 patients was used for designing trial simulations in the mild-to-moderate Alzheimer's disease stage. The validation set contained data from the remaining 61 Alzheimer's patients, 86 MCI-C patients and 133 MCI-NC patients, and was used for confirmatory latent variable analysis and evaluating the performance of the developed biomarker.

7.2.2 Structural MR Analysis

The structural MR volumes of patients were analyzed using the FreeSurfer image analysis suite (http://surfer.nmr.mgh.harvard.edu). The longitudinal stream of the FreeSurfer was used to create unbiased within-patient templates using a robust inverse registration method [133, 134]. All MR volumes underwent processing steps of motion correction and averaging, removal of non-brain tissue, spatial and intensity normalization, segmentation of subcortical white matter and gray matter structures [51, 52], and delineation of boundaries between white matter, gray matter and cerebrospinal fluid [39, 50]. The cerebral cortices were parcellated into 34 gyral structures using deformation procedures [53, 44, 54] and cortical thickness measurements were obtained by calculating the closest distances between the gray matter/white matter and gray matter/cerebrospinal fluid boundaries at each vertex of the cortical surfaces. The accuracy of cortical thickness measurement in FreeSurfer has been validated against histological [139] and manual measurements from MR imaging [90, 145] with good test-retest reliability across scanner manufactures and field strengths [70, 134]. The use of the within-patient templates in the longitudinal stream significantly improves the reliability and statistical power of structural measurements obtained from MR volumes [134].

Cortical thickness measurements have been reported to be better predictors of cognitive impairment [128, 92] . Therefore, mean cortical thickness measurements inside the 34 gyral structures were used in this study. Addition-

ally, volume measurements of the hippocampus and amygdala were included, which have been repeatedly validated as promising biomarkers for tracking Alzheimer's disease progression [177, 154, 152, 84]. For all the brain regions, structural measurements in the two hemispheres were added together. As recommended [189], the volume measurements were normalized with patients' intracranial volumes to account for random inter-patient differences in brain sizes while the cortical thickness measurements were used without any normalization.

7.2.3 Latent Variable Analysis of Atrophy and the ADAS-Cog

Since both cerebral atrophy and responses to the ADAS-Cog items are closely related to underlying cognitive impairment, we conducted a combined latent variable analysis of the structural MR measurements and the ADAS-Cog responses to investigate the relationship between cerebral atrophy and cognitive impairment. The continuous MR measurements were analyzed using latent factor analysis where the k^{th} MR measurement in i^{th} patient was modeled as:

$$y_{ik} = d_{ik} + \boldsymbol{\alpha_k^T \theta_i} + \epsilon_{ik} \tag{7.1}$$

where, $\boldsymbol{\theta_i} = (\theta_{i1}, \ldots, \theta_{im})$ denote m latent traits underlying the MR measurements with associated slopes as $\boldsymbol{\alpha_k} = (\alpha_{k1}, \ldots, \alpha_{km})$, d_{ik} represents patient-specific intercept of k^{th} brain measurement, and ϵ_{ik} is the associated zero-mean residual error term. Factor analysis estimates underlying latent traits by an-

alyzing covariances between the MR measurements y_{ik}. However, besides the systematic component of covariances related to cognitive impairment, the MR measurements also have a random component of covariances due to variability in patients' brain sizes. Therefore, the intercept $d_{ik} = d_k + \delta_{ik}$ was split into a population-level intercept d_k and a patient-level random effect δ_{ik} to account for random covariances between the MR measurements. If the random covariance is not accounted in the model, the estimated latent traits are biased and additionally measure variability in brain sizes, which are not representative of cognitive impairment in Alzheimer's disease patients.

The categorical ADAS-Cog responses were probabilistically modeled using item response theory (IRT) where the probability of an incorrect response $x_{ij} = 1$ by i^{th} patient to a dichotomous item j with response categories $x_{.j} \in \{0, 1\}$ was modeled as

$$P(x_{ij} = 1|\boldsymbol{\theta_i}, \boldsymbol{\alpha_j}, d_j, g_j) = g_j + \frac{(1 - g_j)}{1 + \exp[-(\boldsymbol{\alpha_j^T \theta_i} + d_j)]} \tag{7.2}$$

where, $\boldsymbol{\alpha_j}$ and d_j represent slope and intercept of the j^{th} ADAS-cog item. The lower asymptotes g_j were included to account for difficult ADAS-cog items, which are answered incorrectly even by cognitively normal individuals. The relationship in (7.2) was extended to polytomous items with $C_j \geq 2$ response categories $k = \{0, \ldots, C_j - 1\}$ by modeling boundaries between the response categories.

7.2.3.1 Latent Traits Underlying Atrophy and the ADAS-Cog

The structure of latent traits underlying the MR measurements and the ADAS-Cog responses of patients was investigated using a combination of the following criteria:

- *Traditional techniques*: A combination of traditional techniques of Kaiser's rule (number of eigenvalues ≥ 1) [87], scree plot, and parallel analysis [77, 73] were used for estimating the number of latent traits.
- *Model fit*: The latent trait structure should illustrate good global and item-level fit to the MR measurements and the ADAS-Cog responses. The global-level fit was assessed using root mean squared error of approximation (RMSEA) [31], Tucker Lewis index (TLI) [171], and comparative fit index (CFI) statistics. The criteria of RMSEA$\leq$0.05, TLI$\geq$0.95 and CFI$\geq$0.95 indicates a good global fit. The item-level fit was assessed using the S-X^2 statistic [119, 194] for the ADAS-Cog items and the coefficient of determination (R^2) statistic for the continuous MR measurements. A good item-level fit required S-X^2 statistic to be insignificant for all the ADAS-Cog items and $R^2 \geq 0.70$ for all the MR measurements.
- *Clinical relevance*: The latent traits should be clinically relevant.

The structure of latent traits and the loading parameters were validated by evaluating model fit on the validation set as part of the confirmatory latent variable analysis.

7.2.3.2 Measurement Invariance of the Latent Variable Models

The latent variable models of the MR measurements and the ADAS-Cog items were tested for measurement invariance across disease stages and patient characteristics. Some prior studies have reported nonlinear profiles of cerebral atrophy in brain regions [144, 93]. Therefore, we investigated the significance of higher order polynomial terms in the latent variable models of the structural MR measurements. We also investigated the effects of patient-level factors of gender (men/women), APOE genotype (presence/absence of an $\epsilon 4$ allele), age, and education level (less/greater than 13 years) on the latent variable models of the MR measurements and the ADAS-Cog items. For every patient-level factor, the parameters of the latent variable models of the MR measurements and the ADAS-Cog items were estimated separately for patient groups and compared using the Wald chi-square test with false discovery rate correction [101].

7.2.4 Biomarker for Tracking Alzheimer's Disease Progression

A biomarker based on combined latent variable modeling of the ADAS-Cog responses and cerebral atrophy on MR imaging (ADAS-CogMRI) was developed for more accurate measurement of cognitive impairment and progression rate in patients. The latent variable models of the MR measurements

(7.1) and the ADAS-Cog items (7.2) were extended to longitudinal settings:

$$y_{ik}^t = d_{ik} + \boldsymbol{Z_i}\boldsymbol{\delta_k} + \boldsymbol{\alpha_k^T}\boldsymbol{\theta_i^t} + \boldsymbol{W_i}\boldsymbol{\tau_k^T}\boldsymbol{\theta_i^t} + \epsilon_{ikt} \tag{7.3}$$

$$P(x_{ij}^t = 1|\boldsymbol{\theta_i^t}, \boldsymbol{\alpha_j}, d_j, g_j) = g_j + \frac{(1-g_j)}{1+\exp[-(\boldsymbol{\alpha_j^T}\boldsymbol{\theta_i^t} + \boldsymbol{W_i}\boldsymbol{\tau_j^T}\boldsymbol{\theta_i^t} + d_j + \boldsymbol{Z_i}\boldsymbol{\delta_j})]}$$

where x_{ij}^t, y_{ik}^t, $\boldsymbol{\theta_i^t}$ represent the j^{th} ADAS-Cog item response, the k^{th} MR measurement, and cognitive impairment of i^{th} patient at time t. The fixed effects $\boldsymbol{\tau}$ and $\boldsymbol{\delta}$ denote adjustments in slopes and intercepts of the MR measurements and the ADAS-Cog items to account for measurement bias due to patient-level factors with $\boldsymbol{W_i}$ and $\boldsymbol{Z_i}$ as the associated design matrices. Since follow-up durations of patients are typically too short (∼2-3 years) to observe any complex patterns of progression, a linear progression of cognitive impairment was assumed:

$$\boldsymbol{\theta_i^t} = \boldsymbol{\theta_i^0} + \boldsymbol{r_i} \times t \tag{7.4}$$

where $\boldsymbol{\theta_i^t}$ and $\boldsymbol{\theta_i^0}$ denote cognitive impairment in i^{th} patient at follow-up time t and baseline visit, and $\boldsymbol{r_i}$ denotes the rate of progression of cognitive impairment in i^{th} patient. Given i^{th} patient's longitudinal ADAS-Cog responses $\boldsymbol{x_i^t} = (x_{i1}^t, \ldots, x_{iJ}^t)$ and MR measurements $\boldsymbol{y_i^t} = (y_{i1}^t, \ldots, x_{iK}^t)$ at follow-up times $\boldsymbol{t} = (T_1, \ldots, T_i)$, the proposed ADAS-CogMRI biomarker estimates baseline cognitive impairment $\boldsymbol{\theta_i^0}$ and progression rate $\boldsymbol{r_i}$ of the patient by the

maximum likelihood estimates for observing the data:

$$L(\boldsymbol{X_i}, \boldsymbol{Y_i}|\boldsymbol{\theta_i^0}, \boldsymbol{r_i}) = \sum_{t=T_i}^{T_i} \sum_{j=1}^{J} \log(P(x_{ij}^t|\boldsymbol{\theta_i^0}, \boldsymbol{r_i}, \boldsymbol{\Psi}))+ \sum_{t=T_i}^{T_i} \sum_{k=1}^{K} \log(P(y_{ik}^t|\boldsymbol{\theta_i^0}, \boldsymbol{r_i}, \boldsymbol{\Psi}))$$

$$\{\hat{\boldsymbol{\theta_i^0}}, \hat{\boldsymbol{r_i}}\} = \arg\max_{\boldsymbol{\theta}, \boldsymbol{r}} L(\boldsymbol{X_i}, \boldsymbol{Y_i}|\boldsymbol{\theta_i^0}, \boldsymbol{r_i}) \tag{7.5}$$

where $L(\boldsymbol{X_i}, \boldsymbol{Y_i}|\boldsymbol{\theta_i^0}, \boldsymbol{r_i})$ denotes the log-likelihood of observing the longitudinal ADAS-Cog item responses $\boldsymbol{X_i} = (\boldsymbol{x_i^{T_1}}, \ldots, \boldsymbol{x_i^{T_i}})$ and MR measurements $\boldsymbol{Y_i} = (\boldsymbol{y_i^{T_1}}, \ldots, \boldsymbol{y_i^{T_i}})$ in a patient with baseline cognitive impairment $\boldsymbol{\theta_i^0}$ and progression rate $\boldsymbol{r_i}$. $\boldsymbol{\Psi}$ denotes the set of parameters of the ADAS-Cog item characteristic functions and latent variable models of MR measurements.

7.2.5 Application of the Biomarker in Clinical Trials

Significant inter-patient variability in baseline cognitive impairment and progression rates is typically observed in clinical trials. While some variability is systematic due to patient-level factors and treatment effects, random variability across patients is also substantial. Therefore, we developed a generalized mixed-effects approach for using the proposed ADAS-CogMRI biomarker in clinical trials, where the baseline cognitive impairment $\boldsymbol{\theta_i^{t_0}}$, progression rates $\boldsymbol{r_i}$, and baseline MR measurements $\boldsymbol{d_i} = \{d_{i1}, \ldots, d_{iK}\}$ of pa-

tients are modeled as mixed effects.

$$\begin{aligned}
\boldsymbol{\theta_i^0} &= \boldsymbol{\mu_\theta} + \boldsymbol{\beta_{Arm}} \times (\mathrm{Arm}_i) + \boldsymbol{\beta_{Patient}} \times \boldsymbol{P_i} + \boldsymbol{\varepsilon_{i,\theta}} \\
\boldsymbol{r_i} &= \boldsymbol{\mu_r} + \boldsymbol{\gamma_{Arm}} \times (\mathrm{Arm}_i) + \boldsymbol{\gamma_{Patient}} \times \boldsymbol{P_i} + \boldsymbol{\varepsilon_{i,r}} \\
\boldsymbol{d_i} &= \boldsymbol{\mu_d} + \boldsymbol{\chi_{Patient}} \times \boldsymbol{P_i} + \boldsymbol{\varepsilon_{i,d}}
\end{aligned} \tag{7.6}$$

where, $\boldsymbol{\mu_\theta}$, $\boldsymbol{\mu_r}$, and $\boldsymbol{\mu_d}$ denote the average values of baseline cognitive impairment, progression rates, and baseline MR measurements across patients. The categorical covariate Arm_i encodes information on the trial arm of i^{th} patient as $\mathrm{Arm}_i = \begin{cases} 0 & \text{if placebo} \\ 1 & \text{if treatment} \end{cases}$. While $\boldsymbol{\beta_{Arm}}$ controls for differences in average baseline cognitive impairment between placebo and treatment arm patients, $\boldsymbol{\gamma_{Arm}}$ measures the effect of treatment on progression rates. Patient-level covariates $\boldsymbol{P_i}$ were included to model systematic variability in baseline cognitive impairment, progression rates, and baseline MR measurements with $\boldsymbol{\beta_{Patient}}$, $\boldsymbol{\gamma_{Patient}}$, and $\boldsymbol{\chi_{Patient}}$ representing the associated fixed effects. Random effects $\boldsymbol{\varepsilon_{i,\theta}}$, $\boldsymbol{\varepsilon_{i,r}}$, and $\boldsymbol{\varepsilon_{i,d}}$ were included in the model to account for random inter-patient variations in baseline cognitive impairment, progression rates and baseline MR measurements. The baseline cognitive impairment and progression rates in Alzheimer's disease patients are inter-correlated and, therefore, the random effects $\boldsymbol{\varepsilon_{i,\theta}}$ and $\boldsymbol{\varepsilon_{i,r}}$ were allowed to covary $\begin{pmatrix} \boldsymbol{\varepsilon_{i,\theta}} \\ \boldsymbol{\varepsilon_{i,r}} \end{pmatrix} \sim N\big(0, \begin{bmatrix} \Sigma_\theta & \Sigma_{\theta,r} \\ \Sigma_{\theta,r} & \Sigma_r \end{bmatrix}\big)$. Similarly, since differences in brain sizes affect brain regions similarly, the random effects of baseline MR measurements $\boldsymbol{\varepsilon_{i,d}}$ were also allowed to covary.

