UNIVERSITY OF CALIFORNIA SAN DIEGO

Doubly Robust Imputation for Longitudinal Data with Monotone Dropouts: Applications in Alzheimer's Randomized Trials

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

in

Biostatistics

by

Yuqi Qiu

Committee in charge:

Professor Karen S. Messer, Chair
Professor Howard H. Feldman, Co-Chair
Professor Steven D. Edland
Professor David P. Salmon
Professor Xin Tu

2021

The dissertation of Yuqi Qiu is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2021

DEDICATION

To my wife Siyun: for your 13-years accompany and selfless love.

To my parents Mr. Qiu and Mrs. Zu: for always promoting me to be stronger and better.

To my daughter to be born: nice to meet you and we'll see you in October.

EPIGRAPH

Life is like a double-crostic;

we can do far more than we know.

—John Tukey

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

This dissertation represents not only my development in biostatistical methods, but also a milestone for my five years of Ph.D study in the division of Biostatistics at University of California San Diego (UCSD). It has been an enjoyable journey to obtain doctoral training in this wonderful program and work with so many intelligent people. I have been enlightened by dozens of remarkable individuals at UCSD, whom I sincerely wish to acknowledge.

First and foremost, it is a genuine pleasure to express my deep sense of thanks and gratitude to my mentor and committee chair, Professor Karen Messer. She has been supportive since I first entered this program in 2016. Her enthusiasm for research and work has deeply inspired me. As a professor and an experienced biostatistician, she helped me sharpen my abilities in critical thinking and skills in statistical analysis, and she taught me how to collaborate with researchers from other fields effectively to complete clinical trials. As an advisor and my mentor, she always guided me to discover my research interests and encouraged me to explore novel methods. She spent countless hours proofreading my research papers, providing timely feedback, and brainstorming approaches to improve my research work. She has always been there to navigate me towards my educational and career goals. Throughout my doctoral studies, we had numerous meetings where we spent hours discussing various topics ranging from my research work and career plans to daily issues. No matter how busy she was, she would always find time to meet with me and help me overcome the difficulties I encountered. I am indebted for her timely guidance, constant encouragement, and tremendous support and opportunities she provided throughout my Ph.D study.

I owe a deep sense of gratitude to my committee co-chair Professor Howard Feldman, director of the Alzheimer's Disease Cooperative Study (ADCS), for his keen interest in me at every stage of my research. His wisdom, meticulous scrutiny, and insightful suggestions have guided me to complete this dissertation. His suggestions and comments on both my research as well as on my career have been absolutely invaluable. Professor Howard Feldman introduced me to the research field of clinical trials on Alzheimer's disease, which also motivated me to write this

dissertation. He directed my first first-author publication (Chapter 2), provided valuable feedback to my first statistical paper (Chapter 3), and made insightful comments on my second statistical paper (Chapter 4). His scholarly advice and critical thinking inspired me to be a better researcher.

Moreover, I would also like to thank my committee members Professors Steven Edland, David Salmon, and Xin Tu, for their continuous support and encouragement throughout my study period. Professor Steven Edland, who was my academic advisor, provided intelligent comments on my dissertation, especially in Chapter 3, and was very generous in sharing his experience on academic life and beyond. Professor David Salmon, who is an excellent and knowledgeable collaborator from the Department of Neuroscience, guided me to complete my first first-author publication (Chapter 2), which helped me shape this dissertation. Professor Xin Tu, who taught the causal inference class in our division, stimulated my interests in causal inference and missing data, which are the essential statistical methods that I will be continuously benefiting from in my future research work.

It was my great privilege to work for the Biostatistics Shared Resource (BSR) at Moores Cancer Center (MCC) for the past five years. It was a delightful and memorable period in my Ph.D journey. I am grateful to Professors Loki Natarajan, Lin Liu, Xinlian Zhang, and Jingjing Zou for their extensive personal and professional guidance. I would also like to thank Dr. Emily Pittman, Minya Pu, Jing Zhang, and Ruifeng Chen for their constant help over these years.

It was my great pleasure to work with ADCS. I would like to thank Dr. Diane Jacobs for being extremely helpful and being an outstanding collaborator. I would also like to thank ADCS colleagues Carol Evans, Tatiana Herold, Jennifer Mason, and Carolyn Revta for being very organized and helpful throughout our collaborations.

Furthermore, I wanted to thank our outstanding Biostatistics program at UCSD. I am incredibly grateful for my experience in this program. The USCD Ph.D program in biostatistics taught me not only statistical knowledge and programming skills, but also how to apply statistics as a collaborator in scientific projects. These skills will form the foundation of my future scientific

career. As a member of the first cohort of our doctoral program, I am glad that I was able to witness our program expand and grow over the past five years. I would like to thank Professors Armin Schwartzman, Florin Vaida, Wesley Thompson, Sonia Jain, Charles Berry, and Lily Xu for teaching me and helping me practice my statistical knowledge. I also wanted to express my great appreciation to Stella Tripp, Melody Bazyar, and Sarah Dauchez. This program would not have developed so fast without their contributions.

Finally, I would like to dedicate this dissertation to my wife Siyun Chen and my parents. I could not have accomplished this without their love and support.

Chapter 2, in full, has been published and may be found as "Qiu, Yuqi; Jacobs, Diane M.; Messer, Karen S.; Salmon, David P.; Feldman, Howard H. *Cognitive Heterogeneity in Probable Alzheimer's Disease: Clinical and Neuropathological Features*, Neurology, 93 (8), e778-e790, 2019". The dissertation author was the primary author on this paper.

Chapter 3, in full, has been submitted for publication as "Qiu, Yuqi; Messer, Karen S. *Doubly robust imputation in longitudinal studies, with an application to an Alzheimer's clinical trial*, submitted to Annals of Applied Statistics". The dissertation author was the primary author on this paper.

Chapter 4, in full, has been prepared for submission for publication as "Qiu, Yuqi; Feldman, Howard H.; Messer, Karen S. *Doubly Robust Imputations for Randomized Trials with Monotone Dropout under Missing not at Random: Applications in Alzheimer's Trials*". The dissertation author was the primary author on this paper.

Data collection and sharing for projects in chapter 3 and 4 were obtained from the Alzheimer's Disease Cooperative Study (ADCS), funded by the National Institutes of Health Grant U19 AG010483.

VITA

2014	Bachelor of Economics. Beijing Normal University, Beijing, China
2016	Master of Science in Biostatistics. Case Western Reserve University, Cleveland, U.S
2016-2021	Graduate Student Researcher. University of California San Diego Health
2021	Doctor of Philosophy in Biostatistics. University of California San Diego, U.S

PUBLICATIONS

Y. Qiu, D. Jacobs, K. Messer, D. Salmon, H. Feldman. *Cognitive Heterogeneity in Probable Alzheimer's Disease: Clinical and Neuropathological Features* Neurology, 93 (8), e778-e790, 2019.

Y. Qiu, T. Arbogast, H. Li, S. Tang, E. Richardson, O. Hong, T. Pang, S. Cho, Simons VIP Consortium, C. Corsello, C. Deutsch, C. Chevalier, S.M. Lorenzo, E. Davis, Y. Herault, N. Katsanis, K. Messer, J. Sebat. *Oligogenic effects of 16p11.2 copy number variation on craniofacial development*. Cell Reports, 28 (13), 2019.

Y. Qiu and K. Messer. *Doubly robust imputation in longitudinal studies, with an application to an Alzheimer's clinical trial*. Annals of Applied Statistics, Submitted.

L. Liu, **Y. Qiu**, L. Natarajan, K. Messer. *Imputation and Post-selection Inference in Models with Missing Data: An Application to Colorectal Cancer Surveillance Guidelines*. Annals of Applied Statistics, 13 (3), 1370-1396, 2019.

S. Banks, **Y. Qiu**, C.C. Fan, A. Dale, J. Zou, K. Messer, B. Askew, H. Feldman. *Enriching the Design of Alzheimer Disease Clinical Trials: Application of the Polygenic Hazard Score and Composite Outcome Measures*. Alzheimer's & Dementia: Translational Research & Clinical Interventions, 6 (1), e12071, 2020.

B. Panuganti, **Y. Qiu**, B. Messing, G. Lee, C. Fakhry, R. Blanco, P. Ha, K. Messer, J.A. Califano. *Effects of a Comprehensive Performance Improvement Strategy on Postoperative Adverse Events in Head and Neck Surgery*. Otolaryngology - Head and Neck Surgery, 160 (5), 799-809, 2018.

J.M. Lewis, A. Vyas, **Y. Qiu**, K.S. Messer, R. White, M.J. Heller. *Rapid AC Electrokinetic Capture of Exosomes Enables On-Chip Detection of Pancreatic Cancer Biomarkers in Patient Blood Samples*. ACS nano, 12 (4), 3311-3320, 2018.

E. Bayram, R. Yilmaz, **Y. Qiu**, K. Messer, O. Yalap, O. Aydin, H. Ergenc, M. Akbostanci. *No Effect of Bilateral or Unilateral Subthalamic Nucleus Deep Brain Stimulation on Verb and Noun Naming in Parkinson's Disease*. Brain and Language, 212, 104865, 2021.

J.F. Scott, L.M. Das, S. Ahsanuddin, **Y. Qiu**, A.M. Binko, Z.P. Traylor, S.M. Debanne, K.D. Cooper, R. Boxer, K.Q. Lu. *Oral vitamin D rapidly attenuates inflammation from sunburn: an interventional study*. The Journal of Investigative Dermatology, 137 (10), 2078-2086, 2017.

S. Pu, B.O. Oluyede, **Y. Qiu**, D. Linder. *A Generalized Class of Exponentiated Modified Weibull Distribution with Applications*. Journal of Data Science, 14 (4), 585, 2016.

B.O. Oluyede, S. Pu, B. Makubate, **Y. Qiu**. *The Gamma-Weibull-G Family of Distributions with Applications*. Austrian Journal of Statistics, 47 (1), 45-76, 2018.