7.2.5.1 Sensitivity Analysis using Simulated Clinical Trials

We evaluated and compared the sensitivities of the ADAS-CogMRI biomarker and the ADAS-Cog scored using the ADAS-CogIRT and the ANCOVA methodologies by simulating clinical trials focused in the MCI stage (total ADAS-Cog scores: 10±5) and the mild-to-moderate Alzheimer's disease stage (total ADAS-Cog scores: 25±10). The clinical trials were simulated to mimic the complexity of real-world clinical trials by considering unbalanced patient samples in trial arms, systematic and random inter-patient variability in baseline cognitive impairment, progression rates and baseline structural MR measurements, errors in structural MR measurements and dropout of patients during the trials. The parameters for simulating these characteristics were obtained by analyzing longitudinal MR measurements and ADAS-Cog responses of patients using the developed generalized mixed-effects model approach in (7.6). A Cox proportional hazards model was developed for modeling hazard of patient dropout with baseline cognitive impairment, progression rates, and patient-level factors as potential covariates.

In each of the MCI and the mild-to-moderate Alzheimer's disease stages, we performed two simulation experiments to evaluate the sensitivities of the ADAS-CogMRI biomarker, the ADAS-CogIRT, and the ANCOVA methodologies. In the first experiment, the statistical power of the methodologies was evaluated for different sample sizes of 200, 400, 600, 800, and 1000 patients considered in clinical trials of fixed 24-months long duration. For the second

experiment, the sample size was fixed as 400 patients and statistical power of the methodologies was evaluated for different trial durations of 12, 24, 36, and 48 months. These fixed values were selected based on the characteristics of the clinical trial conducted in the past. Both experiments were repeated for four hypothetical treatment effects of Cohen's $d = 0$ (no effect), 0.2 (mild effect), 0.5 (moderate effect), and 0.8 (large effect) simulated in treatment arms of clinical trials [32]. The case of no treatment effect (Cohen's $d = 0$) evaluated the type-I error rates of the methodologies. In both the experiments, patients were followed up biannually until the duration of each trial. The longitudinal ADAS-Cog responses and the MR measurements of patients were simulated using the estimated latent variable models.

In each trial, the simulated data were analyzed using the ADAS-CogMRI biomarker, the ADAS-CogIRT and the ANCOVA methodologies. The statistical significance of the treatment effect was assessed using z-statistic with correction for multiple comparisons. In each simulation experiment, 300 clinical trials were simulated for every possible combination of treatment effect, sample size, and trial duration. The statistical power was evaluated as the proportion of clinical trials wherein a statistically significant treatment effect on patients' progression rates was detected.

7.2.5.2 Sensitivity Analysis in Detecting Differences between Progression Rates of MCI-C and MCI-NC Patients

None of the previous clinical trials that involved MR imaging showed any evidence of a treatment effect, which did not allow validation of the simulation results in a real clinical trial study. Instead, we considered a real world problem of detecting differences between progression rates of MCI-C and MCI-NC patients as a sample clinical trial. The progression rates of MCI-C patients are higher than the progression rates of MCI-NC patients [161, 8, 61, 46, 94]. We posed this problem similar to a real clinical trial, where the MCI-C and MCI-NC patients were assigned to the control and the treatment arms, respectively. This mimics clinical trial of a disease-modifying treatment hypothesized to prevent progression of MCI patients to Alzheimer's disease. We considered only those MCI-NC patients that did not show an evidence of conversion for at least 3 years of follow-up after baseline. Similar to the simulated clinical trials, the statistical power of the ADAS-CogMRI biomarker, and the ADAS-CogIRT and the ANCOVA methodologies were evaluated in detecting differences between progression rates of MCI-C and MCI-NC patients using different sample sizes of 50, 100, 150, 200, and 300 patients and trial durations of 6, 12, 18, and 24 months. The smaller sample sizes and trial durations were considered due to the large treatment effect size involved in this problem. For each possible combination of trial duration and sample size, 300 repetitions were performed with patients selected randomly using bootstrapping with replacement. The statistical power was evaluated as the proportion of repetitions wherein a sta-

tistically significant difference in progression rates of MCI-C and MCI-NC was observed.

7.2.5.3 Enrichment of Clinical Trials in the MCI Stage

Clinical trials in the MCI stage would benefit from sample enrichment by specifically including MCI patients that would convert to Alzheimer's disease in near future. We evaluated the abilities of the ADAS-CogMRI biomarker, the ADAS-CogIRT methodology, and the sole use of cerebral atrophy in predicting MCI patients that would convert to Alzheimer's disease based on patients' MR measurements and ADAS-Cog responses at baseline. In the sole use of cerebral atrophy, dimensionality reduction in the MR measurements was conducted using factor analysis, as described in [43]. Using the training set, Cox-proportional hazard models were developed to test whether baseline cognitive impairment estimated by the methods and patient-level factors are associated with time to conversion. The covariates that were not found to be significant predictors were removed from the model and the hazard models were re-estimated. The re-estimated hazard models were used for estimating survival likelihoods for all patients in the validation set by the end of 3 years of patient follow-up and used for generating the receiver operating characteristic curves. The performance of the three methods in predicting MCI patients that will convert to Alzheimer's disease was calculated in terms of area under the ROC curve (AUC), sensitivity, and specificity.

All data analyses in this study were conducted using the Mplus v6.12 and R software version 3.0.2 environment for statistical computing. The main R scripts for the implementation and evaluation of the ADAS-CogMRI biomarker are available at the following repository: `https://github.com/nishant3115/ADAS-CogMRI-Biomarker`.

7.3 Results

7.3.1 Latent Variable Analysis of Atrophy and the ADAS-Cog

7.3.1.1 Latent Traits Underlying Atrophy and the ADAS-Cog

All the traditional techniques of Kaiser's rule, scree plot, and parallel analysis suggested $m = 4$ latent traits underlying the MR measurements and the ADAS-Cog responses. The ADAS-Cog items loaded only on three out of the four latent traits. The fourth latent trait was determined by the caudal anterior, rostral anterior, posterior, and isthmus portions of the cingulate gyrus, which are associated with executive functioning [163]. The lack of items probing executive functioning in patients is a commonly discussed limitation of the ADAS-Cog [137]. Since none of the ADAS-Cog items loaded on the fourth trait, the trait and the corresponding four MR measurements loading on that trait (thickness measurements of cingulate gyrus) were dropped from further analysis. The three latent traits illustrated acceptable global fit (RMSEA = 0.028, TLI = 0.937, and CFI = 0.930) and good item level fits to the ADAS-Cog items (S-X^2 not significant) and the MR measurements (R^2 values$\geq$0.70

for all brain regions except cuneus, frontal pole, lingual gyrus, and parahippocampal gyrus, which had R2 values $0.60 \leq R2 \leq 0.70$). MR measurements show significant random inter-patient variability, which cannot be explained using cognitive impairment. Therefore, the observed global and item-level fits were considered as acceptable and the latent variable models were used for subsequent analysis.

Based on the nature of the ADAS-Cog items and functions of the brain regions loading on the traits (figure 7.1), the three traits were clinically interpreted as measuring cognitive impairment in the memory, language, and praxis domains. The first trait is determined by the memory-related ADAS-Cog items (word recall, delayed word recall, word recognition, and orientation) and cortical thickness measurements in regions of temporal lobe (entorhinal cortex, hippocampus, amygdala, temporal pole, parahippocampal gyrus, fusiform gyrus, and regions of temporal gyrus), which play important roles in learning and memory [162]. Since most of the ADAS-Cog items probing memory impairment are based on word recall and recognition, lingual gyrus was also associated with the first trait, which is primarily concerned with identification and recognition of words [109].

The language-related ADAS-Cog items and cortical thickness measurements in regions of inferior frontal gyrus (pars opercularis, pars triangularis, and pars orbitalis) dominantly loaded on the second trait. Pars triangularis and pars opercularis are important for speech-language production and dam-

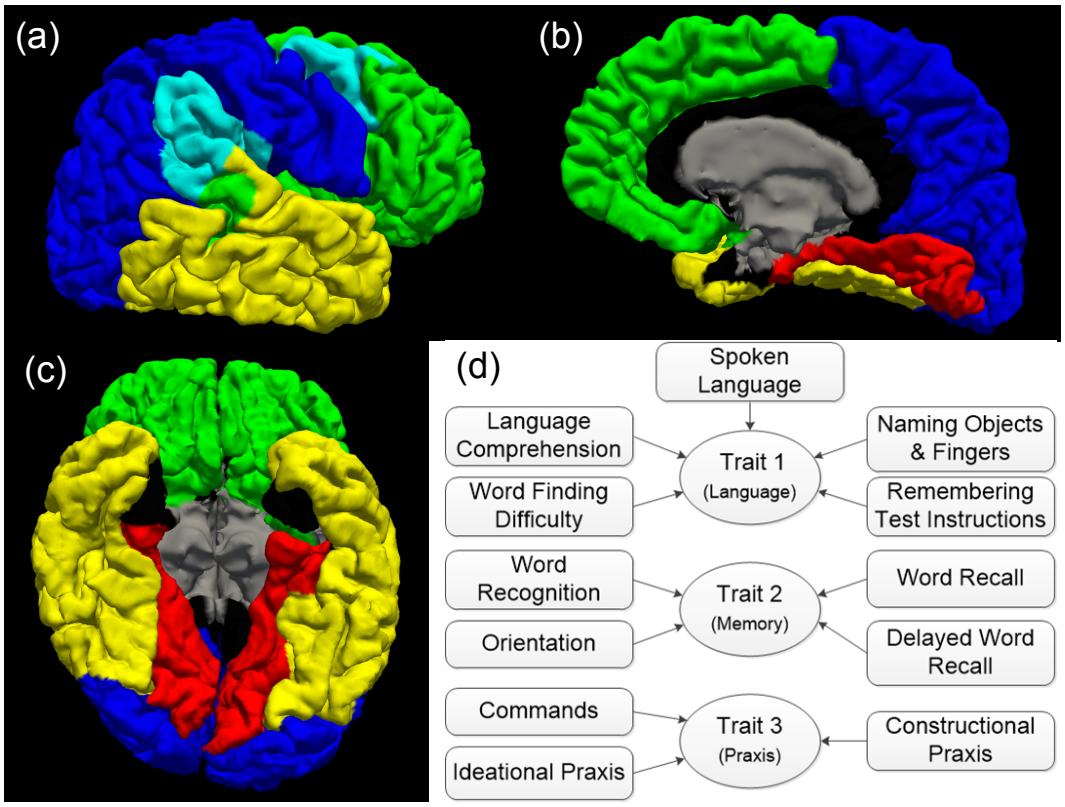


Figure 7.1: Latent traits loading on cerebral atrophy and the ADAS-Cog items: A sample patient's brain showing (a) lateral and (b) medial views of right hemisphere, and (c) inferior view with brain regions color coded as red, green, and blue, based on their loadings on the three traits, which represent cognitive impairment in the memory, language, and praxis domains. The brain regions that cross-load across multiple traits are color coded as cyan (cross-loading on language and praxis factors) and yellow (cross-loading on memory and language factors). The gray and black colors represent regions that are either not brain tissue or were dropped from analysis. Subfigure (d) shows the ADAS-Cog items that load on the three latent traits.

age in these regions has been associated with aphasia [57, 114]. Other brain regions that loaded on the second latent trait included regions in frontal lobe [26, 7], insula [45, 156], and medial temporal lobe [68, 159], which have functions in language comprehension and expression. The third trait was primarily loaded by the praxis-related ADAS-Cog items and regions in the parietal lobe (precuneus cortex, inferior and superior parietal cortices, postcentral gyrus, supramarginal gyrus) and occipital lobe (pericalcarine cortex, cuneus cortex, lateral occipital cortex), which play important roles in motor control and sensory skills [15, 37, 190]. In particular, constructional apraxia in patients has been associated with pathology in occipital cortex [116].

When evaluated on validation set as part of the confirmatory factor analysis, the factor loading structure in figure 7.1 illustrated an acceptable global (RMSEA = 0.032, CFI = 0.921, and TLI = 0.917) and item-level fit to the ADAS-Cog items (S-X^2 not significant) and the MR measurements (similar R^2 values as in the training set).

7.3.1.2 Measurement Invariance of the Latent Variable Models

Several brain regions associated with memory, language, and praxis impairment showed evidence of nonlinear profiles of cerebral atrophy. While tissue loss accelerates in fusiform gyrus, supramarginal gyrus and insula gyrus, it decelerates in lingual gyrus, amygdala, hippocampus, and pericalcarine gyrus with progression of cognitive impairment. Similar nonlinear patterns of cere-

bral atrophy have been reported in previous studies [144, 93]. Among patient-level factors, presence of an APOE-ϵ4 allele and aging are associated with reduced baseline structural measurements in every brain region. Gender was also found to affect baseline structural measurements in certain brain regions. While men have thicker cuneus and larger amygdala, they have thinner frontal pole, pars orbitalis, and smaller hippocampus. Education level was not found to be associated with variations in the MR measurements. As found earlier in chapter 5, several items on the ADAS-Cog illustrate measurement bias due to patient-level factors. While naming the object 'rattle' is easier for women, they are less likely to correctly name 'harmonica' and have more difficulty in drawing a cube. A strong measurement bias due to gender was also observed for the item 'Remembering test instructions', where women are more likely to forget test instructions during administration of the ADAS-Cog. No measurement bias in the ADAS-Cog items was observed due to education level and APOE-ϵgenotype.

7.3.2 Application of the Biomarker in Clinical Trials

The ADAS-CogMRI biomarker provides significant improvement in statistical power to detect treatment effects in clinical trials over the ADAS-Cog scored using the ADAS-CogIRT and the ANCOVA methodologies (figures 7.2–7.5) while maintaining a type-I error rate of $\sim$5% (not shown).