ABSTRACT OF THE DISSERTATION

Doubly Robust Imputation for Longitudinal Data with Monotone Dropouts: Applications in Alzheimer's Randomized Trials

by

Yuqi Qiu

Doctor of Philosophy in Biostatistics

University of California San Diego, 2021

Professor Karen S. Messer, Chair
Professor Howard H. Feldman, Co-Chair

The objective of this dissertation is to utilize statistical methods to obtain consistent estimates from longitudinal data with monotone dropouts. This dissertation is comprised of three main studies. In the first study, which aims to identify the heterogeneity of cognition profiles in probable Alzheimer's disease (AD) and determine if cognitive profiles are systematically related to the clinical course and neuropathological features of the disease, we explored a comprehensive data set from the National Alzheimer's Coordinating Center (NACC) and successfully classified AD patients into 80% "typical" versus 20% "atypical" profiles, across two independent cohorts

and one subset of subjects with autopsy available. We found that the atypical cognition profile was associated with lower Braak stage at autopsy and slower cognitive decline.

Observing an increasing attrition rate after two years with apparent informative dropout in the first study, and being motivated by the informative dropout that is common in FDA regulated trials for AD, in the second study, we proposed a *doubly robust* imputation approach to adjust for dropout-related bias in longitudinal studies. We illustrated this approach with an application in a prodromal AD trial conducted by Alzheimer's Disease Cooperative Study (ADCS), which is a major center for AD clinical trials. We believe the imputation approach we presented has the advantage of computational simplicity and transparency compared to existing approaches in the literature, and may be suitable for use in FDA-regulated trials and a variety of other applications.

As an extension to this topic, in the third study, we investigated the doubly robust imputation method within a pattern mixture model framework, in order to deal with sensitivity analysis for randomized trials under several missing-not-at-random scenarios. We applied the proposed approach to two ADCS trials with different stages of disease, and compared the approach with other well-known methods to evaluate the performance. This study supports that the doubly robust imputation method is a competitive method for handling longitudinal data with monotone dropout, and may be suitable for use in randomized trials to obtain valid estimates of treatment effect.

Chapter 1

Introduction and Background

In the past decades, the International Council on Harmonization (ICH) and the Food and Drug Administration (FDA) have released a series of regulatory reports that emphasize the importance of applying a correct statistical analysis in randomized trials. Statistical considerations are involved in the whole spectrum of clinical trials, including study design, protocol development, data monitoring and reporting, as well as interim and final data analyses. An explicit statistical analysis plan can enhance the reliability and validity of the results in clinical trials. Statistical methods have been developed covering different scenarios in clinical trials, with the aim to obtain an unbiased estimator of the treatment effect finally. In the last few years, the ICH and FDA held several workshops and published guidance elucidating how to appropriately describe the *estimand* as a novel approach to improve the rigor of statistical analysis in clinical trials [66]. In the context of the guidance, the estimand is particularly referring to the treatment effect associated with a clinical trial objective, rather than the general understanding of what is being estimated. Four essential components are required to be explicitly demonstrated to describe the estimand, which are (1) defining the targeted study population; (2) defining the endpoint of interest; (3) describing the handling of any intercurrent events in details (i.e., events which occur between randomization and final assessments); (4) summarizing the variable of interest at the population level.

Alzheimer's disease (AD) is a progressive disease that destroys memory and other important mental functions, and is the most common neurodegenerative disorder and cause of dementia (60%-80%). An estimated 5.7 million Americans of all ages are living with Alzheimer's dementia in 2018, and by 2050 this number is projected to rise to nearly 14 million; among people age 65 and older, 1 in every 10 have AD in 2018. Numerous studies have investigated AD regarding its cause, diagnostics, and treatment [1, 16, 15, 14, 33, 56, 61]. Unfortunately, Alzheimer's has no current cure. AD clinical trials play a most crucial role in order to develop effective therapies. There are hundreds of ongoing AD trials, but since 2003 no new Alzheimer's drug has been approved. According to the Alzheimer's Association, recruiting and retaining trial participants is now the greatest obstacle to developing the next generation of Alzheimer's treatments. Dropout rates are usually higher than 20% in AD trials over two years and may differ between the treatment and placebo groups.

Such disease-related discontinuation is a common problem in AD trials and many other randomized trials. Discontinuation from the study is a typical class of intercurrent events, and the statistical treatment of study discontinuation is required to be clarified as one of the four core components of estimand. Investigators have developed various approaches to deal with this issue, including several imputation methods. In the primary analysis, missing at random (MAR) is always assumed in current practice; hence a usual way of addressing the dropout is to fit a statistical regression model by maximum likelihood and to assume that the model is correctly specified, in which case it will correctly adjust for the dropouts and give consistent and efficient estimates. However, if the outcome model is misspecified, these model-based estimates would be biased, leading to invalid results. The mixed model for repeated measures (MMRM) and multiple imputation are two current principal methods for estimating the outcome model while accounting for the dropout in randomized trials. An alternative approach, modeling the probability of dropouts and giving weights to the observed data so that it can be corrected to represent the complete data, may be helpful; this method is named inverse probability weighting (IPW) or propensity scores

adjustment. A different approach under semi-parametric theory would be the use of a "doubly robust" estimator. The fundamental idea here is to combine a missingness model (i.e., IPW) and an outcome model (or imputation model) so that the estimator would be consistent even if one of the models is incorrect, and with good efficiency properties if both models are correct. In the sensitivity analysis for the primary outcome of a randomized trial, which has been emphasized by regulatory reports in recent years, the assumption of missing not at random (MNAR) is required to be investigated. Under the MNAR assumption, dropouts are presumed to be related to the unobserved data. A commonly used approach constructs a link function between the distributions of observed data and unobserved data, for example, assuming a mean shift in the unobserved data, then fitting a pattern mixture model (PMM) to model the observed data and missing data jointly. We introduced the PMM framework with details in Chapter 4. This PMM framework can be incorporated with an imputation method such as multiple imputation and doubly robust imputation.

1.1 Aims and Organization of this Dissertation

This dissertation focuses on reviewing and constructing a novel statistical method, namely a doubly robust approach to estimating longitudinal data with monotone dropouts, with several applications in AD clinical trials. Chapter two presents a study to investigate the heterogeneity of cognition in probable AD and study the clinical and neuropathological features in the heterogeneity. This study is conducted with comprehensive data from the National Alzheimer's Coordinating Center (NACC) database. Two independent cohorts and a subset with available pathology data are used for external validations. The attrition rate after two years is increasing, with potentially informative missingness, which motivates us to explore the statistical methods for imputing missing responses in longitudinal data.

Chapter three conducts a study investigating doubly robust (DR) imputation methods in

longitudinal data to adjust for dropout-related bias under MAR. We elucidated that most DR estimators can be written in substitution (plug-in) form and proposed a more straightforward DR form that is easier to interpret and obtain. Simulation studies support the theoretical properties of the estimators and provide comparisons with alternative approaches. We also illustrate the imputation approach using historical data from the Alzheimer's Disease Cooperative Study (ADCS).

Chapter four investigates how to perform sensitivity analysis in randomized trials under MNAR, comparing a new DR imputation method, the more usual multiple imputation method, and a sequential mean imputation method developed by Paik. We extend the three imputation methods within the PMM framework and compared them with simulation studies. Two historical AD randomized trials conducted by ADCS are used to demonstrate applications in both primary analysis and sensitivity analysis.

Chapter five presents the overall discussion and future work.

Chapter 2

Cognitive Heterogeneity in Probable Alzheimer's Disease: Clinical and Neuropathological Features

2.1 Abstract

Objective of this study is to identify heterogeneity in cognitive profiles of patients with probable Alzheimer's disease (AD) who have mild-to-moderate dementia and satisfy inclusion and exclusion criteria for a typical AD clinical trial, and determine if cognitive profiles are systematically related to the clinical course and neuropathological features of the disease.

Neuropsychological test data from patients with mild-to-moderate probable AD (n=4,711) were obtained from the National Alzheimer's Coordinating Center (NACC). Inclusion and exclusion criteria usually used in AD clinical trials were applied. Principal component analysis (PCA) and model-based clustering were used to identify cognitive profiles in a subset of patients with autopsy-verified AD (n=800), and validated in the overall (non-autopsy) sample and an independent cohort with similar test data. Relationships between cognitive profile, clinical characteristics,

and rate of decline were examined using mixed-effects models.

In the autopsy-confirmed sample, 79.6% of patients had a "typical" AD cognitive profile (greater impairment of episodic memory than other cognitive functions), and 20.4% had an "atypical" profile (comparable impairment across cognitive domains). Similar results were obtained in the overall (typical: 79.8%; atypical: 20.2%) and validation (typical: 71.8%; atypical: 28.2%) samples. Atypicality was associated with younger age, male sex, lower probability of APOE e4, less severe global dementia, higher depression scores, lower Braak stage at autopsy, and slower cognitive decline.

In conclusion, we can reliably identify distinct cognitive profiles among clinically-diagnosed probable AD patients that are associated with tangle pathology and with different rates of decline. This may have implications for clinical trials in AD, especially therapies targeting tau.

2.2 Introduction

An important source of variability that may bias the results of therapeutic trials for Alzheimer's dementia (AD) is heterogeneity in the presentation and course of cognitive deficits that are expressed within its clinical syndrome. Although AD is most commonly characterized by initial predominant impairment in learning and memory, variants of AD with primary deficits in language (i.e., logopenic primary progressive aphasia), visuospatial abilities (i.e., posterior cortical atrophy) or executive functions (i.e., executive variant AD) with relative sparing of memory have been identified [10, 12, 45, 59, 60, 65, 3, 13, 36, 73]. The use of best practices for assessment of dementia [24] and standardized clinical criteria for the diagnosis of AD dementia [34, 16, 14] allow many of these variants to be identified, and they are usually excluded from AD clinical trials because they are likely to progress in an atypical fashion and also have predominant deficits that are not sensitively measured by widely-used clinical trial cognitive outcome measures. AD trial cohorts often are further restricted to patients with mild-to-moderate dementia severity to ensure

sufficient range to measure change over time. While these inclusion criteria provide a relatively homogeneous sample, the relative degree of impairment in memory versus other cognitive abilities may still vary in a systematic manner that can be identified and considered when analyzing and interpreting AD clinical trial results. Therefore, the present study investigates cognitive heterogeneity in mildly-to-moderately demented probable AD patients who satisfy inclusion and exclusion criteria of a typical AD clinical trial, and whether cognitive profiles are systematically related to the clinical course and neuropathological features of the disease.