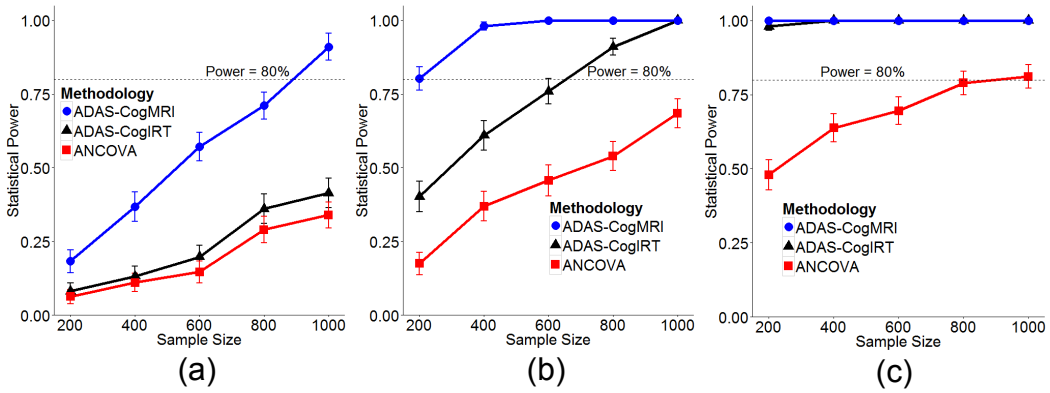


Figure 7.2: Statistical power against sample size in the MCI stage: Plots showing statistical power of the ADAS-CogMRI, the ADAS-CogIRT, and the ANCOVA methodologies in detecting (a) mild ($d = 0.2$), (b) moderate ($d = 0.5$), and (c) large ($d = 0.8$) treatment effects for different sample sizes of 200, 400, 600, and 800 patients considered in simulated clinical trials of 24-months duration.

7.3.2.1 Simulated Clinical Trials in the MCI Stage

For detecting a mild treatment effect, all the three methodologies suffer from low power (figures 7.2-7.3a) due to large inter-patient variability in progression rates, which confounds the detection of a mild treatment effect. However, with an increase in sample size, the performance of the ADAS-CogMRI biomarker improves much more quickly than the ADAS-Cog and achieves the desirable power threshold of 80% with ∼900 patients (figure 7.2a). On the other hand, the ADAS-Cog scored using either the ADAS-CogIRT or the ANCOVA methodologies is unable to achieve 80% power even with large sample size of 1000 patients. When the trial duration is increased, little improvements in statistical power of the methodologies is observed for all the methodologies

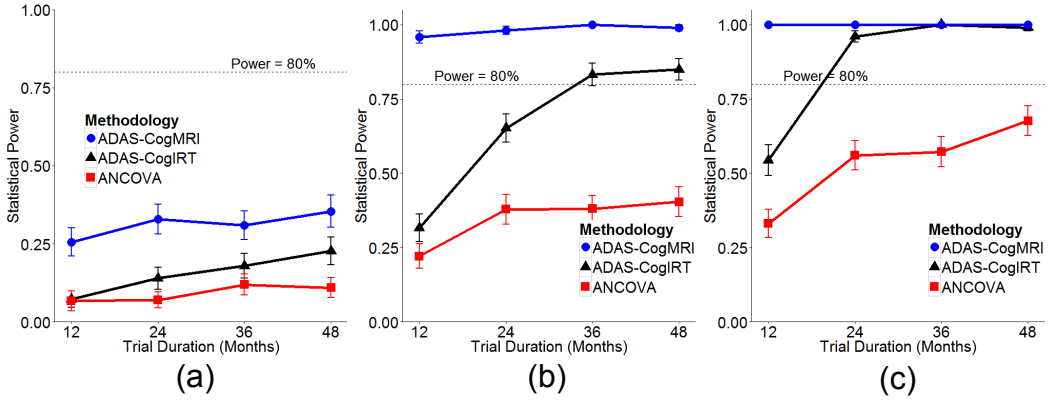


Figure 7.3: Statistical power against trial duration in the MCI stage: Plots showing statistical power of the ADAS-CogMRI, the ADAS-CogIRT, and the ANCOVA methodologies in detecting (a) mild ($d = 0.2$), (b) moderate ($d = 0.5$), and (c) large ($d = 0.8$) treatment effects for different trial durations of 12, 24, 36, and 48 months considered in simulated clinical trials involving 400 patients.

(figure 7.2b). This is because a mild treatment effect is difficult to detect in presence of large inter-patient variability in progression rates even if the measurement accuracy of progression rates is improved by measuring at several follow-up visits.

For a moderate treatment effect, the ADAS-CogMRI methodology illustrates $\geq$80% power for all sample sizes and trial durations (figures 7.2-7.3b). The ADAS-CogIRT methodology also achieves 80% power with a sample size of $\sim$600 patients (figure 7.2b) in a 24-months long trial or 400 patients in a $\sim$36-months long trial. However, the ANCOVA methodology never achieves 80% power even with large sample size of 1000 patients and long trial duration of 4 years. The performance of the ADAS-CogIRT and the ANCOVA method-

ologies significantly improves in detecting a large treatment effect. However, the ADAS-CogIRT methodology requires a minimum longitudinal follow-up of at least ∼24 months in a clinical trial (figure 7.3c) because the ADAS-Cog item scores do not sufficiently change in the MCI stage over short durations. This limitation is also shared in the cases of mild and moderate treatment effects, where the power of the ADAS-CogIRT methodology improves significantly as the trial duration is increased to 24 months (figures 7.3a-b). The power of the ANCOVA methodology also approaches 80% with a large sample size of 1000 patients in a 24-months long trial; however, with a small sample size of 400 patients, the ANCOVA methodology is unable to achieve 80% power even with 4-year long trials.

7.3.2.2 Simulated Clinical Trials in the Mild-to-moderate Alzheimer's Disease Stage

Similar to the MCI stage, all the three methodologies show low power in detecting a mild treatment effect (figures 7.4-7.5a). While the ADAS-CogMRI methodology is able to achieve 80% power with a sample size of ∼800 patients, the ADAS-CogIRT and the ANCOVA methodologies are unable to do so even with large sample sizes (figure 7.4a) and trial durations (figure 7.5a). The ADAS-CogMRI biomarker shows ∼100% power in detecting a moderate treatment effect. The ADAS-CogIRT methodology also shows good performance and achieves 80% statistical power in detecting a moderate treatment effect with a small sample size of ∼300 patients in a 24-months long trial

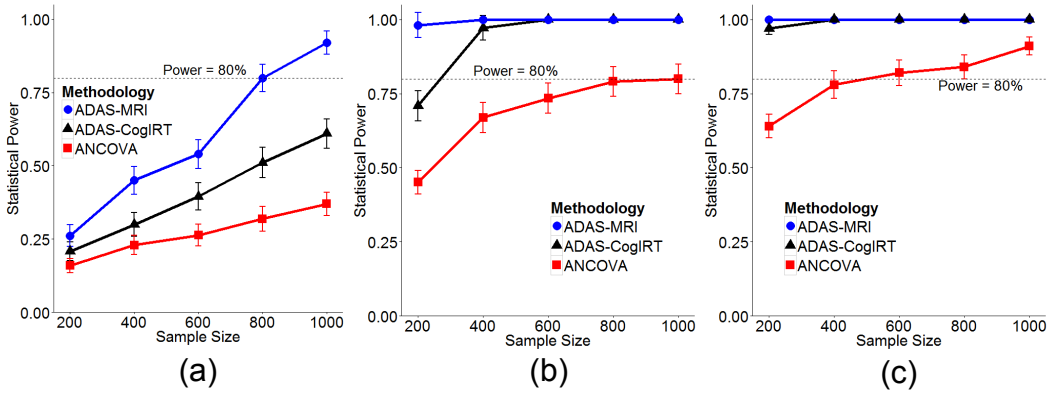


Figure 7.4: Statistical power against sample size in the mild-to-moderate Alzheimer's disease stage: Plots showing statistical power of the ADAS-CogMRI, the ADAS-CogIRT, and the ANCOVA methodologies in detecting (a) mild ($d = 0.2$), (b) moderate ($d = 0.5$), and (c) large ($d = 0.8$) treatment effects for different sample sizes of 200, 400, 600, and 800 patients considered in simulated clinical trials of 24-months long duration.

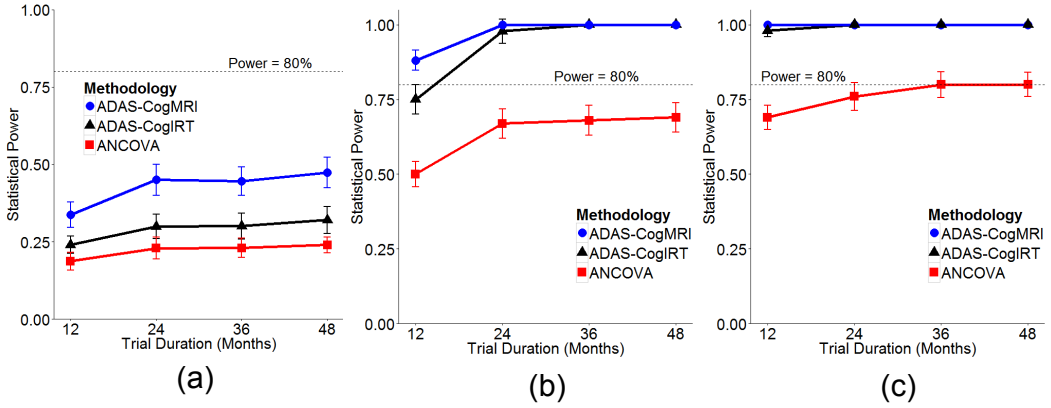


Figure 7.5: Statistical power against trial duration in the mild-to-moderate Alzheimer's disease stage: Plots showing statistical power of the ADAS-CogMRI, the ADAS-CogIRT, and the ANCOVA methodologies in detecting (a) mild ($d = 0.2$), (b) moderate ($d = 0.5$), and (c) large ($d = 0.8$) treatment effects for different trial durations of 12, 24, 36, and 48 months considered in simulated clinical trials involving 400 patients.

(figure 7.4b) or with 400 patients in a ∼15-months long trial (figure 7.5b). The ANCOVA methodology also achieves 80% power in detecting a moderate treatment effect with ∼800 patients in a 24-months trial; however, its performance is low with small sample size even after the trial duration is increased to 4 years (figure 7.5b). For a large treatment effect, both the ADAS-CogMRI biomarker and the ADAS-CogIRT methodology show ∼100% power for all sample sizes and trial durations (figures 7.4-7.5c). The statistical power of the ANCOVA methodology also significantly improves and reaches 80% power with a sample size of ∼500 patients in a 24-months long trial or with 400 patients in a ∼36-months long trial.

7.3.2.3 Sensitivity Comparison in the MCI and the Mild-to-moderate Alzheimer's Disease Stages

When compared to the mild-to-moderate Alzheimer's disease stage, the statistical power of the ADAS-CogIRT and ANCOVA methodologies decrease significantly in the MCI stage due to the language and praxis-related ADAS-Cog items suffering from severe floor effects. In contrast, the performance of the ADAS-CogMRI biomarker stays approximately consistent across the two disease stages, which is a desirable property for an outcome measure in clinical trials. Moreover, while the ADAS-CogIRT and ANCOVA methodologies require at least 24 months of longitudinal follow-up in order to detect progression in language and praxis impairment (figure 7.3a-c) in the MCI stage, the ADAS-CogMRI biomarker measures progression in cognitive impairment

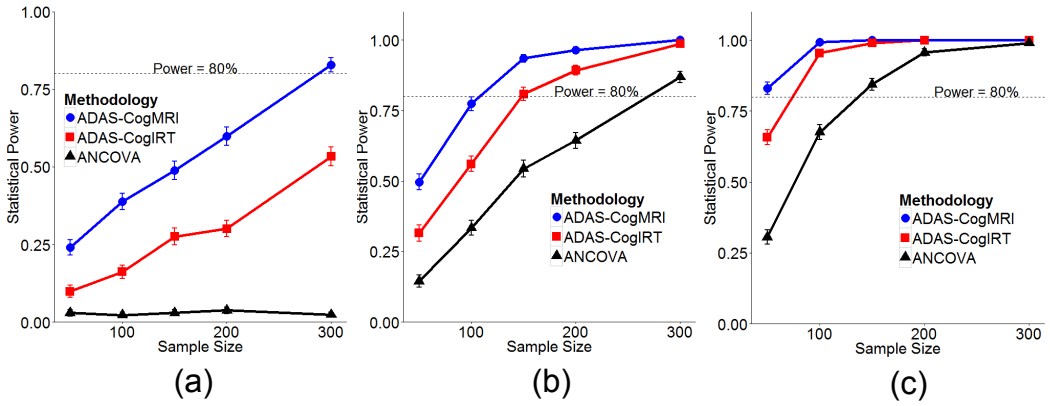


Figure 7.6: Statistical power in detecting differences between MCI-C and MCI-NC patients: Plots showing statistical power of the ADAS-CogMRI, the MRI-FA, the ADAS-CogIRT, and the ANCOVA methodologies in detecting differences between progression rates of MCI-C and MCI-NC patients for varying sample sizes and longitudinal follow-up durations of (a) 6 months, (b) 12 months, and (c) 24 months.

much more consistently. This is also evident from the little improvement in statistical power of the ADAS-CogMRI as the trial duration is increased in the MCI stage (figure 7.3a-c).

7.3.2.4 Sensitivity Analysis in Detecting Differences in Progression Rates between MCI-C and MCI-NC Patients

When validated on detecting differences between progression rates of MCI-C and MCI-NC patients, the ADAS-CogMRI biomarker illustrated better statistical power than the ADAS-CogIRT and the ANCOVA methodologies (figure 7.6). The effect sizes corresponding to differences between progression rates of MCI-C and MCI-NC patients were $d = 0.92$ in the memory, $d = 0.37$

in the language, and $d = 0.28$ in the praxis domains. Due to the large effect size in the memory domain, all the methodologies showed good statistical power in detecting differences between progression rates of MCI-NC and MCI-C patients in a 24-months long study with 200 patients (figure 7.6c). However, when the follow-up duration and sample size are reduced, the statistical powers of the methodologies show a similar pattern as observed in simulation experiments in the MCI stage (figure 7.2a). For a short follow-up duration of 6 months, the ANCOVA and the ADAS-CogIRT methodologies show poor statistical power due to little change in the ADAS-Cog scores in the MCI stage (figure 7.6a). However, the ADAS-CogMRI biomarker shows much better statistical power and is able to achieve 80% power with ∼300 patients. For a follow-up duration of 12 months, the performance of the ADAS-CogIRT and the ANCOVA methodologies improve. While the ADAS-CogMRI biomarker achieves 80% power with ∼100 patients, the ADAS-CogIRT and the ANCOVA methodologies also achieve the threshold with ∼150 patients and ∼250 patients, respectively (figure 7.6b).

7.3.2.5 Enrichment of Clinical Trials in the MCI Stage

Among patient-level factors, APOE genotype is significant in predicting conversion of MCI patients to Alzheimer's disease. The baseline memory (hazards ratio = 3.37, 95% CI: 2.46-4.61, p-value $< 10^{-13}$) and language impairment (hazards ratio = 1.40, 95% CI: 1.037-1.896, p-value = 0.002) calculated using the ADAS-CogMRI biomarker were found to be significant predictors

of conversion to Alzheimer's disease. Baseline memory impairment was found to be a significant predictor also when calculated using the ADAS-CogIRT methodology (hazards ratio = 3.14, 95% CI: 2.37-4.16, p-value $< 10^{-13}$) and the sole use of cerebral atrophy (hazards ratio = 2.31, 95% CI: 1.83-2.90, p-value $< 10^{-11}$). However, baseline impairment in the language domain was not found to be significant using the ADAS-CogIRT methodology and the sole use of cerebral atrophy. Table 7.2 compares the accuracies of the ADAS-CogMRI biomarker, the ADAS-Cog scored using the ADAS-CogIRT methodology, and the sole use of cerebral atrophy in predicting MCI patients that will convert to Alzheimer's disease within 3 years of longitudinal follow-up. The AUC values of the ADAS-CogMRI were significantly better than the AUC values of the ADAS-CogIRT methodology (training p-value = 0.0046, test p-value = 0.01) and the sole use of cerebral atrophy (training p-value = 0.02, test p-value = 0.0044).

7.4 Conclusion

The patients diagnosed with amnestic MCI do not show any noticeable impairment in language and praxis domains [6]. However, as MCI patients progress to Alzheimer's disease, language and praxis abilities of patients deteriorate to an extent that decreases patients' abilities to independently function [108]. Since disease-modifying treatments in the MCI stage aim towards slowing progression to Alzheimer's disease, clinical trials should additionally

Table 7.2: Performance comparison: Table comparing the accuracies of the ADAS-CogMRI biomarker, the ADAS-CogIRT scoring methodology, and the sole use of cerebral atrophy (Atrophy) in predicting MCI patients that will convert to Alzheimer's disease.

Training Set			
Method	AUC	Sensitivity	Specificity
ADAS-CogMRI	0.869	84.04%	76.02%
ADAS-CogIRT	0.832	77.65%	78.76%
Atrophy	0.829	81.91%	75.34%
Validation Set			
ADAS-CogMRI	0.868	83.87%	76.71%
ADAS-CogIRT	0.839	73.11%	80.13%
Atrophy	0.801	78.49%	73.97%

evaluate the effects of treatments in the language and praxis domains. However, the ADAS-Cog outcome measure does not allow evaluation of treatment effects in the language and praxis domains due to the inherent limitations of its items.

Cerebral atrophy is closely related to cognitive impairment and, therefore, can serve as a substrate for tracking progression of cognitive impairment in the MCI stage. A latent variable analysis revealed four latent traits underlying cerebral atrophy due to Alzheimer's disease. These four traits are consistent with the factor analysis results from previous studies [43] and the hierarchical pattern of neurodegeneration observed in Alzheimer's disease [22]. The ADAS-Cog items loaded on three out of the four traits with the same loading structure as observed in an independent psychometric analysis of the ADAS-Cog [180, 164, 117, 89]. The good model fit of the consistent load-

ing structure to the ADAS-Cog items validates the close relationship between cerebral atrophy and cognitive impairment in Alzheimer's disease patients. We extended the latent variable modeling framework to utilize this relationship and developed the proposed ADAS-CogMRI biomarker for more accurate measurement of progression of cognitive impairment in the MCI stage.

When compared with the sole use of the ADAS-Cog, the ADAS-CogMRI biomarker significantly improves the efficacy of clinical trials focused in the MCI stage by reducing sample size and trial duration required for detecting treatment effects (figures 7.2-7.3). Similar improvements were also observed in detecting differences in progression rates between MCI-C and MCI-NC patients (figure 7.6), which validated the results from the simulation experiments. While the use of the ADAS-CogMRI biomarker provides clear benefits over the ADAS-Cog in the MCI stage, the improvement in statistical power in the mild-to-moderate Alzheimer's disease stage is not significant. The ADAS-Cog items that assess language and praxis domains become sensitive in the mild-to-moderate Alzheimer's disease stage and, therefore, the ADAS-Cog scored using the ADAS-CogIRT methodology shows comparable performance as the ADAS-CogMRI biomarker. However, the ADAS-CogMRI methodology may still provide some improvement in statistical power for the case of mild treatment effects in the mild-to-moderate Alzheimer's disease stage.

While not presented here, we also compared the sensitivity of the ADAS-CogMRI biomarker with the sole use of cerebral atrophy in clinical

trials. As described in [43], the structural MR measurements were reduced in dimensionality using factor analysis and the resulting latent traits were used as neuroanatomical scores for patients. The sole use of cerebral atrophy illustrated similar statistical power as the ADAS-CogMRI biomarker in simulated clinical trials and in detecting differences in progression rates between MCI-C and MCI-NC patients. However, the neuroanatomical scores are biased representatives of underlying cognitive impairment in patients. As we briefly discussed in Section 7.2.3, significant random variability exists in the baseline MR measurements after accounting for the systematic variability that is related to disease processes and cognitive impairment. If not accounted in the factor analysis model, the latent traits get biased due to the random inter-patient variability such that the latent traits are smaller in a patient with larger baseline MR measurements than a patient with smaller baseline MR measurements, even if the underlying impairment were the same. This bias in the latent traits is the primary reason behind the lower accuracy of cerebral atrophy in predicting MCI-C patients as compared to the ADAS-CogMRI biomarker (table 7.2). However, this bias is eliminated in longitudinal studies when patients are used as their own controls in clinical trials. Therefore, the sole use of cerebral atrophy results in similar statistical power as the ADAS-CogMRI biomarker in clinical trials.

The combined latent variable modeling of the ADAS-Cog and cerebral atrophy in the ADAS-CogMRI biomarker allows separating the systemic component of covariance due to cognitive impairment from the random component

due to inter-patient variability in baseline MR measurements. Therefore, the ADAS-Cog and cerebral atrophy address each others' limitations. While cerebral atrophy improves measurement of progression of cognitive impairment in the MCI stage, the ADAS-Cog helps in controlling random inter-patient variability in the baseline MR measurements. The proposed ADAS-CogMRI biomarker fulfills all the four requirements of an Alzheimer's disease biomarker laid down by the regulatory agencies [69]. Firstly, the automated image analysis and statistical techniques enable an accurate and reliable measurement of cognitive impairment in Alzheimer's patients [65, 47, 139, 90, 145, 70, 134, 177]. Secondly, the ADAS-CogMRI biomarker shows an acceptable sensitivity ($\sim$84%) and specificity ($\sim$76%) in diagnosing MCI patients that will convert to Alzheimer's disease in future. Thirdly, the ADAS-CogMRI biomarker also shows good sensitivity in detecting treatment effects in clinical trials. Moreover, it allows evaluation of treatment effects in the language and praxis domains, which is not possible using the ADAS-Cog in the MCI stage. Fourthly, the ADAS-CogMRI biomarker measures the extent of cognitive impairment in Alzheimer's patients, which is a clinically important outcome. Besides these properties, the ADAS-CogMRI biomarker is easy to implement since both MR imaging and the ADAS-Cog are already utilized in clinical trials. MR imaging is routinely used in clinical trials for patient screening at baseline, and evaluating safety of treatment during the study.

The ADAS-CogMRI biomarker and this study suffer from several limitations. Firstly, the ADNI data used in this study contains strictly screened

MCI patients, which do not represent the heterogeneity typically encountered in clinical trials such as the presence of multiple pathologies. As a consequence, when evaluated in a real clinical trial, the statistical powers of the proposed ADAS-CogMRI biomarker and the ADAS-Cog are expected to be lower than reported in this study. Secondly, reduction in inflammation from treatments may result in overestimation of progression rates using the ADAS-CogMRI biomarker and confound detection of treatment effects, a limitation shared by all imaging biomarkers. The treatments that focus at removing amyloid plaques may reduce inflammation in patients' brains in a region-specific or widespread manner, resulting in increased atrophy rates at the start of trials. If the inflammation reduction is regional, the effect on the ADAS-CogMRI biomarker will be little because the ADAS-CogMRI biomarker combines information from several brain regions to estimate progression rates within each cognitive domain. However, if the effect is widespread, additional strategies (such as modeling inflammation reduction or analyzing data from initial phase of trial separately) may be required to separate the overlapping effects of inflammation reduction and neurodegeneration. Thirdly, due to the lack of the ADAS-Cog items probing executive functioning in patients, we dropped the fourth latent trait from subsequent analysis since it was biased due to random inter-patient variability in baseline measurements. However, as executive functioning items are developed and added in the ADAS-Cog [158], the fourth latent trait may be included in the ADAS-CogMRI biomarker because cingulate cortex has been reported to be involved in early stages of Alzheimer's

disease [110].

Despite these limitations, the proposed ADAS-CogMRI biomarker is highly significant for improving efficacy of clinical trials in the MCI stage. The ADAS-CogMRI biomarker has significantly better sensitivity than the ADAS-Cog in the MCI stage and allows evaluation of treatment effects in the language and praxis cognitive domains. Since both the ADAS-Cog and structural MR imaging are routinely utilized, future clinical trials should consider the use of the proposed ADAS-CogMRI biomarker as part of the secondary efficacy analysis to establish the improvement in statistical power obtained over the use of the ADAS-Cog.

Chapter 8

Conclusion and Future Work

This dissertation attempts to advance two active area of research in Alzheimer's disease. The first research area deals with the development of automatic algorithms for analysis of brain MR volumes. The measurement of cerebral atrophy on brain MR volumes requires a sequence of low-level image analysis tasks as prerequisite steps. As a result, performance of any atrophy based biomarker is directly impacted by the accuracy achieved in these low-level image analysis tasks. MR tissue segmentation is one such low-level task, which is required for the measurement of cerebral atrophy within subcortical and cortical brain structures. The measurement of cortical thickness across the brain mantle also requires MR tissue segmentation for delineating the boundary between white matter and gray matter and the boundary between gray matter and cerebrospinal fluid. While an easy task for humans, the presence of image corruptions in MR volumes makes automatic MR tissue segmentation a difficult task due to significant intensity overlap between the tissue classes. In this dissertation, we present a new knowledge-driven decision theory (KDT) approach for MR tissue segmentation, which embeds prior knowledge on relative extents of intensity overlaps between the tissue classes

in the segmentation framework. When evaluated and compared with existing segmentation approaches, the strategy of incorporating prior intensity overlap knowledge is found to be promising in correctly classifying voxels that belong in the intensity overlap spectrum without needing a preprocessing step for removal of intensity inhomogeneities.

The second area of Alzheimer's disease research and the main focus of this dissertation pertains to improving the efficiency of clinical trials of disease-modifying treatments. The currently utilized ADAS-Cog outcome measure has inadequate sensitivity in measuring progression of cognitive impairment, which severely affects the efficiency of clinical trials [27, 131, 75, 74]. The 99.6% failure rate in over 400 clinical trials conducted during the last decade is highest among any therapeutic area and provides a testament to the low efficiency of Alzheimer's clinical trials [38]. While none of the disease-modifying treatments were successful in the last decade, the only drug that got approved is a symptomatic cognitive enhancer. The limitations associated with the current ADAS-Cog scoring methodology is one of the primary reasons behind the low sensitivity of the ADAS-Cog. One of the contributions of this dissertation is an improved ADAS-CogIRT scoring methodology for the ADAS-Cog. The ADAS-CogIRT scoring methodology measures cognitive impairment more accurately in Alzheimer's disease patients and makes clinical trials more efficient by reducing the sample size and follow-up duration required for investigating treatments. More importantly, as validated in the huperzine A trial, the ADAS-CogIRT scoring methodology allows for the detection of treatment ef-

fects that may be missed by using the current scoring methodology. With an increasing prevalence of Alzheimer's disease and the lack of a treatment, such a boost in the efficiency of clinical trials is highly desirable. It would enable rapid testing of future treatments, while making the clinical trials more cost-effective. An improvement in clinical trial efficiency may also provide a boost to the development of new disease-modifying treatments. The failure of all treatments investigated till date and the high cost of clinical trials has significantly impacted industry-led research efforts towards developing novel treatments for Alzheimer's disease.

Addressing the other primary reason behind the low sensitivity of the ADAS-Cog requires modification of its existing items and/or addition of new items to the ADAS-Cog, which probe more subtle levels of cognitive impairment. It is worthwhile to note that addressing the floor effects of the ADAS-Cog items does not reduce the significance associated with using the ADAS-CogIRT scoring methodology. Even if more sensitive items are included in the ADAS-Cog, the limitations of the current scoring methodology would still persist. Therefore, the use of the ADAS-CogIRT scoring methodology would still provide improvements in the sensitivity of the ADAS-Cog in clinical trials. In fact, due to the scale-independent property of its parameters, the ADAS-CogIRT methodology provides a convenient framework for easily adding or removing items from the ADAS-Cog without the need for re-estimation of parameters or measurement scale properties.