2.3 Methods

2.3.1 Participants

Data from participants in the NIH Alzheimer's Disease Centers program were downloaded from the National Alzheimer's Coordinating Center (NACC) database, as of the September 2017 data freeze. Sample 1 consisted of 4,711 participants diagnosed between 2005 and 2015 who met NINCDS-ADRDA diagnostic criteria for probable AD [32], scored 16-24 (inclusive) on the Mini-Mental State Exam (MMSE), and had neuropsychological test data available. Sample 2 consisted of 692 participants enrolled after March 2015 (when NACC Uniform Data Set (UDS) version 3 was implemented and some study procedures were changed) who met NIA-AA diagnostic criteria for probable AD dementia [34], scored 7-19 (inclusive) on the Montreal Cognitive Assessment (MoCA), and had neuropsychological test data available. A MoCA range of 7-19 is equivalent to an MMSE range of 16-24 [40], and both ranges are indicative of mild to moderate dementia severity. Participants were not included in either sample if they had a clinical diagnosis of primary progressive aphasia (PPA), posterior cortical atrophy, frontotemporal dementia, Parkinson's disease, corticobasal syndrome, dementia with Lewy bodies (DLB), cerebrovascular disease (CVD), prion disease, traumatic brain injury, normal pressure hydrocephalus, or other neurological disease contributing to the cognitive impairment. Clinical data from the UDS visit at which a

participant was initially diagnosed with probable AD dementia were used in analyses.

2.3.2 Procedure

At each approximately annual visit, participants received a standardized dementia evaluation [42] that included medical and family history, physical and neurological examination, neuropsychological testing, functional assessment with the Clinical Dementia Rating (CDR) Scale [41] and the Functional Assessment Questionnaire (FAQ) [47], and assessment of depressive symptomology with the Geriatric Depression Scale (GDS) [76]. A categorical decision of whether or not depression was contributing to cognitive dysfunction was also made based on a clinician's judgement.

Neuropsychological Measures

The NACC neuropsychological test battery administered from 2005 to 2015 (UDS version 2) consisted of the MMSE, and measures of verbal learning and memory (Logical Memory Test (story A only) I and II), attention and executive function (Digit Span Forward and Backward; Trail-Making A and B; Digit Symbol Substitution), and language/semantic memory (30-item Boston Naming Test, Animal Fluency) [75]. In March 2015, the NACC protocol (UDS version 3) replaced several tests with comparable non-proprietary measures and added three additional tests to broaden the scope of the battery. Details of the revised battery and demonstration of its comparability to the original are described elsewhere [40, 74].

Neuropathological Diagnosis

Neuropathological findings were available on 976 participants from Sample 1. Diagnostic classification was made based on Braak staging of neurofibrillary tangles [5] and CERAD scoring of neocortical neuritic plaque density [38]. "Definite AD" was defined as Braak stage III-VI plus moderate or frequent neocortical plaques. With these criteria the UDS clinical diagnosis

of probable AD has been shown to have approximately 71% sensitivity and specificity, and 83% positive predictive value [4]. Other pathological diagnoses such as frontotemporal lobar degeneration (Picks, corticobasal ganglionic degeneration, progressive supranuclear palsy), DLB, and CVD (e.g., infarct, lacune, hemorrhage, microbleed, microinfarct) were made according to standard published criteria.

2.3.3 Statistical Methods

Data from Sample 1 were used to identify potential cognitive subtypes and to investigate their association with clinical features, rate of decline, and pathological findings. Data from Sample 2 were used to investigate whether the subtypes identified from Sample 1 would replicate in an independent sample with potentially different baseline characteristics and similar, but not identical, neuropsychological test measures. Neuropsychological test scores were standardized within each sample and all measures were coded so that a higher test score indicated better performance. Because time-to-completion scores on the Trail-Making Test were truncated (maximum time of 150 sec for Part A and 300 sec for Part B), a rate measure was calculated by dividing the number of correct lines drawn by testing time. Computations used R version 3.4.1

Identifying and validating AD Cognitive Subgroups

Principal Component Analysis (PCA, using the "stats:princomp" function), and model-based clustering (using the "mclust" package [21, 20], allowing for different means and covariance structures with number of clusters determined by Bayesian information criterion [62]) were used to identify characteristic patterns of performance on neuropsychological test scores, within each sample. First, PCA was used to identify the linear combinations, or factor loadings, of the original test scores which best constructed two informative and independent components. Then, these first 2 principal component scores for each participant were used in model-based clustering to identify subgroups of participants with distinct patterns of test performance. Analyses initially

were performed using Sample 1 participants with definite AD pathology, then using all of Sample 1 (with restriction to 2 clusters), and finally repeated again using Sample 2 as an independent validation cohort.

Characterizing AD Cognitive Subgroups

Linear regression was used to explore association between cluster membership (i.e., AD cognitive subtype) and demographic variables, clinical characteristics, global cognitive test scores, and neuropathologic features. AD pathology (Braak stage and CERAD plaque score) was coded as an ordinal factor.

Linear mixed effects models were used to examine the association of AD cognitive subtype with the rate of longitudinal decline, by investigating the interaction between time slope and subtype. Models controlled for baseline cognitive score and APOE genotype, and included a random intercept and slope for each participant. Only the first two years of follow-up data were used in these analyses because of apparent informative censoring after two years (i.e., more severely impaired participants were more likely to discontinue). Similar analyses were used to investigate the stability of the AD cognitive subtype classification over time.

Missing Data

Missing test scores in the UDS dataset are designated as missing due to 1) cognitive problems, 2) other, non-cognitive problems, or 3) not collected at that assessment. For scores missing due to cognitive problems, we imputed the worst possible score for the measure. For scores missing due to non-cognitive reasons, we imputed a score using linear regression with predictors age, education, MMSE or MoCA, CDR sum of boxes and all available neuropsychological measures, computed in the R package "mice" [71].

2.4 Results

2.4.1 Participant Characteristics

The pathologically verified subsample from Sample 1 did not differ from the overall Sample 1 in age, MMSE score, CDR global rating, GDS score, or % APOE e4 genotype, but was more educated and had worse FAQ scores (Table 2.1). Sample 2 did not differ from Sample 1 in CDR global rating, FAQ score, GDS score, sex distribution or % APOE e4 genotype, but was younger and more educated.

2.4.2 Principal Components Analysis of Neuropsychological Tests

The results of the principal component analyses were consistent across Sample 1 (the 'discovery' dataset), the autopsy-confirmed subset of Sample 1 (the 'gold standard' dataset), and Sample 2 (the independent 'validation' dataset). Each PCA identified two independent, mean zero, principal components that together explained 47% to 58% of variance. Factor loadings for each principal component were similar in the three analyses, and were not sensitive to the imputation (see Table 2.2). The first principal component (PC1) had positive factor loadings for all neuropsychological tests, thus reflecting overall cognitive performance or dementia severity (i.e., a positive score for PC1 indicates above the mean on cognitive performance, a negative score indicates below the mean). The second principal component (PC2) had positive factor loadings for neuropsychological tests of episodic memory and naming, and negative factor loadings for non-memory tests; thus, a higher PC2 score reflects relatively better performance on memory-related cognitive tests and relatively worse performance on non-memory-related tests. Therefore, PC2 can be interpreted as a continuous measure that discriminates between "cognitive profiles" within probable AD, independently of cognitive severity (i.e., PC1). This pattern of factor loadings for PC2 was reproduced in all three samples. Results were nearly identical in each sample after excluding the approximately 20% of participants for whom depression was considered to

Table 2.1: Cohort characteristics: Mean (SD)

	Sample 1	Sample 2	Neuropath confirmed AD[a]
Overall N	4,711	692	800
Participant Characteristics			
Age[b]	76.15 (9.45)	74.25 (9.78)	76.98 (10.04)
Sex: Female – N (%)	2680 (56.9)	380 (54.9)	386 (48.2)
Educational attainment (yrs.)[b,c]	14.24 (3.54)	15.42 (3.05)	15.12 (3.12)
APOE e4 Positive – N (%)	2207 (59.1)	270 (61.5)	443 (60.4)
	(N = 3732)	(N = 439)	(N = 733)
Clinical Ratings			
MMSE	20.77 (2.45)	NA	20.53 (2.55)
MOCA	NA	14.07 (3.49)	NA
CDR Global rating	0.97 (0.45)	0.97 (0.48)	1.07 (0.50)
CDR Sum of boxes	5.58 (2.62)	5.61 (2.77)	6.23 (2.81)
Geriatric Depression Scale	2.41 (2.52)	2.35 (2.62)	2.27 (2.29)
FAQc	17.16 (7.73)	17.18 (7.45)	19.63 (7.06)
Depression – N (%)	1010 (21.4)	162 (23.4)	140 (17.5)
Neuropsychological Test Measures			
Paragraph Immediate recall	3.66 (3.0)	5.66 (3.9)	3.37 (3.1)
Paragraph Delayed recall	1.44 (2.4)	2.12 (3.2)	1.32 (2.4)
Benson Figure Delayed recall	NA	1.87 (2.9)	NA
Benson Figure Copy	NA	12.8 (4.4)	NA
Confrontation Naming	19.1 (6.9)	22.9 (6.7)	19.3 (6.9)
Category Fluency	17.0 (7.2)	16.9 (7.0)	15.9 (6.8)
Letter Fluency	NA	18.2 (8.5)	NA
Digit Span Forward	6.84 (2.2)	6.37 (2.2)	6.80 (2.4)
Digit Span Backward	4.31 (1.9)	4.18 (2.0)	4.35 (1.9)
WAIS-R Digit Symbol	21.6 (13.4)	NA	20.2 (13.4)
Trail Making A –Time to Completion	76.4 (41.7)	71.3 (39.9)	83.2 (43.3)
Trail Making A – Correct Lines	21.7 (5.2)	22.4 (5.2)	20.7 (6.5)
Trail Making B – Time to Completion	249.5 (75.3)	248.2 (78.0)	254.2 (72.8)
Trail Making B – Correct Lines	12.2 (9.9)	12.7 (10.5)	11.1 (10.0)

Abbreviations: MMSE = Mini-mental state examination; MOCA = Montreal Cognitive Assessment score; CDR = Clinical Dementia Rating; FAQ = Functional Activities Questionnaire; Depression = Clinician rating of depression as contributing to cognitive impairment.
a. Sample 1 participants with Braak stage III-VI plus moderate or frequent neocortical plaques.
b. Age and education are significantly different between Sample 1 and Sample 2.
c. Education and FAQ scores are significantly different between the pathologically verified subsample from Sample 1 and the overall Sample 1.