Besides the limitations of the ADAS-Cog, another reason behind the low efficiency of clinical trials is the advanced stage of Alzheimer's disease, where clinical trials have traditionally focused. The scope for improvement in clinical performance of dementia patients is low due to the significant amount of neurodegeneration that has already occurred in their brains. Noting this limitation, clinical trials have started to shift their focus towards the prodromal MCI stage of Alzheimer's disease. However, clinical trials in the MCI stage also suffer from several limitations. The sensitivity of the ADAS-Cog is even lower in the MCI stage as compared to the mild-to-moderate Alzheimer's disease stage. Moreover, the inability to specifically select MCI patients that will convert to Alzheimer's disease in future further impacts the efficiency of clinical trials in the MCI stage. Towards this end, the last contribution of this dissertation presents a biomarker, which uses cerebral atrophy as a proxy measure of cognitive impairment in clinical trials. The ADAS-CogMRI biomarker is designed based on a combined latent variable analysis of the ADAS-Cog and cerebral atrophy, which revealed that the spatio-temporal patterns of brain-wide atrophy are closely related with cognitive impairment assessed on the ADAS-Cog. When compared with the sole use of the ADAS-Cog, the proposed biomarker provides significant improvements in efficiency of clinical trials focused in the MCI stage. The ADAS-CogMRI biomarker also improves efficiency of clinical trials by facilitating early detection of MCI patients that will convert to Alzheimer's disease in future with an acceptable sensitivity of $\sim$84% and specificity of $\sim$76%.

The work in this dissertation sets up a number of research questions, which should be considered as part of the future work. While the developed KDT segmentation approach shows better ability in handling intensity overlaps between the tissue classes, some sensitivity to the presence of high levels of intensity inhomogeneities was observed. As part of the future work, simultaneous correction of intensity inhomogeneities in MR volumes can be investigated by including additional terms in the energy function. Tissue segmentation and correction of intensity inhomogeneities are interdependent tasks and, therefore, tissue segmentation is expected to benefit from the simultaneous correction of intensity inhomogeneities. Since the relative extents of intensity overlap between the tissue classes stay consistent across different levels of intensity inhomogeneities, simultaneous correction should not impact the Bayesian decision theory energy function that drives tissue segmentation. Intensity inhomogeneities are smoothly varying in nature and, therefore, regularizing terms imposing smoothness constraints on the estimated inhomogeneities should be included. A related idea of simultaneously correcting for intensity inhomogeneities in a level set framework has been investigated in a prior work from our group [179].

Our research work towards improving the efficacy of clinical trials of disease-modifying treatments also opens several avenues of future research. In the development of the ADAS-CogIRT scoring methodology, we could not investigate measurement invariance of the ADAS-Cog across patient-level factors of race and ethnicity due to the lack of patient heterogeneity. Race and

ethnicity are common considerations in any medical problem as diversity with respect to these factors is typically a norm rather than an exception. If not investigated as part of the future work, the ADAS-CogIRT scoring methodology may produce biased measurements of cognitive impairment in patient groups found to respond to the ADAS-Cog items differently than the patient population considered in this dissertation. While addressing this would likely involve data collection by the researchers as most public datasets contain predominantly non-Hispanic Caucasian patients, it is a key research consideration before the ADAS-CogIRT scoring methodology can be utilized for any clinical purposes. Another area of future research can be to investigate nonlinear trends in progression in clinical trials. While we assumed a linear profile based on support from the existing literature, the developed generalized mixed-effects model framework does provide the flexibility for investigating more complex progression trends in clinical trials. The use of nonlinear progression profiles in clinical trials has been argued in literature and, therefore, the flexibility of the ADAS-CogIRT scoring methodology may be significant, specially for long term studies.

Most of the disease-modifying treatments target amyloid plaque depositions in patients' brains. The removal of amyloid plaques is sometimes associated with a side-effect of reduction in brain inflammation resulting in a region-specific or brain-wide shrinkage of brain tissue. While inflammation reduction is desirable and an early indicator of the disease-modifying effects of the treatments, the associated shrinkage in brain tissue confounds with pro-

gressive atrophy due to Alzheimer's disease. If an atrophy based biomarker does not include the effects of inflammation reduction, it may overestimate progression rates in the treatment arm during the initial phase of the trial. This may give an indication about the treatment either not working or even worsening neurodegeneration due to Alzheimer's disease, resulting in withdrawal of the treatment from clinical trial. On the other hand, clinical rating scales are not affected by inflammation reduction in patients' brains. This further supports the combined utilization of cerebral atrophy and the ADAS-Cog as it reduces the sensitivity of a cerebral atrophy based outcome measure to effects of inflammation reduction in clinical trials. As part of the future work, dynamics of brain tissue shrinkage from inflammation reduction can be studied through statistical modeling of data from clinical trials of amyloid targeting treatments. Such models of brain tissue shrinkage would be clinically significant as they can be incorporated in the design of any atrophy based biomarker (such as the ADAS-CogMRI biomarker) to account for the effects of inflammation reduction in future clinical trials. While data from clinical trials of disease-modifying treatments is not yet available, the expanding ADCS cohort is likely to include data from such clinical trials in future. In the absence of inflammation reduction models, an alternative strategy could be to only consider the ADAS-Cog responses and drop cerebral atrophy measurements from the initial part of clinical trials. However, inflammation reduction still needs to be better understood in order to determine the initial period of the clinical trials, where cerebral atrophy should not be included to study

treatment effects.

Another possible future research direction is to extend the ADAS-CogMRI biomarker to additionally measure impairment in executive functioning in Alzheimer's disease patients. We considered the most commonly employed version of the ADAS-Cog for developing the ADAS-CogMRI biomarker, which does not contain any items that probe executive functioning in patients. Even though a latent trait measuring executive functioning was identified in our analysis, we dropped the trait in order to avoid the bias introduced from the sole use of cerebral atrophy. A prior work has proposed additional items for the ADAS-Cog (such as the maze task), which specifically assess executive functioning [158]. As part of the future work, these items can be included to define the fourth latent trait in the ADAS-CogMRI biomarker for assessing impairment in executive functioning in patients. The response data for these new items already exists for a fraction of patients in the ADNI cohort. The brain regions associated with executive functioning have been reported to be early sites of neurodegeneration and, therefore, the inclusion of executive functioning in the ADAS-CogMRI biomarker is expected to improve not only its sensitivity in clinical trials but also its ability to early detecting MCI patients that will convert to Alzheimer's disease in near future.

Bibliography

[1] 2015 Alzheimer's disease facts and figures. *Alzheimer's & Dementia*, 11(3):332 – 384, 2015.

[2] Richard A Grove, Conn M Harrington, Andreas Mahler, Isabel Beresford, Paul Maruff, Martin T Lowy, Andrew P Nicholls, Rebecca L Boardley, Alienor C Berges, Pradeep J Nathan, et al. A Randomized, Double-Blind, Placebo-Controlled, 16-Week Study of the H3 Receptor Antagonist, GSK239512 as a Monotherapy in Subjects with Mild-to-Moderate Alzheimer's Disease. *Current Alzheimer Research*, 11(1):47–58, 2014.

[3] Paul S Aisen, Lon S Schneider, Mary Sano, Ramon Diaz-Arrastia, Christopher H van Dyck, Myron F Weiner, Teodoro Bottiglieri, Shelia Jin, Karen T Stokes, Ronald G Thomas, et al. High-dose B vitamin supplementation and cognitive decline in Alzheimer disease: a randomized controlled trial. *Journal of American Medical Association*, 300(15):1774–1783, 2008.

[4] Ayelet Akselrod-Ballin, Meirav Galun, John Moshe Gomori, Achi Brandt, and Ronen Basri. Prior knowledge driven multiscale segmentation of brain MRI. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI*, pages 118–126. Springer, 2007.

[5] Ayelet Akselrod-Ballin, Meirav Galun, Moshe John Gomori, Ronen Basri, and Achi Brandt. Atlas guided identification of brain structures by combining 3D segmentation and SVM classification. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 209–216. Springer, 2006.

[6] Marilyn S. Albert, Steven T. DeKosky, Dennis Dickson, Bruno Dubois, Howard H. Feldman, Nick C. Fox, Anthony Gamst, David M. Holtzman, William J. Jagust, Ronald C. Petersen, and others. The diagnosis of mild cognitive impairment due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & Dementia*, 7(3):270–279, 2011.

[7] Michael P Alexander, D Frank Benson, and Donald T Stuss. Frontal lobes and language. *Brain and Language*, 37(4):656–691, 1989.

[8] HA Archer, J Kennedy, J Barnes, T Pepple, R Boyes, K Randlesome, S Clegg, KK Leung, S Ourselin, C Frost, et al. Memory complaints and increased rates of brain atrophy: risk factors for mild cognitive impairment and Alzheimer's disease. *International Journal of Geriatric Psychiatry*, 25(11):1119–1126, 2010.

[9] John Ashburner and Karl J Friston. Voxel-based morphometry–the methods. *Neuroimage*, 11(6):805–821, 2000.

[10] Julius Ashkin and Edward Teller. Statistics of two-dimensional lattices with four components. *Physical Review*, 64(5-6):178, 1943.

[11] Suyash P Awate, Tolga Tasdizen, Norman Foster, and Ross T Whitaker. Adaptive Markov modeling for mutual-information-based, unsupervised MRI brain-tissue classification. *Medical Image Analysis*, 10(5):726–739, 2006.

[12] Steve Balsis, Alexis A Unger, Jared F Benge, Lisa Geraci, and Rachelle S Doody. Gaining precision on the Alzheimers Disease Assessment Scale-cognitive: A comparison of item response theory-based scores and total scores. *Alzheimer's & Dementia*, 8(4):288–294, 2012.

[13] Laurel A Beckett, Danielle J Harvey, Anthony Gamst, Michael Donohue, John Kornak, Hao Zhang, Julie H Kuo, Alzheimer's Disease Neuroimaging Initiative, et al. The Alzheimer's Disease Neuroimaging Initiative: Annual change in biomarkers and clinical outcomes. *Alzheimer's & Dementia*, 6(3):257–264, 2010.

[14] Jared F Benge, Steve Balsis, Lisa Geraci, Paul J Massman, and Rachelle S Doody. How well do the ADAS-cog and its subscales measure cognitive dysfunction in Alzheimer's disease ? *Dementia and Geriatric Cognitive Disorders*, 28(1):63, 2009.

[15] D Frank Benson, William A Sheremata, Remi Bouchard, Joseph M Segarra, Donald Price, and Norman Geschwind. Conduction aphasia: a clinicopathological study. *Archives of Neurology*, 28(5):339–346, 1973.

[16] James O Berger. *Statistical decision theory and Bayesian analysis*. Springer, 1985.

[17] Julian Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society, Series B: Methodological*, pages 192–236, 1974.

[18] M Bobinski, MJ De Leon, J Wegiel, S Desanti, A Convit, LA Saint Louis, H Rusinek, and HM Wisniewski. The histological validation of post mortem magnetic resonance imaging-determined hippocampal volume in Alzheimer's disease. *Neuroscience*, 95(3):721–725, 1999.

[19] R Darrell Bock and Murray Aitkin. Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, 46(4):443–459, 1981.

[20] ZI Botev, JF Grotowski, and DP Kroese. Kernel density estimation via diffusion. *The Annals of Statistics*, 38(5):2916–2957, 2010.

[21] Yuri Boykov, Olga Veksler, and Ramin Zabih. Markov random fields with efficient approximations. In *Proc. IEEE computer society conference on Computer vision and pattern recognition*, pages 648–655, 1998.

[22] H. Braak and E. Braak. Neuropathological stageing of Alzheimer-related changes. *Acta Neuropathologica*, 82(4):239–259, 1991.

[23] Peter A Brex, Olga Ciccarelli, Jonathon I O'Riordan, Michael Sailer, Alan J Thompson, and David H Miller. A longitudinal study of abnor-

malities on MRI and disability from multiple sclerosis. *New England Journal of Medicine*, 346(3):158–164, 2002.

[24] Mariano Cabezas, Arnau Oliver, Xavier Lladó, Jordi Freixenet, and Meritxell Bach Cuadra. A review of atlas-based segmentation for magnetic resonance brain images. *Computer Methods and Programs in Biomedicine*, 104(3):e158–e177, 2011.

[25] Li Cai. Metropolis-Hastings Robbins-Monro algorithm for confirmatory item factor analysis. *Journal of Educational and Behavioral Statistics*, 35(3):307–335, 2010.

[26] Qing Cai, Michal Lavidor, Marc Brysbaert, Yves Paulignan, and Tatjana A Nazir. Cerebral lateralization of frontal lobe language processes and lateralization of the posterior visual word processing system. *Journal of Cognitive Neuroscience*, 20(4):672–681, 2008.

[27] Stefan J Cano, Holly B Posner, Margaret L Moline, Stephen W Hurt, Jina Swartz, Tim Hsu, and Jeremy C Hobart. The ADAS-cog in Alzheimer's disease clinical trials: psychometric evaluation of the sum and its parts. *Journal of Neurology, Neurosurgery & Psychiatry*, 81(12):1363–1368, 2010.

[28] Tony F Chan and Luminita A Vese. Active contours without edges. *IEEE Transactions on Image Processing*, 10(2):266–277, 2001.

[29] Kunal N Chaudhury and KR Ramakrishnan. Stability and convergence of the level set method in computer vision. *Pattern Recognition Letters*, 28(7):884–893, 2007.

[30] Wen-Hung Chen and David Thissen. Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, 22(3):265–289, 1997.

[31] Gordon W Cheung and Roger B Rensvold. Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, 9(2):233–255, 2002.

[32] Jacob Cohen. *Statistical power analysis for the behavioral sciences*. Academic press, 2013.

[33] Donald J Connor and Marwan N Sabbagh. Administration and scoring variance on the ADAS-Cog. *Journal of Alzheimer's Disease*, 15(3):461–464, 2008.

[34] Paul K Crane, Gerald van Belle, and Eric B Larson. Test bias in a cognitive test: differential item functioning in the CASI. *Statistics in Medicine*, 23(2):241–256, 2004.

[35] Daniel Cremers, Mikael Rousson, and Rachid Deriche. A review of statistical approaches to level set segmentation: integrating color, texture, motion and shape. *International Journal of Computer Vision*, 72(2):195–215, 2007.

[36] Meritxell Bach Cuadra, Leila Cammoun, Torsten Butz, Olivier Cuisenaire, and J-P Thiran. Comparison and validation of tissue modelization and statistical classification methods in T1-weighted MR brain images. *IEEE Transactions on Medical Imaging*, 24(12):1548–1565, 2005.

[37] Jody C Culham, Cristiana Cavina-Pratesi, and Anthony Singhal. The role of parietal cortex in visuomotor control: what have we learned from neuroimaging ? *Neuropsychologia*, 44(13):2668–2684, 2006.

[38] Jeffrey L Cummings, Travis Morstorf, and Kate Zhong. Alzheimers disease drug-development pipeline: few candidates, frequent failures. *Alzheimers Research Therapy*, 6(37):10–1186, 2014.

[39] Anders M Dale, Bruce Fischl, and Martin I Sereno. Cortical surface-based analysis: I. Segmentation and surface reconstruction. *Neuroimage*, 9(2):179–194, 1999.

[40] Pritam Das, M Paul Murphy, Linda H Younkin, Steven G Younkin, and Todd E Golde. Reduced effectiveness of Aβ1-42 immunization in APP transgenic mice with significant amyloid deposition. *Neurobiology of Aging*, 22(5):721–727, 2001.

[41] Rafael Jaime De Ayala. *Theory and practice of item response theory*. Guilford Publications, 2013.

[42] Steven T DeKosky and Stephen W Scheff. Synapse loss in frontal cortex biopsies in Alzheimer's disease: correlation with cognitive severity.

Annals of Neurology, 27(5):457–464, 1990.

[43] Rahul S. Desikan, Howard J. Cabral, Fabio Settecase, Christopher P. Hess, William P. Dillon, Christine M. Glastonbury, Michael W. Weiner, Nicholas J. Schmansky, David H. Salat, Bruce Fischl, and others. Automated MRI measures predict progression to Alzheimer's disease. *Neurobiology of Aging*, 31(8):1364–1374, 2010.

[44] Rahul S Desikan, Florent Ségonne, Bruce Fischl, Brian T Quinn, Bradford C Dickerson, Deborah Blacker, Randy L Buckner, Anders M Dale, R Paul Maguire, Bradley T Hyman, et al. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *Neuroimage*, 31(3):968–980, 2006.

[45] Hugues Duffau, Luc Bauchet, Stéphane Lehéricy, and Laurent Capelle. Functional compensation of the left dominant insula for language. *Neuroreport*, 12(10):2159–2163, 2001.

[46] Sahar Elahi, Alvin H Bachman, Sang Han Lee, John J Sidtis, and Babak A Ardekani. Corpus Callosum Atrophy Rate in Mild Cognitive Impairment and Prodromal Alzheimer's Disease. *Journal of Alzheimer's Disease*, 45(3):921–931, 2015.

[47] M. Ewers, S. J. Teipel, O. Dietrich, S. O. Schonberg, F. Jessen, R. Heun, P. Scheltens, L. van de Pol, N. R. Freymann, H.-J. Moeller, and others. Multicenter assessment of reliability of cranial MRI. *Neurobiology of Aging*, 27(8):1051–1059, 2006.

[48] Howard H. Feldman, Steven Ferris, Bengt Winblad, Nikolaos Sfikas, Linda Mancione, Yunsheng He, Sibel Tekin, Alistair Burns, Jeffrey Cummings, Teodoro del Ser, and others. Effect of rivastigmine on delay to diagnosis of Alzheimer's disease from mild cognitive impairment: the InDDEx study. *The Lancet Neurology*, 6(6):501–512, 2007.

[49] Adelino R Ferreira da Silva. A Dirichlet process mixture model for brain MRI tissue classification. *Medical Image Analysis*, 11(2):169–182, 2007.

[50] Bruce Fischl and Anders M Dale. Measuring the thickness of the human cerebral cortex from magnetic resonance images. *Proceedings of the National Academy of Sciences*, 97(20):11050–11055, 2000.

[51] Bruce Fischl, David H Salat, Evelina Busa, Marilyn Albert, Megan Dieterich, Christian Haselgrove, Andre Van Der Kouwe, Ron Killiany, David Kennedy, Shuna Klaveness, et al. Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. *Neuron*, 33(3):341–355, 2002.

[52] Bruce Fischl, David H Salat, André JW van der Kouwe, Nikos Makris, Florent Ségonne, Brian T Quinn, and Anders M Dale. Sequence-independent segmentation of magnetic resonance images. *Neuroimage*, 23:S69–S84, 2004.

[53] Bruce Fischl, Martin I Sereno, Roger BH Tootell, Anders M Dale, et al. High-resolution intersubject averaging and a coordinate system for the cortical surface. *Human Brain Mapping*, 8(4):272–284, 1999.

[54] Bruce Fischl, André van der Kouwe, Christophe Destrieux, Eric Halgren, Florent Ségonne, David H Salat, Evelina Busa, Larry J Seidman, Jill Goldstein, David Kennedy, et al. Automatically parcellating the human cerebral cortex. *Cerebral Cortex*, 14(1):11–22, 2004.

[55] Adam S Fleisher, Michael Donohue, Kewei Chen, James B Brewer, Paul S Aisen, Alzheimers Disease Neuroimaging Initiative, et al. Applications of neuroimaging to disease-modification trials in Alzheimer's disease. *Behavioural Neurology*, 21(1-2):129–136, 2009.

[56] Food, Drug Administration, et al. Guidance for industry Alzheimers disease: developing drugs for the treatment of early stage disease. *Food and Drug Administration, Silver Springs, MD*, 2013.

[57] Anne L Foundas, Kathy F Eure, Laura F Luevano, and Daniel R Weinberger. MRI asymmetries of Broca's area: the pars triangularis and pars opercularis. *Brain and Language*, 64(3):282–296, 1998.

[58] NC Fox, RI Scahill, WR Crum, and MN Rossor. Correlation between rates of brain atrophy and cognitive decline in AD. *Neurology*, 52(8):1687–1687, 1999.

[59] Richard Frank and Richard Hargreaves. Clinical biomarkers in drug discovery and development. *Nature Reviews Drug Discovery*, 2(7):566–580, 2003.

[60] Richard A Frank, Douglas Galasko, Harald Hampel, John Hardy, Mony J de Leon, Pankaj D Mehta, Joseph Rogers, Eric Siemers, and John Q Trojanowski. Biological markers for therapeutic trials in Alzheimers disease: proceedings of the biological markers working group; NIA initiative on neuroimaging in Alzheimers disease. *Neurobiology of Aging*, 24(4):521–536, 2003.

[61] Edit Frankó, Olivier Joly, Alzheimers Disease Neuroimaging Initiative, et al. Evaluating Alzheimer's disease progression using rate of regional hippocampal atrophy. *PloS one*, 8(8):e71354, 2013.

[62] Giovanni B Frisoni, Nick C Fox, Clifford R Jack, Philip Scheltens, and Paul M Thompson. The clinical use of structural MRI in Alzheimer disease. *Nature Reviews Neurology*, 6(2):67–77, 2010.

[63] Monica Garcia-Alloza, Meenakshi Subramanian, Diana Thyssen, Laura A Borrelli, Abdul Fauq, Pritam Das, Todd E Golde, Bradley T Hyman, and Brian J Bacskai. Existing plaques and neuritic abnormalities in APP: PS1 mice are not affected by administration of the gamma-secretase inhibitor LY-411575. *Molecular Neurodegeneration*, 4:19, 2009.

[64] Douglas J Gelb, Eugene Oliver, and Sid Gilman. Diagnostic criteria for Parkinson disease. *Archives of Neurology*, 56(1):33, 1999.

[65] Jay N. Giedd, Patricia Kozuch, Debra Kaysen, A. Catherine Vaituzis, Susan D. Hamburger, John J. Bartko, and Judith L. Rapoport. Reliability of cerebral measures in repeated examinations with magnetic

resonance imaging. *Psychiatry Research: Neuroimaging*, 61(2):113–119, 1995.

[66] Teresa Gómez-Isla, Richard Hollister, Howard West, Stina Mui, John H Growdon, Ronald C Petersen, Joseph E Parisi, and Bradley T Hyman. Neuronal loss correlates with but exceeds neurofibrillary tangles in Alzheimer's disease. *Annals of Neurology*, 41(1):17–24, 1997.

[67] Hayit Greenspan, Amit Ruf, and Jacob Goldberger. Constrained Gaussian mixture model framework for automatic segmentation of MR brain images. *IEEE Transactions on Medical Imaging*, 25(9):1233–1245, 2006.

[68] Michael M Haglund, Mitchel S Berger, Michael Shamseldin, Etorre Lettich, and George A Ojemann. Cortical localization of temporal lobe language sites in patients with gliomas. *Neurosurgery*, 34(4):567–576, 1994.

[69] Harald Hampel, Richard Frank, Karl Broich, Stefan J. Teipel, Russell G. Katz, John Hardy, Karl Herholz, Arun LW Bokde, Frank Jessen, Yvonne C. Hoessler, and others. Biomarkers for Alzheimer's disease: academic, industry and regulatory perspectives. *Nature Reviews Drug Discovery*, 9(7):560–574, 2010.

[70] Xiao Han, Jorge Jovicich, David Salat, Andre van der Kouwe, Brian Quinn, Silvester Czanner, Evelina Busa, Jenni Pacheco, Marilyn Albert, Ronald Killiany, et al. Reliability of MRI-derived measurements of

human cerebral cortical thickness: the effects of field strength, scanner upgrade and manufacturer. *Neuroimage*, 32(1):180–194, 2006.

[71] John Harrison, Sonia L Minassian, Lisa Jenkins, Ronald S Black, Martin Koller, and Michael Grundman. A neuropsychological test battery for use in Alzheimer disease clinical trials. *Archives of Neurology*, 64(9):1323–1329, 2007.

[72] Liesi E Hebert, Jennifer Weuve, Paul A Scherr, and Denis A Evans. Alzheimer disease in the united states (2010–2050) estimated using the 2010 census. *Neurology*, 80(19):1778–1783, 2013.

[73] Robin K Henson and J Kyle Roberts. Use of exploratory factor analysis in published research common errors and some comment on improved practice. *Educational and Psychological measurement*, 66(3):393–416, 2006.

[74] Jeremy Hobart, Stefan Cano, Holly Posner, Ola Selnes, Yaakov Stern, Ronald Thomas, and John Zajicek. Putting the Alzheimers cognitive test to the test I: Traditional psychometric methods. *Alzheimer's & Dementia*, 9(1):S4–S9, 2013.

[75] Jeremy Hobart, Stefan Cano, Holly Posner, Ola Selnes, Yaakov Stern, Ronald Thomas, and John Zajicek. Putting the Alzheimer's cognitive test to the test II: Rasch Measurement Theory. *Alzheimer's & Dementia*, 9(1):S10–S20, 2013.

[76] Paul W Holland and Howard Wainer. *Differential item functioning*. Routledge, 2012.

[77] John L Horn. A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30(2):179–185, 1965.

[78] Li-tze Hu and Peter M Bentler. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6(1):1–55, 1999.

[79] Xue Hua, Suh Lee, Igor Yanovsky, Alex D Leow, Yi-Yu Chou, April J Ho, Boris Gutman, Arthur W Toga, Clifford R Jack, Matt A Bernstein, et al. Optimizing power to track brain degeneration in Alzheimer's disease and mild cognitive impairment with tensor-based morphometry: an ADNI study of 515 subjects. *Neuroimage*, 48(4):668–681, 2009.

[80] M Ibrahim, N John, M Kabuka, and A Younis. Hidden Markov models-based 3D MRI brain segmentation. *Image and Vision Computing*, 24(10):1065–1079, 2006.

[81] Clifford R Jack, David S Knopman, William J Jagust, Ronald C Petersen, Michael W Weiner, Paul S Aisen, Leslie M Shaw, Prashanthi Vemuri, Heather J Wiste, Stephen D Weigand, et al. Tracking pathophysiological processes in Alzheimer's disease: an updated hypothetical model of dynamic biomarkers. *The Lancet Neurology*, 12(2):207–216, 2013.

[82] Clifford R Jack, David S Knopman, William J Jagust, Leslie M Shaw, Paul S Aisen, Michael W Weiner, Ronald C Petersen, and John Q Trojanowski. Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade. *The Lancet Neurology*, 9(1):119–128, 2010.

[83] Clifford R Jack, Val J Lowe, Stephen D Weigand, Heather J Wiste, Matthew L Senjem, David S Knopman, Maria M Shiung, Jeffrey L Gunter, Bradley F Boeve, Bradley J Kemp, et al. Serial PIB and MRI in normal, mild cognitive impairment and Alzheimer's disease: implications for sequence of pathological events in Alzheimer's disease. *Brain*, page awp062, 2009.

[84] Clifford R Jack, Ronald C Petersen, Peter C O'Brien, and Eric G Tangalos. MR-based hippocampal volumetry in the diagnosis of Alzheimer's disease. *Neurology*, 42(1):183–183, 1992.

[85] Mark Jenkinson and Stephen Smith. A global optimisation method for robust affine registration of brain images. *Medical Image Analysis*, 5(2):143–156, 2001.

[86] Juan Ramón Jiménez-Alaniz, Verónica Medina-Bañuelos, and Oscar Yáñez-Suárez. Data-driven brain MRI segmentation supported on edge confidence and a priori tissue information. *IEEE Transactions on Medical Imaging*, 25(1):74–83, 2006.

[87] Henry F Kaiser. The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, 1960.

[88] Taehoon Kang and Troy T Chen. Performance of the Generalized S-X2 Item Fit Index for Polytomous IRT Models. *Journal of Educational Measurement*, 45(4):391–406, 2008.

[89] Yong S Kim, Donald W Nibbelink, and John E Overall. Factor structure and reliability of the Alzheimer's Disease Assessment Scale in a multicenter trial with linopirdine. *Journal of Geriatric Psychiatry and Neurology*, 7(2):74–83, 1994.

[90] Gina R Kuperberg, Matthew R Broome, Philip K McGuire, Anthony S David, Marianna Eddy, Fujiro Ozawa, Donald Goff, W Caroline West, Steven CR Williams, Andre JW van der Kouwe, et al. Regionally localized thinning of the cerebral cortex in schizophrenia. *Archives of General Psychiatry*, 60(9):878–888, 2003.

[91] EL Lehmann. Model Specification: The Views of Fisher and Neyman, and Later Developments. In *Selected Works of EL Lehmann*, pages 955–963. Springer, 2012.

[92] Jason P. Lerch, Jens Pruessner, Alex P. Zijdenbos, D. Louis Collins, Stefan J. Teipel, Harald Hampel, and Alan C. Evans. Automated cortical thickness measurements from MRI can accurately separate Alzheimer's patients from normal elderly controls. *Neurobiology of Aging*, 29(1):23–30, 2008.