Table 2.2: Factor loadings and missing data rates for each sample

Cognitive Domain	Neuropsychological Tests		Sample 1 (N=4,711)			Sample 2 (N=692)			Neuropath confirmed AD (N=800)		
	Cumulative variance explained		56.50%			48.30%			57.70%		
	UDS 2	UDS 3	PC1	PC2	Missing rate	PC1	PC2	Missing rate	PC1	PC2	Missing rate
Memory related domains											
Episodic Memory	Logical Memory Immediate recall	Craft Story 21 Immediate recall	0.23	0.58	0.02	0.3	0.43	0.01	0.25	0.56	0.03
Episodic Memory	Logical Memory Delayed recall	Craft Story 21 Delayed recall	0.08	0.66	0.03	0.13	0.6	0.02	0.08	0.66	0.03
Language/ Semantic Memory	Boston Naming	Multilingual Naming	0.31	0.22	0.02	0.28	0.14	0.02	0.3	0.25	0.03
Language/ Semantic Memory	Category Fluency	Category Fluency	0.35	0.18	0.02	0.38	0.09	0.03	0.37	0.15	0.02
Episodic Memory	-	Benson Figure Recall	-	-	-	0.11	0.49	0.01	-	-	-
Non-memory related domains											
Attention/Executive	Digit Span Forward	Number Span Forward	0.27	-0.16	0.01	0.28	-0.2	0.01	0.26	-0.13	0.02
Attention/Executive	Digit Span Backward	Number Span Backward	0.34	-0.23	0.01	0.34	-0.27	0.02	0.33	-0.23	0.02
Attention/Executive	WAIS-R Digit Symbol	-	0.44	-0.16	0.05	-	-	-	0.43	-0.18	0.06
Attention/Executive	Trail Making A rate	Trail Making A rate	0.42	-0.18	0.44	0.39	-0.15	0.01	0.42	-0.19	0.59
Attention/Executive	Trail Making B rate	Trail Making B rate	0.41	-0.13	0.37	0.36	-0.16	0.05	0.41	-0.17	0.47
Language/Executive	-	Letter Fluency	-	-	-	0.33	-0.12	0.01	-	-	-
Visuospatial	-	Benson Figure Copy	-	-	-	0.28	-0.13	0.01	-	-	-

a. Trail Making rates were derived by dividing the number of correct lines completed by the time to completion.
b. Missing rate is the proportion of missing data due to non-cognitive problems for each neuropsychological test measure in each sample. Missing rates were less than .06 (i.e., 6%) with the exception of the Trail-Making Test, parts A and B in Sample 1 and in the autopsy confirmed subset of Sample 1 (due to correct lines completed on the Trail-Making Test not being recorded until late in the study). When PCA was performed on Sample 1 without the Trail-Making Test, and on Sample 1 and Sample 2 without imputation, results were consistent with those presented.

have contributed to the cognitive impairment. The correlation between factor loadings for PC2 computed separately from Samples 1 and 2 (using only tests that were identical or analogous in both samples) was .99, indicating strong reproducibility. Thus, we conclude that about 50% of the variance in UDS neuropsychological test scores in patients with probable AD is due jointly to overall cognitive severity and to the relative pattern of severity of memory versus non-memory deficits. Furthermore, these two factors can be consistently measured across independent cohorts using standardized data from similar (though different) neuropsychological tests.

2.4.3 Identification of Subtypes of Neuropsychological Deficit Patterns

To identify subgroups with distinct neuropsychological profiles, model-based clustering was applied to the autopsy-verified subset of Sample 1 (as the gold standard), using the derived factor scores for cognitive severity (PC1) and cognitive profile (PC2). Two clusters of participants were determined by the model (Fig. 2.1A): Cluster 1 (n=637; 79.6% of participants) was generally above the sample mean on the cognitive profile factor (PC2, mean=2.0) and nearly centered at the mean on cognitive severity (PC1, mean=0.3); Cluster 2 (n=163, 20.4% of participants) was generally below the mean on the cognitive profile factor (mean=-0.5) and also nearly centered at the mean on cognitive severity (mean=-0.08). Thus, the two clusters are separated by cognitive profile, but are largely overlapping on cognitive severity. The approximately 80% of participants in Cluster 2 were thus classified as having a "typical" AD cognitive profile, while the remaining 20% of participants (Cluster 1) and were classified as having an "atypical" AD cognitive profile.

When the same clustering procedure was applied to Sample 1 (the entire sample) and the number of clusters was set at two (based on results from the autopsy-verified subset), similar clusters were observed and a similar proportion of typical (n=3761; 79.8%) versus atypical (n=950; 20.2%) classification was obtained (Fig. 2.1C). An analysis without pre-determining the number of clusters gave similar results, but further split the two main clusters into sub-clusters. When the identical clustering procedure was applied to Sample 2, two clusters ('typical' and atypical') were again identified, with similar characteristics (Figure 2.1E; typical n=497; 71.8% of sample versus atypical n=195; 28.2% of sample).

To further explore the consistency of the classification procedure across different cohorts, we applied the Sample 1 classification rule to the PC scores identified from participants in Sample 2 and investigated whether the Sample-1 based rule and the Sample-2 based classification agreed in their assignment to a typical or atypical AD cognitive profile. The agreement rate was 94.7%: all 497 "typical" Sample 2 participants (as shown in Figure 2.1E) were also called typical using the Sample 1-determined rule; of the 195 "atypical" Sample 2 participants, 37 were misclassified

as typical using the Sample 1 rule. For completeness, the classification rules from Samples 1 and 2 are given below and R code to compute these assignments from the underlying raw neuropsychological test scores is available on the Alzheimer's Disease Cooperative Study website (https://data-archive.adcs.ucsd.edu).

The classification rules are given in the equation 2.1

$$If\ Pr(Y = Typical|X) = \frac{\hat{\pi}_1 f_1(X)}{\hat{\pi}_1 f_1(X) + \hat{\pi}_2 f_2(X)} > 0.5 \tag{2.1}$$

we classify Y as typical, otherwise we classify it as atypical

Here, X is the pair of scores $(PC1, PC2)$ for the participant, computed from the standardized neuropsychological test scores using the factor loadings given in Table 2.2, and standardized using the data in Table 2.1; and $f_1(X)$ and $f_2(X)$ are two-dimensional normal densities with $\mu_1 = (-0.03 - 0.50)^T$, $\mu_2 = (0.08, 1.29)^T$, $\Sigma_1 = \begin{pmatrix} 3.76 & -0.25 \\ -0.25 & 0.54 \end{pmatrix}$, $\Sigma_2 = \begin{pmatrix} 2.57 & 0.51 \\ 0.51 & 2.23 \end{pmatrix}$ for typical AD and atypical AD, in Sample 1. In Sample 1, $\hat{\pi}_1$ is 0.72 and $\hat{\pi}_2$ is 0.28; these are the proportions for typical AD and atypical AD. In Sample 2, $\hat{\pi}_1$ is 0.65 and $\hat{\pi}_2$ is 0.35, and $\mu_1 = (-0.13 - 0.72)^T$, $\mu_2 = (0.24, 1.32)^T$, $\Sigma_1 = \begin{pmatrix} 3.79 & -0.62 \\ -0.62 & 0.51 \end{pmatrix}$, $\Sigma_2 = \begin{pmatrix} 2.45 & 0.65 \\ 0.65 & 1.97 \end{pmatrix}$.

Mean differences on the individual neuropsychological test scores for the typical and atypical AD subtypes are illustrated graphically in Figures 2.1(B), (D) and (F). Raw test scores are scaled to z-scores based on normative data from approximately 3600 cognitively normal UDS participants21,22. Atypical participants had better performance than typical participants on episodic and semantic memory measures on average, but slightly worse performance on measures of attention and executive function. Table 2.3 compares demographic features and clinical and neuropsychological test scores for typical and atypical participants within each of the three datasets.