[93] Kelvin K Leung, Jonathan W Bartlett, Josephine Barnes, Emily N Manning, Sebastien Ourselin, Nick C Fox, et al. Cerebral atrophy in

mild cognitive impairment and Alzheimer disease Rates and acceleration. *Neurology*, 80(7):648–654, 2013.

[94] Kelvin K Leung, Kai-Kai Shen, Josephine Barnes, Gerard R Ridgway, Matthew J Clarkson, Jurgen Fripp, Olivier Salvado, Fabrice Meriaudeau, Nick C Fox, Pierrick Bourgeat, et al. Increasing power to predict mild cognitive impairment conversion to Alzheimer's disease using hippocampal atrophy rate and statistical shape models. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 125–132. Springer, 2010.

[95] Chunming Li, Rui Huang, Zhaohua Ding, Chris Gatenby, Dimitris Metaxas, and John Gore. A variational level set approach to segmentation and bias correction of images with intensity inhomogeneity. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 1083–1091. Springer, 2008.

[96] Chunming Li, Chiu-Yen Kao, John C Gore, and Zhaohua Ding. Minimization of region-scalable fitting energy for image segmentation. *IEEE Transactions on Image Processing*, 17(10):1940–1949, 2008.

[97] Chunming Li, Chenyang Xu, Changfeng Gui, and Martin D Fox. Distance regularized level set evolution and its application to image segmentation. *IEEE Transactions on Image Processing*, 19(12):3243–3254, 2010.

[98] Hua Li, Anthony Yezzi, and Laurent D Cohen. Fast 3D brain segmentation using dual-front active contours with optional user-interaction. In *Computer Vision for Biomedical Image Applications*, pages 335–345. Springer, 2005.

[99] Jun Li, Hong Mei Wu, Rongle L Zhou, Guan Jian Liu, and Bi Rong Dong. Huperzine A for Alzheimer's disease. *The Cochrane Library*, 2008.

[100] Lei Lin, Daniel Garcia-Lorenzo, Chong Li, Tianzi Jiang, and Christian Barillot. Adaptive pixon represented segmentation (APRS) for 3D MR brain images based on mean shift and Markov random fields. *Pattern Recognition Letters*, 32(7):1036–1043, 2011.

[101] Frederic M Lord. *Applications of item response theory to practical testing problems*. Routledge, 1980.

[102] Clement Loy and Lon Schneider. Galantamine for Alzheimer's disease and mild cognitive impairment. *The Cochrane Library*, 2006.

[103] J Steve Marron and Matt P Wand. Exact mean integrated squared error. *The Annals of Statistics*, 20(2):712–736, 1992.

[104] José L Marroquín, Baba C Vemuri, Salvador Botello, E Calderon, and Antonio Fernandez-Bouzas. An accurate and efficient Bayesian method for automatic segmentation of brain MRI. *IEEE Transactions on Medical Imaging*, 21(8):934–945, 2002.

[105] Arnaldo Mayer and Hayit Greenspan. An adaptive mean-shift framework for MRI brain segmentation. *IEEE Transactions on Medical Imaging*, 28(8):1238–1250, 2009.

[106] John Mazziotta, Arthur Toga, Alan Evans, Peter Fox, Jack Lancaster, Karl Zilles, Roger Woods, Tomas Paus, Gregory Simpson, Bruce Pike, et al. A probabilistic atlas and reference system for the human brain: International Consortium for Brain Mapping (ICBM). *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences*, 356(1412):1293–1322, 2001.

[107] Guy McKhann, David Drachman, Marshall Folstein, Robert Katzman, Donald Price, and Emanuel M. Stadlan. Clinical diagnosis of Alzheimer's disease Report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. *Neurology*, 34(7):939–939, 1984.

[108] Guy M. McKhann, David S. Knopman, Howard Chertkow, Bradley T. Hyman, Clifford R. Jack, Claudia H. Kawas, William E. Klunk, Walter J. Koroshetz, Jennifer J. Manly, Richard Mayeux, and others. The diagnosis of dementia due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & Dementia*, 7(3):263–269, 2011.

[109] Andrea Mechelli, Glyn W. Humphreys, Kate Mayall, Andrew Olson, and Cathy J. Price. Differential effects of word length and visual contrast in the fusiform and lingual gyri during. *Proceedings of the Royal Society of London, Series B: Biological Sciences*, 267(1455):1909–1913, 2000.

[110] S. Minoshima, B. Giordani, S. Berent, K. A. Frey, N. L. Foster, and D. E. Kuhl. Metabolic reduction in the posterior cingulate cortex in very early Alzheimer's disease. *Annals of Neurology*, 42(1):85–94, July 1997.

[111] Alex J Mitchell and Mojtaba Shiri-Feshki. Rate of progression of mild cognitive impairment to dementia–meta-analysis of 41 robust inception cohort studies. *Acta Psychiatrica Scandinavica*, 119(4):252–265, 2009.

[112] Richard C Mohs, David Knopman, Ronald C Petersen, Steven H Ferris, Chris Ernesto, Michael Grundman, Mary Sano, Linas Bieliauskas, David Geldmacher, Chris Clark, et al. Development of cognitive instruments for use in clinical trials of antidementia drugs: additions to the Alzheimer's Disease Assessment Scale that broaden its scope. *Alzheimer Disease & Associated Disorders*, 11:13–21, 1997.

[113] EC Mormino, JT Kluth, CM Madison, GD Rabinovici, SL Baker, BL Miller, RA Koeppe, CA Mathis, MW Weiner, WJ Jagust, et al. Episodic memory loss is related to hippocampal-mediated β-amyloid deposition in elderly subjects. *Brain*, 132(5):1310–1323, 2009.

[114] Margaret A Naeser, Paula I Martin, Hugo Theoret, Masahito Kobayashi, Felipe Fregni, Marjorie Nicholas, Jose M Tormos, Megan S Steven, Errol H Baker, and Alvaro Pascual-Leone. TMS suppression of right pars triangularis, but not pars opercularis, improves naming in aphasia. *Brain and Language*, 119(3):206–213, 2011.

[115] Sean M Nestor, Raul Rupsingh, Michael Borrie, Matthew Smith, Vittorio Accomazzi, Jennie L Wells, Jennifer Fogarty, Robert Bartha, Alzheimer's Disease Neuroimaging Initiative, et al. Ventricular enlargement as a possible measure of Alzheimer's disease progression validated using the Alzheimer's disease neuroimaging initiative database. *Brain*, 131(9):2443–2454, 2008.

[116] Kristy A Nielson, Brian J Cummings, and Carl W Cotman. Constructional apraxia in Alzheimer's disease correlates with neuritic neuropathology in occipital cortex. *Brain Research*, 741(1):284–293, 1996.

[117] Jason T Olin and Lon S Schneider. Assessing response to tacrine using the factor analytic structure of the Alzheimer's disease assessment scale (ADAS) cognitive subscale. *International Journal of Geriatric Psychiatry*, 10(9):753–756, 1995.

[118] Maria Orlando and David Thissen. Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement*, 24(1):50–64, 2000.

[119] Maria Orlando and David Thissen. Further investigation of the performance of S-X2: An item fit index for use with dichotomous item response theory models. *Applied Psychological Measurement*, 27(4):289–298, 2003.

[120] BU Park, Seok-Oh Jeong, MC Jones, and Kee-Hoon Kang. Adaptive variable location kernel density estimators with good performance at boundaries. *Journal of Nonparametric Statistics*, 15(1):61–75, 2003.

[121] Xavier Pennec, Pascal Cachier, and Nicholas Ayache. Understanding the "demon's algorithm": 3D non-rigid registration by gradient descent. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI*, pages 597–605. Springer, 1999.

[122] Cecilia M Persson, Åsa K Wallin, Sten Levander, and Lennart Minthon. Changes in cognitive domains during three years in patients with Alzheimer's disease treated with donepezil. *BMC Neurology*, 9(1):7, 2009.

[123] Ronald C Petersen. Mild cognitive impairment as a diagnostic entity. *Journal of Internal Medicine*, 256(3):183–194, 2004.

[124] Ronald C Petersen, Ronald G Thomas, Michael Grundman, David Bennett, Rachelle Doody, Steven Ferris, Douglas Galasko, Shelia Jin, Jeffrey Kaye, Allan Levey, et al. Vitamin E and donepezil for the treatment of mild cognitive impairment. *New England Journal of Medicine*, 352(23):2379–2388, 2005.

[125] Ronald C Petersen and John Q Trojanowski. Use of alzheimer disease biomarkers: potentially yes for clinical trials but not yet for clinical practice. *Journal of American Medical Association*, 302(4):436–437, 2009.

[126] Marcel Prastawa, John H Gilmore, Weili Lin, and Guido Gerig. Automatic segmentation of MR images of the developing newborn brain. *Medical image analysis*, 9(5):457–466, 2005.

[127] Jerry L Prince, Dzung Pham, and Qing Tan. Optimization of MR pulse sequences for Bayesian image segmentation. *Medical Physics*, 22:1651, 1995.

[128] Olivier Querbes, Florent Aubry, Jeremie Pariente, Jean-Albert Lotterie, Jean-Francois Demonet, Veronique Duret, Michele Puel, Isabelle Berry, Jean-Claude Fort, Pierre Celsis, et al. Early diagnosis of Alzheimer's disease using cortical thickness: impact of cognitive reserve. *Brain*, 132(8):2036–2047, 2009.

[129] Joseph F Quinn, Rema Raman, Ronald G Thomas, Karin Yurko-Mauro, Edward B Nelson, Christopher Van Dyck, James E Galvin, Jennifer Emond, Clifford R Jack, Michael Weiner, et al. Docosahexaenoic acid supplementation and cognitive decline in Alzheimer disease: a randomized trial. *Journal of American Medical Association*, 304(17):1903–1911, 2010.

[130] MS Rafii, S Walsh, JT Little, K Behan, B Reynolds, C Ward, S Jin, R Thomas, PS Aisen, et al. A phase II trial of huperzine A in mild to moderate Alzheimer disease. *Neurology*, 76(16):1389–1394, 2011.

[131] Nandini Raghavan, Mahesh N Samtani, Michael Farnum, Eric Yang, Gerald Novak, Michael Grundman, Vaibhav Narayan, and Allitia DiBernardo. The ADAS-Cog revisited: Novel composite scales based on ADAS-Cog to improve efficiency in MCI and early AD trials. *Alzheimer's & Dementia*, 9(1):S21–S31, 2013.

[132] Jagath C Rajapakse and Frithjof Kruggel. Segmentation of MR images with intensity inhomogeneities. *Image and Vision Computing*, 16(3):165–180, 1998.

[133] Martin Reuter and Bruce Fischl. Avoiding asymmetry-induced bias in longitudinal image processing. *Neuroimage*, 57(1):19–21, 2011.

[134] Martin Reuter, Nicholas J Schmansky, H Diana Rosas, and Bruce Fischl. Within-subject template estimation for unbiased longitudinal image analysis. *Neuroimage*, 61(4):1402–1418, 2012.

[135] Mariano Rivera, Omar Ocegueda, and Jose L Marroquin. Entropy-controlled quadratic Markov measure field models for efficient image segmentation. *IEEE Transactions on Image Processing*, 16(12):3047–3057, 2007.

[136] David Rivest-Hénault and Mohamed Cheriet. Unsupervised MRI segmentation of brain tissues using a local linear model and level set. *Magnetic Resonance Imaging*, 29(2):243–259, 2011.

[137] Kenneth Rockwood and Serge Gauthier. *Trial designs and outcomes in dementia therapeutic research*. CRC Press, 2005.

[138] SL Rogers, MR Farlow, RS Doody, R Mohs, LT Friedhoff, et al. A 24-week, double-blind, placebo-controlled trial of donepezil in patients with Alzheimer's disease. *Neurology*, 50(1):136–145, 1998.

[139] HD Rosas, AK Liu, S Hersch, M Glessner, RJ Ferrante, DH Salat, A van Der Kouwe, BG Jenkins, AM Dale, and B Fischl. Regional and progressive thinning of the cortical ribbon in Huntington's disease. *Neurology*, 58(5):695–701, 2002.

[140] Wilma G Rosen, Richard C Mohs, and Kenneth L Davis. A new rating scale for Alzheimer's disease. *The American Journal of Psychiatry*, 1984.

[141] Michael Rösler, Ravi Anand, Ana Cicin-Sain, Serge Gauthier, Yves Agid, Peter Dal-Bianco, Hannes B Stähelin, Richard Hartman, Marguirguis Gharabawi, and Tony Bayer. Efficacy and safety of rivastigmine in patients with Alzheimer's disease: international randomised controlled trialCommentary: Another piece of the Alzheimer's jigsaw. *Bmj*, 318(7184):633–640, 1999.

[142] Snehashis Roy, Aaron Carass, Pierre-Louis Bazin, Susan Resnick, and Jerry L Prince. Consistent segmentation using a Rician classifier. *Medical Image Analysis*, 16(2):524–535, 2012.

[143] Andrea Rueda, Oscar Acosta, Michel Couprie, Pierrick Bourgeat, Jurgen Fripp, Nicholas Dowson, Eduardo Romero, and Olivier Salvado. Topology-corrected segmentation and local intensity estimates for improved partial volume classification of brain cortex in MRI. *Journal of Neuroscience Methods*, 188(2):305–315, 2010.

[144] Mert R Sabuncu, Rahul S Desikan, Jorge Sepulcre, Boon Thye T Yeo, Hesheng Liu, Nicholas J Schmansky, Martin Reuter, Michael W Weiner, Randy L Buckner, Reisa A Sperling, et al. The dynamics of cortical and hippocampal atrophy in Alzheimer disease. *Archives of Neurology*, 68(8):1040–1048, 2011.

[145] David H Salat, Randy L Buckner, Abraham Z Snyder, Douglas N Greve, Rahul SR Desikan, Evelina Busa, John C Morris, Anders M Dale, and Bruce Fischl. Thinning of the cerebral cortex in aging. *Cerebral Cortex*, 14(7):721–730, 2004.

[146] Stephen Salloway, S. Ferris, A. Kluger, R. Goldman, T. Griesing, D. Kumar, S. Richardson, and others. Efficacy of donepezil in mild cognitive impairment A randomized placebo-controlled trial. *Neurology*, 63(4):651–657, 2004.