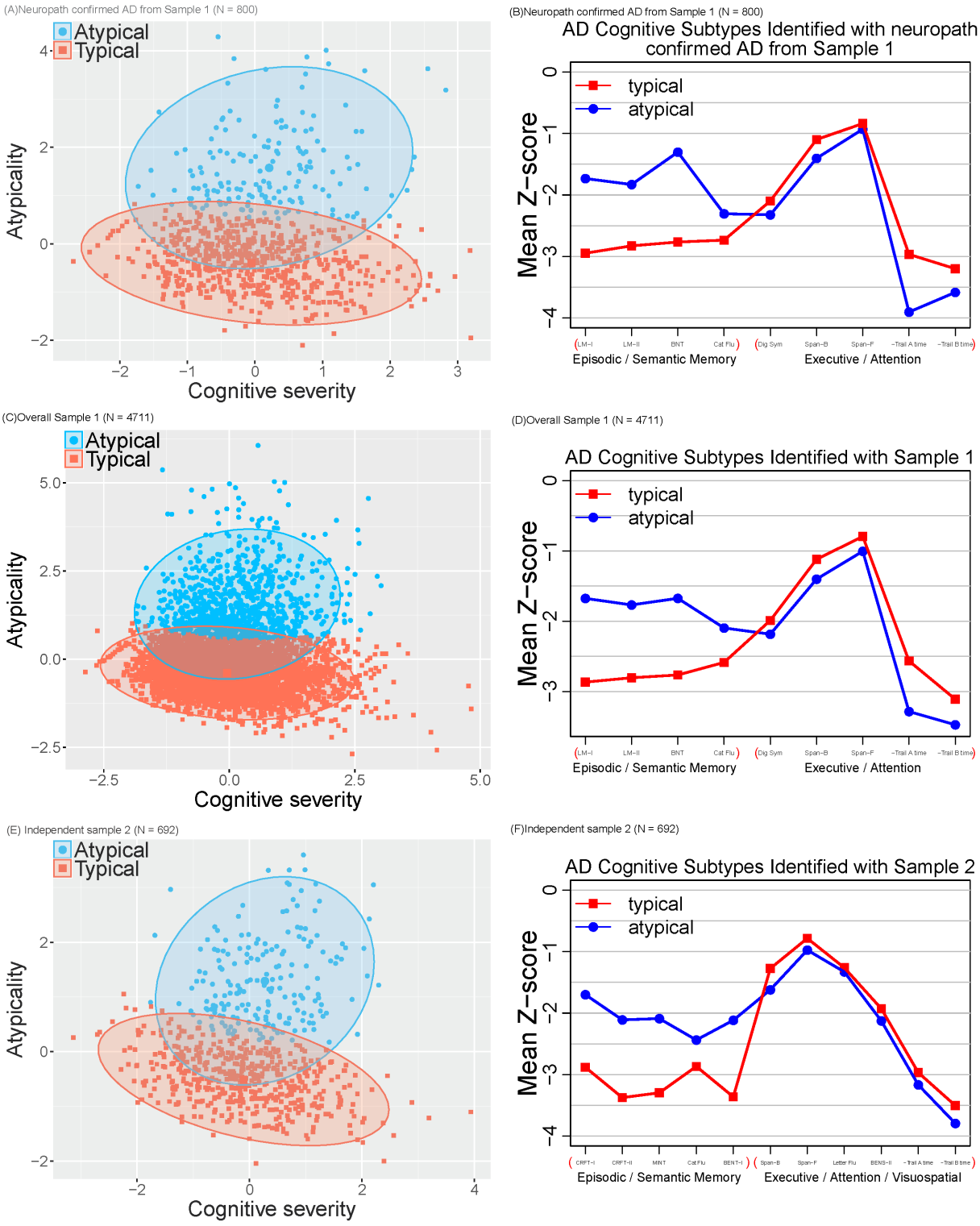

Figure 2.1: Results of model-based clustering and z-scores on neuropsychological test measures

2.4.4 Demographic and Clinical Characteristics of Neuropsychological Subtypes

Univariate linear models showed that higher scores on PC2 (i.e., a higher measure of atypicality in cognitive profile) are significantly associated with younger age, male sex, lower probability of APOE e4, less severe global dementia, higher GDS scores, and greater likelihood that the clinician believes depression contributes to cognitive impairment (all p values <0.001). A multivariate linear regression model (omitting clinician-rated depression due to collinearity with GDS score) showed that all variables significant in the univariate models remained significantly associated with atypicality, with similar effect sizes.

2.4.5 Neuropathological Characteristics of Neuropsychological Subtypes

Approximately 82% (800/976) of individuals diagnosed with probable AD in Sample 1 met neuropathological criteria for AD at autopsy. The 176 misdiagnosed cases included 14 with frontotemporal lobar degeneration (FTLD), 49 with DLB, 62 with CVD, and 56 with low levels of AD pathology that did not reach diagnostic thresholds. Cognitive profile classification of these individuals was as follows. FTLD: 78.6% typical (11 typical, 3 atypical); DLB: 55.1% typical (27 typical, 22 atypical); CVD: 69.3% typical (43 typical, 19 atypical); low level of pathology: 78.6% typical (44 typical, 12 atypical).

A number of the 800 individuals who met neuropathological criteria for AD had a secondary pathological diagnosis. These included 4 with secondary FTLD, 269 with secondary DLB and 271 with secondary CVD. Cognitive profile classification of these individuals was as follows. AD and secondary FTLD: 100% typical (4 typical, 0 atypical); AD and secondary DLB: 82.1% typical (221 typical, 48 atypical); AD and secondary CVD: 81.9% typical (222 typical, 49 atypical). Those with AD and no secondary pathological diagnosis included 79.9% typical (163 typical, 41 atypical).

Overall, a lower Braak stage was associated with a higher likelihood of an atypical cognitive profile (Fig. 2.2A): 32% (76/238) of Braak stage 0 to IV, 17.2% (43/250) of Braak stage V, and 16.0% (77/480) of Braak stage VI were classified as atypical (Braak stage was missing for 8 participants). Linear regression showed a strong inverse relationship between "atypicality" score (PC2) and Braak stage treated as an ordinal predictor ($p<0.001$). There was only a marginally significant negative linear relationship between atypicality score and degree of neocortical plaque pathology ($p=0.04$; Fig. 2.2B). When both neuropathological measures were included in a multivariate model, there was a highly significant negative effect of Braak stage ($p<0.001$) and no significant effect of plaque pathology. Similar results were obtained when these analyses were restricted to the 800 participants with neuropathologically-confirmed AD (Fig. 2.2c, d). Univariate linear regression models showed no significant relationship between degree of atypicality in cognitive profile and the presence of FTLD pathology (n=18), DLB pathology (n=318) or CVD pathology (n=333).

2.4.6 Stability of Neuropsychological Subtype classification over time

Because the UDS clinical and neuropsychological evaluations were repeated approximately annually, 2,944 of the 4,711 participants with probable AD in Sample 1 had two or more evaluations, including the baseline evaluations which were used to determine the Table 2.2 factor loadings and thus define the cognitive atypicality score (PC2). We used these fixed factor loadings in Table 2.2 to calculate the degree of cognitive atypicality from test scores at each subsequent evaluation. For participants in the "atypical" cluster at baseline, the mean cognitive atypicality score was 2.02 at baseline, 1.33 at the second evaluation, and then only decreased by 0.34 over the next two annual evaluations. For those in the "typical" cluster at baseline, the mean cognitive atypicality score was -0.50 at baseline, -0.33 at the second evaluation, and then increased by only 0.11 over the next two evaluations. Thus, the cognitive "atypicality" score is relatively stable over a four-year interval.

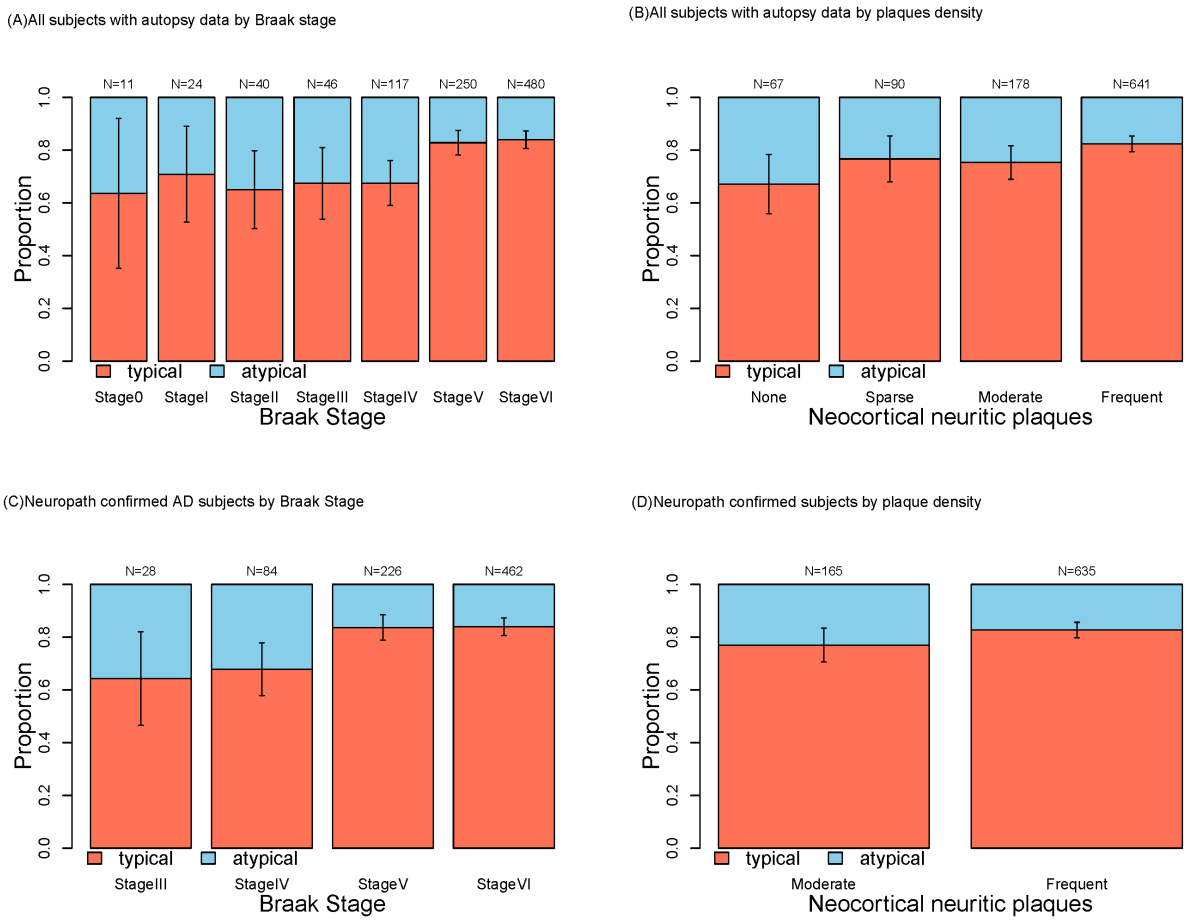

Figure 2.2: Proportion of typical and atypical participants by Braak and CERAD rating

Table 2.3: Mean (SD) demographic characteristics, clinical test scores, and neuropsychological test scores for typical and atypical AD subtypes within each sample

		Sample 1		Sample 2		Neuropath confirmed AD	
		Typical	Atypical	Typical	Atypical	Typical	Atypical
N		3,761	950	497	195	637	163
Age		76.5 (9.1) *	74.6 (10.5) *	75.0 (9.5) *	72.4 (10.2) *	77.5 (9.5) *	74.9 (11.6) *
Sex: Female - N(%)		2191 (58.3) *	489 (51.5) *	272 (54.7)	108 (55.4)	313 (49.1)	73 (44.8)
Education		14.2 (3.5)	14.4 (3.9)	15.5 (3.0)	15.3 (3.3)	15.1 (3.1)	15.1 (3.4)
MMSE		20.6 (2.5) *	21.6 (2.2) *	N/A	N/A	20.3 (2.6) *	21.4 (2.3) *
MoCA		N/A	N/A	13.7 (3.6) *	15.1 (3.1) *	N/A	N/A
CDR Global rating		1.01 (0.5) *	0.82 (0.4) *	1.04 (0.5) *	0.78 (0.4) *	1.11 (0.5) *	0.91 (0.4) *
CDR Sum of boxes		5.84 (2.7) *	4.53 (2.1) *	6.10 (2.8) *	4.36 (2.2) *	6.52 (2.9) *	5.10 (2.1) *
Geriatric Depression Scale		2.29 (2.5) *	2.86 (2.8) *	2.07 (2.5) *	3.05 (2.8) *	2.18 (2.3) *	2.61 (2.4) *
Functional Activities Questionnaire		17.9 (7.6) *	14.0 (7.7) *	18.6 (7.0) *	13.2 (7.2) *	20.4 (7.0) *	16.4 (6.5) *
APOE e4 Positive – N (%)		1818 (60.7) *	389 (52.8) *	221 (68.0) *	49 (43.0) *	355 (60.8)	88 (59.1)
Depression[a] – N (%)		781 (20.8) *	229 (24.1) *	112 (22.5)	50 (25.6)	111 (17.4)	29 (17.8)
Neuropsychological Test Measures							
Memory related cognitive domains							
Epi. Memory	Para. Immed. recall	2.72 (2.2) *	7.37 (2.9) *	4.29 (3.1) *	9.13 (3.3) *	2.41 (2.1) *	7.14 (3.4) *
Epi. Memory	Para. Delayed recall	0.54 (0.9) *	5.00 (3.2) *	0.59 (1.2) *	6.03 (3.3) *	0.44 (0.9) *	4.73 (3.2) *
L./ S. Memory	Confrontation Naming	18.4 (7.0) *	21.8 (5.5) *	22.1 (6.9) *	25.0 (5.6) *	18.4 (7.0) *	23.0 (4.9) *
L./ S. Memory	Category Fluency	16.3 (6.8) *	19.8 (8.3) *	16.0 (7.0) *	19.1 (6.6) *	15.2 (6.6) *	18.3 (6.8) *
Epi. Memory	Benson Figure Recall	N/A	N/A	0.78 (1.5) *	4.64 (3.7) *	N/A	N/A
Non-memory related cognitive domains							
Att./ Exec.	Digit Span Backward	4.44 (1.9) *	3.82 (1.6) *	4.40 (2.1) *	3.63 (1.7) *	4.48 (1.9) *	3.81 (1.6) *
Att./ Exec.	Digit Span Forward	6.93 (2.2) *	6.49 (2.1) *	6.50 (2.3) *	6.06 (1.9) *	6.84 (2.4) *	6.65 (2.4)
Att./ Exec.	WAIS-R Digit Symbol	22.1 (13.7) *	19.7 (12.1) *	-	-	20.8 (13.5) *	18.0 (12.6) *
Att./ Exec.	Trail Making A T.	74.1 (41.4) *	85.2 (41.7) *	70.7 (40.0)	73.3 (40.0)	80.3 (42.7) *	94.7 (43.9) *
Att./ Exec.	Trail Making A CL.	21.8 (5.2)	21.6 (5.3)	22.3 (5.4)	22.4 (4.9)	20.8 (6.4)	20.3 (6.7)
Att./ Exec.	Trail Making B T.	245.8 (77.4) *	264.0 (64.2) *	244.4 (80.8) *	257.8 (69.7) *	250.3 (75.2) *	269.5 (60.9) *
Att./ Exec.	Trail Making B CL.	12.4 (10.1) *	11.3 (9.5) *	12.8 (10.6)	12.3 (10.3)	11.3 (10.0) *	10.1 (9.7)
L./ Exec.	Letter Fluency	-	-	18.4 (8.9)	17.8 (7.5)	-	-
Visuospatial	Benson Figure Copy	-	-	12.9 (4.4)	12.6 (4.4)	-	-

Abbreviations: Epi. = Episodic; Para. = Paragraph; Immed. = Immediate; L./S. = Language/Semantic; Att./Exec. = Attention/Executive; L./Exec. = Language/Executive; T. = Time; CL. = Correct Lines.

a. Clinician impression that depression contributed to participant's cognitive deficits.

*Statistically significant difference between AD subtypes within the sample ($P < 0.05$)

2.4.7 Rates of Decline of Neuropsychological Subtypes over time

The rate of global cognitive/clinical decline was compared between participants with typical or atypical cognitive profile at baseline using linear mixed effects models. Decline was measured by MMSE, CDR sum of boxes and CDR global ratings, over two years with three annual evaluations in Sample 1 (baseline, year 1, year 2). Predictors were time (year), AD cognitive subtype, and time x AD cognitive subtype interaction. As expected, patients worsened significantly over time, with estimated trends for MMSE -2.59 [-2.73, -2.44] points/year, CDR global ratings 0.32 [0.31, 0.34] points/year and CDR sum of boxes 2.03 [1.94, 2.12] points/year (Figure 2.3). The interaction between time and AD cognitive subtype was statistically significant for MMSE (0.35 [0.03, 0.67] points/year) and CDR global ratings (-0.04 [-0.08, -0.001] points/year) indicating that atypical patients declined more slowly than typical patients. Mean two-year decline on the MMSE was 4.47 points for the cognitively atypical AD patients and 5.17 points for the cognitively typical patients, a 15.7% difference in total change between subtypes. The mean two-year increase in CDR global ratings 0.56 points for the cognitively atypical AD patients and 0.65 points for the cognitively typical patients, a 16.1% difference. The mean two-year increase in CDR sum of boxes was 3.75 points for the cognitively atypical AD patients and 4.07 points for the cognitively typical patients, an 8.5% difference.

2.5 Discussion

Our objectives were to identify heterogeneity in cognitive profiles of patients with mild-to-moderate probable AD dementia and to determine whether cognitive profiles are systematically related to the clinical course and neuropathological features of the disease. To achieve our objectives we: (1) determined an empirically-derived classification rule based on neuropsychological test scores that revealed cognitive subgroups within a sample of mildly-to-moderately demented patients with probable AD; (2) validated the classification rule in an independent sample of AD

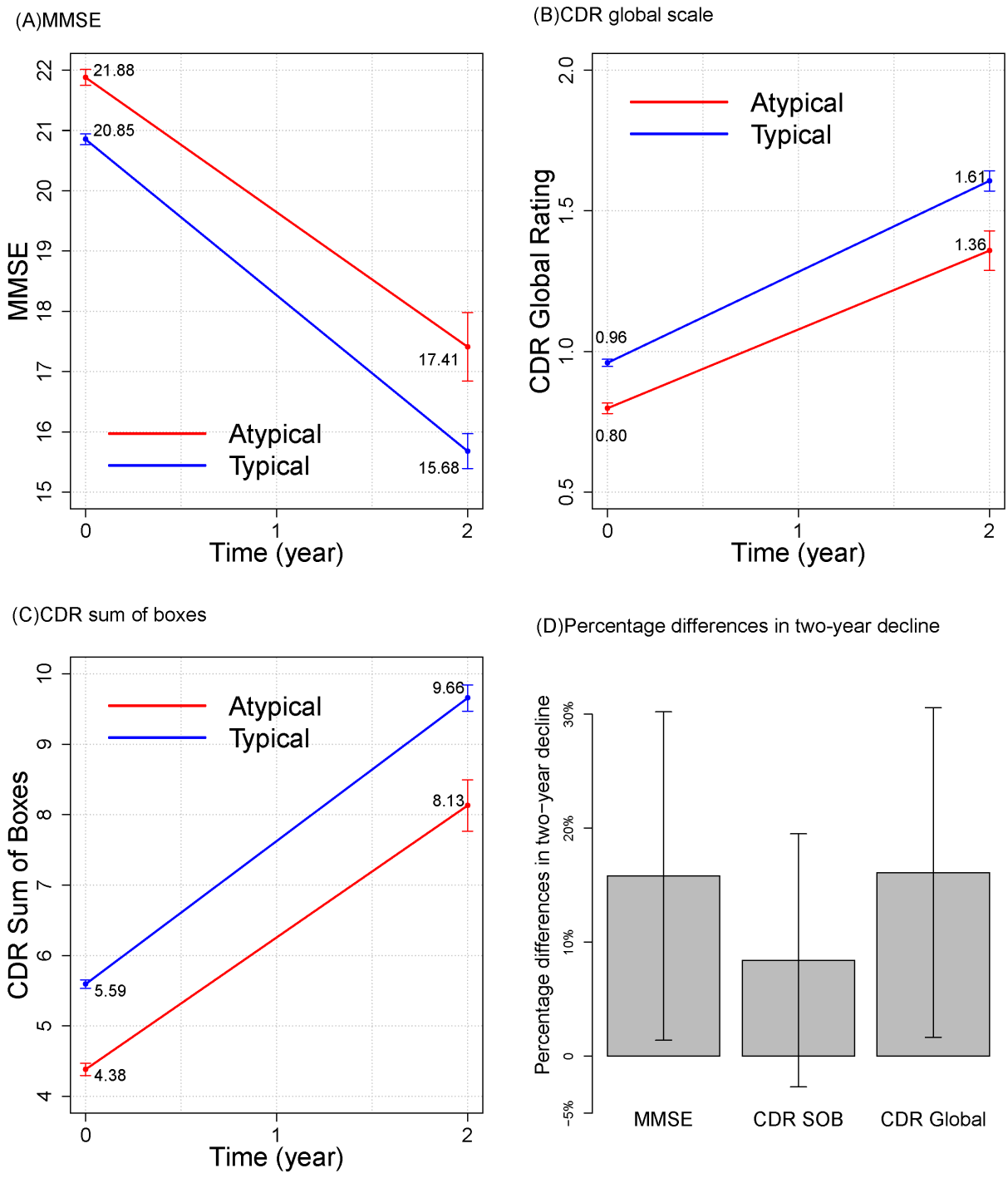

Figure 2.3: Model fitted lines comparing rate of change over two years for typical and atypical AD