[147] Stephen Salloway, Jacobo Mintzer, Myron F Weiner, and Jeffrey L Cummings. Disease-modifying therapies in Alzheimers disease. *Alzheimer's & Dementia*, 4(2):65–79, 2008.

[148] Mahesh N Samtani, Michael Farnum, Victor Lobanov, Eric Yang, Nandini Raghavan, Allitia DiBernardo, and Vaibhav Narayan. An improved model for disease progression in patients from the Alzheimer's disease neuroimaging initiative. *The Journal of Clinical Pharmacology*, 52(5):629–644, 2012.

[149] Mahesh N Samtani, Nandini Raghavan, Yingqi Shi, Gerald Novak, Michael Farnum, Victor Lobanov, Tim Schultz, Eric Yang, Allitia DiBernardo, and Vaibhav A Narayan. Disease progression model in subjects with mild cognitive impairment from the Alzheimer's disease neuroimaging initiative: CSF biomarkers predict population subtypes. *British Journal of Clinical Pharmacology*, 75(1):146–161, 2013.

[150] M Sano, KL Bell, D Galasko, JE Galvin, RG Thomas, CH van Dyck, and PS Aisen. A randomized, double-blind, placebo-controlled trial of simvastatin to treat Alzheimer disease. *Neurology*, 77(6):556–563, 2011.

[151] K Schafer, S De Santi, and L S Schneider. Errors in ADAS-cog administration and scoring may undermine clinical trials results. *Current Alzheimer Research*, 8(4):373–376, 2011.

[152] PH Scheltens, D Leys, F Barkhof, D Huglo, HC Weinstein, P Vermersch, M Kuiper, M Steinling, E Ch Wolters, and J Valk. Atrophy of medial

temporal lobes on MRI in "probable" Alzheimer's disease and normal ageing: diagnostic value and neuropsychological correlates. *Journal of Neurology, Neurosurgery & Psychiatry*, 55(10):967–972, 1992.

[153] Terena Searcey, Linda Bierer, and Kenneth L Davis. A longitudinal study of Alzheimer's disease: measurement, rate, and predictors of cognitive deterioration. *American Journal of Psychiatry*, 1:51, 1994.

[154] Feng Shi, Bing Liu, Yuan Zhou, Chunshui Yu, and Tianzi Jiang. Hippocampal volume and asymmetry in mild cognitive impairment and Alzheimer's disease: Meta-analyses of MRI studies. *Hippocampus*, 19(11):1055–1064, 2009.

[155] Il-Seon Shin, Michele Carter, Donna Masterman, Lynn Fairbanks, and Jeffrey L Cummings. Neuropsychiatric symptoms and quality of life in alzheimer disease. *The American Journal of Geriatric Psychiatry*, 13(6):469–474, 2005.

[156] Jeffrey Shuren. Insula and aphasia. *Journal of Neurology*, 240(4):216–218, 1993.

[157] Mohammed Yakoob Siyal and Lin Yu. An intelligent modified fuzzy c-means based algorithm for bias estimation and segmentation of brain MRI. *Pattern Recognition Letters*, 26(13):2052–2062, 2005.

[158] Jeannine Skinner, Janessa O Carvalho, Guy G Potter, April Thames, Elizabeth Zelinski, Paul K Crane, Laura E Gibbons, Alzheimer's Dis-

ease Neuroimaging Initiative, et al. The Alzheimer's Disease Assessment Scale-Cognitive-Plus (ADAS-Cog-Plus): an expansion of the ADAS-Cog to improve responsiveness in MCI. *Brain Imaging and Behavior*, 6(4):489–501, 2012.

[159] Michael E Smith, June M Stapleton, and Eric Halgren. Human medial temporal lobe potentials evoked in memory and language tasks. *Electroencephalography and Clinical Neurophysiology*, 63(2):145–159, 1986.

[160] Stephen M Smith. Fast robust automated brain extraction. *Human Brain Mapping*, 17(3):143–155, 2002.

[161] Gabriela Spulber, Eini Niskanen, Stuart MacDonald, Oded Smilovici, Kewei Chen, Eric M Reiman, Anne M Jauhiainen, Merja Hallikainen, Susanna Tervo, Lars-Olof Wahlund, et al. Whole brain atrophy rate predicts progression from MCI to Alzheimer's disease. *Neurobiology of Aging*, 31(9):1601–1605, 2010.

[162] Larry R. Squire and Stuart Zola-Morgan. The medial temporal lobe memory system. *Science*, 253(5026):1380–1386, 1991.

[163] P. R. Szeszko, R. M. Bilder, T. Lencz, M. Ashtari, R. S. Goldman, G. Reiter, H. Wu, and J. A. Lieberman. Reduced anterior cingulate gyrus volume correlates with executive dysfunction in men with first-episode schizophrenia. *Schizophrenia Research*, 43(2-3):97–108, June 2000.

[164] Sheela Talwalker, John E Overall, Mandyam K Srirama, and Stephen I Gracon. Cardinal features of cognitive dysfunction in Alzheimer's disease: a factor-analytic study of the Alzheimer's Disease Assessment Scale. *Journal of Geriatric Psychiatry and Neurology*, 9(1):39–46, 1996.

[165] Pierre N Tariot, Lon S Schneider, Jeffrey Cummings, Ronald G Thomas, Rema Raman, Laura J Jakimovich, Rebekah Loy, Barbara Bartocci, Adam Fleisher, M Saleem Ismail, et al. Chronic divalproex sodium to attenuate agitation and clinical progression of Alzheimer disease. *Archives of General Psychiatry*, 68(8):853–861, 2011.

[166] George R Terrell and David W Scott. Variable kernel density estimation. *The Annals of Statistics*, pages 1236–1265, 1992.

[167] Robert D Terry, Eliezer Masliah, David P Salmon, Nelson Butters, Richard DeTeresa, Robert Hill, Lawrence A Hansen, and Robert Katzman. Physical basis of cognitive alterations in Alzheimer's disease: synapse loss is the major correlate of cognitive impairment. *Annals of Neurology*, 30(4):572–580, 1991.

[168] J-P Thirion. Image matching as a diffusion process: an analogy with Maxwell's demons. *Medical Image Analysis*, 2(3):243–260, 1998.

[169] Paul M Thompson, Kiralee M Hayashi, Greig I de Zubicaray, Andrew L Janke, Stephen E Rose, James Semple, Michael S Hong, David H Herman, David Gravano, David M Doddrell, et al. Mapping hippocampal

and ventricular change in Alzheimer disease. *Neuroimage*, 22(4):1754–1766, 2004.

[170] Jussi Tohka, Ivo D Dinov, David W Shattuck, and Arthur W Toga. Brain MRI tissue classification based on local Markov random fields. *Magnetic Resonance Imaging*, 28(4):557–573, 2010.

[171] Ledyard R Tucker and Charles Lewis. A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, 38(1):1–10, 1973.

[172] Sebastian Ueckert, Elodie L Plan, Kaori Ito, Mats O Karlsson, Brian Corrigan, Andrew C Hooker, Alzheimers Disease Neuroimaging Initiative, et al. Improved utilization of ADAS-cog assessment data through Item Response Theory based pharmacometric modeling. *Pharmaceutical Research*, 31(8):2152–2165, 2014.

[173] Koen Van Leemput, Frederik Maes, Dirk Vandermeulen, and Paul Suetens. Automated model-based bias field correction of MR images of the brain. *IEEE Transactions on Medical Imaging*, 18(10):885–896, 1999.

[174] Koen Van Leemput, Frederik Maes, Dirk Vandermeulen, and Paul Suetens. Automated model-based tissue classification of MR images of the brain. *IEEE Transactions on Medical Imaging*, 18(10):897–908, 1999.

[175] Baba C Vemuri, J Ye, Y Chen, and Christiana Morison Leonard. Image registration via level-set motion: Applications to atlas-based segmentation. *Medical Image Analysis*, 7(1):1–20, 2003.

[176] P Vemuri, HJ Wiste, SD Weigand, LM Shaw, JQ Trojanowski, MW Weiner, DS Knopman, RC Petersen, CR Jack, et al. MRI and CSF biomarkers in normal, MCI, and AD subjects predicting future clinical change. *Neurology*, 73(4):294–301, 2009.

[177] Prashanthi Vemuri and Clifford R. Jack Jr. Role of structural MRI in Alzheimer's disease. *Alzheimers Research Therapy*, 2(4):23, 2010.

[178] Prashanthi Vemuri, Jennifer L Whitwell, Kejal Kantarci, Keith A Josephs, Joseph E Parisi, Maria S Shiung, David S Knopman, Bradley F Boeve, Ronald C Petersen, Dennis W Dickson, et al. Antemortem MRI based STructural Abnormality iNDex (STAND)-scores correlate with postmortem Braak neurofibrillary tangle stage. *Neuroimage*, 42(2):559–567, 2008.

[179] N Verma, MC Cowperthwaite, and MK Markey. Variational level set approach for automatic correction of multiplicative and additive intensity inhomogeneities in brain MR Images. In *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 98–101. IEEE, 2012.

[180] Nishant Verma and Mia K Markey. Item response analysis of Alzheimer's disease assessment scale. In *36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 2476–2479. IEEE, 2014.

[181] Nishant Verma and Mia K Markey. Psychometric Analysis of Alzheimer's Disease Assessment Scale. Annual Meeting of the Biomedical Engineering Society, 2014.

[182] Nishant Verma, Gautam S Muralidhar, Alan C Bovik, Matthew C Cowperthwaite, Mark G Burnett, and Mia K Markey. Three-dimensional brain magnetic resonance imaging segmentation via knowledge-driven decision theory. *Journal of Medical Imaging*, 1(3):034001–034015, 2014.

[183] Nishant Verma, Gautam S Muralidhar, Alan C Bovik, Matthew C Cowperthwaite, and Mia K Markey. Model-driven, probabilistic level set based segmentation of magnetic resonance images of the brain. In *Proc. Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 2821–2824, 2011.

[184] Luminita A Vese and Tony F Chan. A multiphase level set framework for image segmentation using the Mumford and Shah model. *International Journal of Computer Vision*, 50(3):271–293, 2002.

[185] Bai-song Wang, Hao Wang, Zhao-hui Wei, Yan-yan Song, Lu Zhang, and Hong-zhuan Chen. Efficacy and safety of natural acetylcholinesterase inhibitor huperzine A in the treatment of Alzheimer's disease: an updated meta-analysis. *Journal of Neural Transmission*, 116(4):457–465, 2009.

[186] Daniel Weintraub, Monique Somogyi, and Xiangyi Meng. Rivastigmine in Alzheimer's disease and Parkinson's disease dementia: an ADAS-cog

factor analysis. *American Journal of Alzheimer's disease and other Dementias*, page 1533317511424892, 2011.

[187] Williams M Wells III, W Eric L Grimson, Ron Kikinis, and Ferenc A Jolesz. Adaptive segmentation of MRI data. *IEEE Transactions on Medical Imaging*, 15(4):429–442, 1996.

[188] Michael Wels, Yefeng Zheng, Martin Huber, Joachim Hornegger, and Dorin Comaniciu. A discriminative model-constrained EM approach to 3D MRI brain tissue classification and intensity non-uniformity correction. *Physics in Medicine and Biology*, 56(11):3269, 2011.

[189] Eric Westman, Carlos Aguilar, J-Sebastian Muehlboeck, and Andrew Simmons. Regional magnetic resonance imaging measures for multivariate analysis in Alzheimer's disease and mild cognitive impairment. *Brain Topography*, 26(1):9–23, 2013.

[190] Kylie J Wheaton, James C Thompson, Ari Syngeniotis, David F Abbott, and Aina Puce. Viewing the motion of human body parts activates different regions of premotor, temporal, and parietal cortex. *Neuroimage*, 22(1):277–288, 2004.

[191] Jennifer L Whitwell, Dennis W Dickson, Melissa E Murray, Stephen D Weigand, Nirubol Tosakulwong, Matthew L Senjem, David S Knopman, Bradley F Boeve, Joseph E Parisi, Ronald C Petersen, et al. Neuroimaging correlates of pathologically defined subtypes of Alzheimer's disease: a case-control study. *The Lancet Neurology*, 11(10):868–877, 2012.

[192] JL Whitwell, KA Josephs, ME Murray, K Kantarci, SA Przybelski, SD Weigand, P Vemuri, ML Senjem, JE Parisi, DS Knopman, et al. MRI correlates of neurofibrillary tangle pathology at autopsy A voxel-based morphometry study. *Neurology*, 71(10):743–749, 2008.

[193] April L Zenisky, Ronald K Hambleton, and Stephen G Sireci. Effects of Local Item Dependence on the Validity of IRT Item, Test, and Ability Statistics. MCAT Monograph. 2003.

[194] Bo Zhang and Clement A Stone. Evaluating item fit for multidimensional item response models. *Educational and Psychological Measurement*, 68(2):181–196, 2008.

[195] Yongyue Zhang, Michael Brady, and Stephen Smith. Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *IEEE Transactions on Medical Imaging*, 20(1):45–57, 2001.

[196] Bochuan Zheng and Zhang Yi. A new method based on the CLM of the LV RNN for brain MR image segmentation. *Digital Signal Processing*, 22(3):497–505, 2012.

Vita

Nishant Verma was born in Pune, Maharashtra, India. He received the Bachelor of Technology degree in Biological Sciences and Bioengineering from the Indian Institute of Technology (IIT) at Kanpur in 2010. He was awarded the Tata Consultancy Services Award and the Proficiency Medal for his research work at IIT Kanpur. Nishant joined The University of Texas (UT) at Austin in 2010 to begin his Ph.D. studies in Biomedical Engineering under the supervision of Dr. Mia K. Markey. In spring 2014, he received the Master of Science degree in Biomedical Engineering from UT Austin. Nishant has been a recipient of the George J. Heuer, Jr. Ph.D. Endowed Graduate Fellowship and the Carol Lewis Heideman Endowed Presidential Fellowship in Biomedical Engineering at UT Austin. During his studies at UT-Austin, Nishant also held an affiliation as a graduate student researcher with the NeuroTexas Research Institute at St. David's HealthCare in Austin.

Email address: nishant3115@gmail.com

This dissertation was typeset with LaTeX[†] by the author.

[†]LaTeX is a document preparation system developed by Leslie Lamport as a special version of Donald Knuth's TeX Program.