patients, tested with a similar set of neuropsychological measures; (3) assessed the stability of subtype classification over time and compared rates of cognitive and functional decline among cognitive subgroups; and (4) compared neuropathological features among cognitive subgroups.

Model-based clustering of PCA factor scores produced "typical" (79.6% of the sample) and "atypical" (20.4% of the sample) cognitive profile subgroups, in a sample with autopsy-verified AD. The typical profile was characterized by greater deficits in episodic and semantic memory than in attention and executive functions, whereas the atypical profile had similar levels of impairment across all cognitive domains. From another perspective, the atypical profile could be viewed as having milder than expected memory impairment given the severity of deficits in attention and executive functions. Similar results were obtained in the discovery (typical: 79.8%; atypical: 20.2%) and independent validation (typical: 71.8%; atypical: 28.2%) samples. Furthermore, the exclusion of participants for whom depression may have contributed to cognitive impairment did not alter this pattern of results. The reproducibility of these subtypes was evidenced by strong correlation in factor loadings between samples as well as the similarity of the identified cognitive profiles, and is notable given minor differences in the specific neuropsychological tests that were administered across the original and validation samples.

The two-cluster solution we observed is consistent with previous results [60]. Although the prevalence of the atypical cognitive subtype was lower (20-28%) in the present study than in some previous reports (e.g., 29%-52%; Scheltens et al., 2017), this difference could be related to our exclusion of clinically diagnosed variant phenotypes of AD (e.g., logopenic PPA, posterior cortical atrophy) that would usually not be included in an AD clinical trial. These phenotypic AD variants were not excluded in other studies [10, 12, 45, 59, 60, 65] which may explain why they identified a "preserved" memory subtype that we did not observe.

A strength of our study is the neuropathological validation of the clinical diagnosis of probable AD in a large subset of patients. Approximately 82% of those diagnosed with probable AD met neuropathological criteria for AD, and 68% of those had a secondary pathological

diagnosis in addition to AD (e.g., DLB, CVD). The distribution of atypical cognitive profiles in patients with autopsy-verified AD was similar in those with (typical: 81.6%; atypical: 18.4%) and without (typical: 79.9%; atypical: 20.1%) secondary pathology. There was a slightly higher percentage of atypical cognitive profiles in those who had been clinically misdiagnosed (i.e., those with non-AD pathology only; typical: 70.5%; atypical: 29.5%) compared to the overall sample, but this difference was not significant. These findings suggest that the atypical cognitive profile we observed in a subset of patients is not due to presence of non-AD pathology, but more likely is driven by variability in the severity and distribution of AD pathology. Furthermore, our results suggest that identification of an atypical cognitive profile will not be particularly helpful in differentiating between those with or without AD pathology, supporting the need to measure amyloid and tau biomarkers in selecting participants for clinical trials.

Consistent with the notion that variability in AD pathology drives the typical-atypical distinction, we found that the typical cognitive profile was associated with higher Braak stages than the atypical cognitive profile. There was a strong inverse relationship between Braak stage and atypicality score that was not modified significantly by level of neocortical plaque pathology. There was no significant independent relationship between level of plaque pathology (i.e., amyloid burden) and atypicality score. These results suggest that the typical-atypical distinction is largely determined by tangle pathology. The milder than expected memory impairment (given the severity of deficits in executive functions and attention) of the atypical group suggests that they have less pathology in the entorhinal cortex, hippocampus, and surrounding temporal lobe neocortex than do those in the typical subgroup. This is consistent with their lower average Braak stage (which reflects both severity and distribution of tangle pathology) and lower likelihood of having an APOE e4 genotype compared to the typical subgroup. Other researchers have also reported more extensive tangle pathology and a higher prevalence of APOE e4 in AD patients with substantial relative memory impairment [10]. The typical-atypical classification does not simply reflect stage of disease, however, since subgroup classification remained stable across longitudinal assessments

(i.e., patients with an atypical profile did not develop a typical profile over time). Our ability to address the effect of comorbid pathologies on degree of atypicality was limited by the lack of systematic characterization of the severity of comorbid neuropathology in the NACC database.

Comparison of rates of cognitive and functional decline among cognitive subgroups showed that the atypical cognitive profile was associated with slower decline on the MMSE (typical: 2.59 points/yr.; atypical: 2.17 points/yr.) and the CDR (global rating, typical: .32/yr.; atypical: .27/yr.; sum of boxes, typical: 2.03 points/yr.; atypical: 1.87 points/yr.). Over 2 years, there was an approximately 16% difference in rate of decline on both the MMSE and the CDR global rating, and an approximately 9% difference on the CDR sum of boxes score. These differences in rate of decline are noteworthy, given the typical therapeutic effect sizes reported in AD trials, suggesting that cognitive heterogeneity may be a source of variability that should be considered in trial design and interpretation. For example, a trial that included 80% typical and 20% atypical patients (the composition of our sample) would need approximately 10% more participants (678 versus 616) than a trial that included only typical patients to achieve 90% power to observe a 25% difference between treatment and control groups in two-year rate of decline on the MMSE. However, if the composition of the group were 50% typical and 50% atypical, a distribution that has been observed in several cohorts [60], approximately 30% more participants (794 versus 616 for 90% power) would be needed. Thus, cognitive heterogeneity could have a considerable effect on statistical power, depending upon the prevalence of atypicality.

The decision rule we developed with model-based clustering methods can be used to classify patients with probable AD into typical and atypical subtypes in other samples tested with the same or similar neuropsychological tests. When we used the rule developed in the original sample to classify patients in the autopsy-confirmed subsample or the independent sample there were low misclassification rates (5% and 1%), and high sensitivity (both 100%) and good specificity (81% and 94%) in relation to classification using their own models. Thus, cognitive subtype can be determined easily in the future by applying our proposed decision rule.

Our study has several limitations. First, data regarding timing and reasons for dropout were limited, restricting the opportunity for modeling of early drop outs who might have had more aggressive decline within the 4-year study timeline. However, the rate of drop-out in the typical and atypical subgroups was similar. Second, the neuropathological diagnosis of AD was based on a relatively low threshold of Braak stage III or higher with a CERAD neuritic plaque rating of moderate or frequent. Our overall pattern of results did not change, however, when we use more stringent neuropathological diagnostic criteria (Braak stage V or higher with a CERAD neuritic plaque rating of frequent, data not shown). Finally, the UDS neuropsychological test battery is brief and limited in scope, which may have precluded our ability to detect additional cognitive subtypes; however, similar 2-cluster solutions have been reported in four different AD cohorts using four different neuropsychological test batteries [60], suggesting that this is a robust finding. Nevertheless, more detailed neuropsychological assessment may reveal additional cognitive subtypes, particularly in early dementia.

Despite these limitations, our results show that we can reliably identify distinct cognitive profiles among clinically-diagnosed probable AD patients. These cognitive profiles are differentially associated with tangle pathology and have different rates of decline; hence cognitive heterogeneity in probable AD may have implications for clinical trials, especially therapies targeting tau.

2.6 Study Funding

The NACC database is funded by NIA/NIH Grant U01 AG016976. NACC data are contributed by the NIA-funded ADCs: P30 AG019610 (PI Eric Reiman, MD), P30 AG013846 (PI Neil Kowall, MD), P50 AG008702 (PI Scott Small, MD), P50 AG025688 (PI Allan Levey, MD, PhD), P50 AG047266 (PI Todd Golde, MD, PhD), P30 AG010133 (PI Andrew Saykin, PsyD), P50 AG005146 (PI Marilyn Albert, PhD), P50 AG005134 (PI Bradley Hyman, MD, PhD),

P50 AG016574 (PI Ronald Petersen, MD, PhD), P50 AG005138 (PI Mary Sano, PhD), P30 AG008051 (PI Thomas Wisniewski, MD), P30 AG013854 (PI M. Marsel Mesulam, MD), P30 AG008017 (PI Jeffrey Kaye, MD), P30 AG010161 (PI David Bennett, MD), P50 AG047366 (PI Victor Henderson, MD, MS), P30 AG010129 (PI Charles DeCarli, MD), P50 AG016573 (PI Frank LaFerla, PhD), P50 AG005131 (PI James Brewer, MD, PhD), P50 AG023501 (PI Bruce Miller, MD), P30 AG035982 (PI Russell Swerdlow, MD), P30 AG028383 (PI Linda Van Eldik, PhD), P30 AG053760 (PI Henry Paulson, MD, PhD), P30 AG010124 (PI John Trojanowski, MD, PhD), P50 AG005133 (PI Oscar Lopez, MD), P50 AG005142 (PI Helena Chui, MD), P30 AG012300 (PI Roger Rosenberg, MD), P30 AG049638 (PI Suzanne Craft, PhD), P50 AG005136 (PI Thomas Grabowski, MD), P50 AG033514 (PI Sanjay Asthana, MD, FRCP), P50 AG005681 (PI John Morris, MD), P50 AG047270 (PI Stephen Strittmatter, MD, PhD).

2.7 Afterthoughts before next Chapter

This chapter demonstrated several essential findings and conclusions, which may contribute to AD clinical studies and neuroscience more broadly. In addition, for the first time, this study introduced me to the world of randomized trials and has motivated me to think about developing statistical methods to solve problems in randomized trials. As described in section 2.3.3, only two years of the available four years of follow-up data were used for longitudinal analysis in this study due to apparent informative censoring after two years. Discontinuation is always a significant problem in randomized trials. In the UDS data set, the reasons for dropout are mainly because of physical or cognitive problems. However, it is more complicated in a real trial. Participants may terminate and withdraw from the study due to many reasons such as adverse events, caregiver issues, guess of the blinded arms, and other factors. Regulatory guidance emphasizes the importance of carefully handling missing data and describing the intercurrent events, which we will discuss in Chapter 4. What can we contribute to this problem from a

statistical aspect?

In the subsequent chapters, we will introduce some classic and commonly used methods for dealing with dropouts in randomized trials with examples in both primary and sensitivity analysis. Then we propose a novel approach called "Doubly Robust" imputation, which can obtain consistent estimators by specifying a missingness model and an imputation model. This method can fix the bias when the imputation model is misspecified by using a correct missingness model.

This chapter, in full, has been published and may be found as "Qiu, Yuqi; Jacobs, Diane M.; Messer, Karen S.; Salmon, David P.; Feldman, Howard H. *Cognitive Heterogeneity in Probable Alzheimer's Disease: Clinical and Neuropathological Features*, Neurology, 93 (8), e778-e790, 2019". The dissertation author was the primary author on this paper.

Chapter 3

Doubly Robust Imputation in Longitudinal Studies, with an Application to an Alzheimer's Clinical Trial

3.1 Abstract

Motivated by the informative dropout that is common in FDA-regulated trials for Alzheimer's disease, we propose a doubly-robust imputation approach to adjusting for dropout-related bias in longitudinal studies. The approach uses standard software for estimating equations and is appropriate for the estimand of interest in clinical trials. We show that most doubly robust estimators can be written in imputation form, an approach that may be easier to understand and apply. We present two particular imputation estimators which are closely related to existing doubly robust estimators for longitudinal data. We illustrate the imputation approach using historical data from the Alzheimer's Disease Cooperative Study (ADCS), a major center for clinical trials. Similar ADCS data is commonly used by us and others to support the design of current clinical trials, increasing the relevance of the illustration. Simulation studies support the theoretical properties of

the estimators and provide comparisons with alternative doubly-robust approaches. The imputation approach we present has the advantage of computational simplicity and transparency compared to existing approaches in the literature and may be suitable for use in FDA regulated trials as well as a variety of other applications.

3.2 Introduction and background

Outcome-related and treatment-related dropouts are common in randomized trials for Alzheimer's disease and other dementias. In planning such studies, it is common to assume dropout rates of 25% or more, with higher dropout expected among patients who progress faster and on active treatment. The primary outcome measures in Alzheimer's clinical trials are typically within-subject change from baseline on several relatively demanding neurocognitive tests; subjects are assessed in the clinic at regular intervals over a period of a year or more. In such FDA-regulated trials, the primary analysis would typically compare model-based estimates of the difference between treatment arms at the final time point, following FDA guidance and practices which support using a restricted set of models, with limited adjustment by covariates. In particular, a mixed model with repeated measures (MMRM) would often be used, parameterized with categorical time. The covariance matrix would be fully parameterized or a simple working covariance matrix with a sandwich estimate of variance in a generalized estimating equations (GEE) approach. Together, these trial characteristics make it likely that substantial dropout will occur. The dropout may not satisfy the missing-at-random (MAR) assumptions needed to obtain consistent estimates from the primary analysis.

In this setting, interest lies in obtaining a robust and consistent estimate of a parameter defined as the solution to an estimating equation U, such as the mean difference between arms at a specified time point, or a model coefficient such as the treatment-by-time interaction. Importantly the parameter of interest is the solution of U applied to the "full" data, by which we mean data

with no dropout. However, the primary intent-to-treat analysis from the trial would apply U to the observed data, which would indeed provide unbiased estimates of the parameter if the data are missing completely at random (MCAR). Because of the complex patterns of dropout and the limited modeling assumptions incorporated into U, the MCAR assumption is unlikely to hold, and this approach will generally provide biased estimates. If the observed data are MAR, then maximum likelihood methods can provide consistent and efficient estimates [68], but here U may not capture the correct likelihood.

In this setting of longitudinal data with informative dropout, Paik (1997) [44] defined a consistent estimator by first applying a sequential regression approach to impute the missing outcome values and then applying U to the completed data. Consistent estimates may also be obtained by inverse probability weighting (IPW) of the estimating equations U applied to data from completers only, using the estimated probability of dropout, as reviewed in Robins et al. (1995) [50]. They showed that any efficient estimator in this setting takes the form of an augmented IPW estimator (AIPW) that combines a regression modeling component, which 'imputes' missing data, and an IPW component, which weights observed data. Scharfstein, Rotnitzky and Robins (1999) [58] noticed that these AIPW estimators enjoy a double robustness property, in that they are consistent when either the regression model is correct or the model for the dropout probability is correct. Bang and Robins (2005) [2] first extended such DR estimators to longitudinal monotone missing data using a sequential regression approach, but the method did not apply to estimating equations U. Tsiatis and colleagues [68, 69] first extended optimal AIPW estimating equations to the longitudinal setting with dropout. However, the estimator is computationally complex and did not demonstrate superiority in simulations. Seaman and Copas (2009) [63] developed a simpler DR generalized estimating equation (GEE) estimator for estimating regression coefficients in longitudinal studies, which still requires specialized optimization routines. Since then, doubly robust estimators have been extensively developed, as reviewed in Seaman and Vansteelandt (2018) [64]. However, doubly robust longitudinal estimating equations seem to be rarely used in practice,

possibly because existing approaches are complex and require specialized software.

Here, we develop two imputation-based doubly robust estimators for longitudinal data based on the full data estimating equations U, using the approaches of Paik (1997) [44] and Seaman and Copas (2009) [63]. These imputation-based estimators require only standard software tools, and one is computationally simpler than existing approaches. We show that any AIPW estimator can be written in substitution (plug-in) form, for U which are linear in the data, and conversely. This development also provides a simple and direct proof of double robustness for these estimators. We also extend the approach of Bang and Robins (2005) [2] to estimating equations and relate it explicitly to Paik (1997) [44]. We compare the performance of these estimators by simulation and by application to data from a prodromal stage trial in Alzheimer's Disease.

3.2.1 The MCI trial of Donepezil

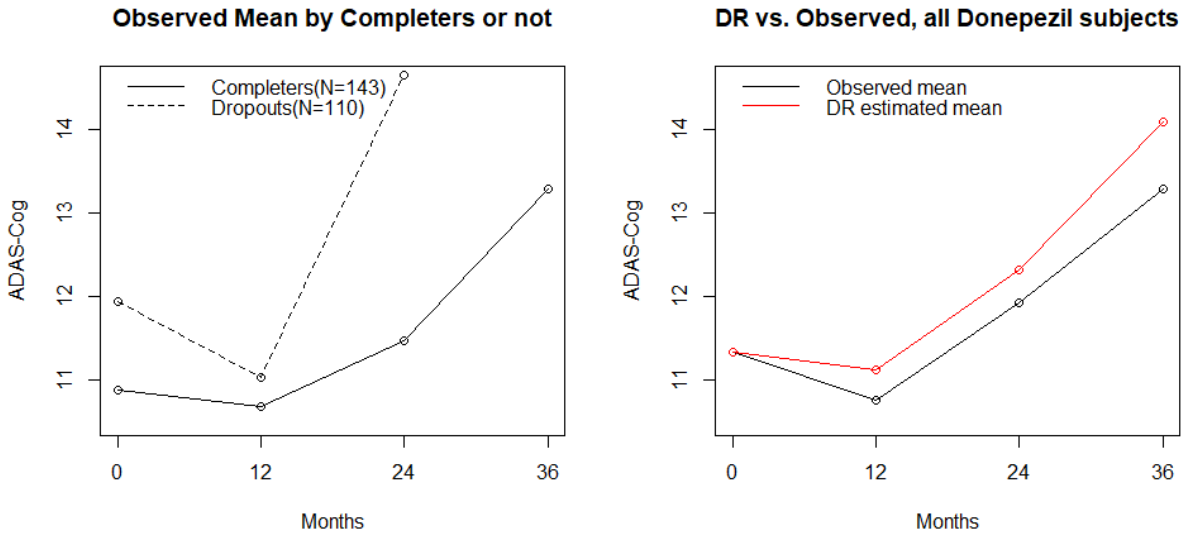

Figure 3.1: Estimated mean of ADAS-Cog for Donepezil group over time

Donepezil is a widely used cholinesterase inhibitor that improves symptoms and might delay the clinical diagnosis of Alzheimer's disease in subjects with the amnestic form of mild

cognitive impairment (MCI). A randomized, double-blind, placebo-controlled, parallel-group trial was conducted by the Alzheimer's Disease Cooperative Study (ADCS) between March 1999 and January 2004 [46]. The study compared the time to progression to possible or probable AD among 769 subjects with MCI randomized to treatment for 36 months with either Donepezil (n=253), Vitamin E (n=257), or placebo (n=259). Dropout rates by month 36 were 42.7%, 38.1%, and 32.0% for the Donepezil, Vitamin E, and placebo groups, respectively. Figure 3.1 displays the apparent bias due to dropouts for the mean score on one of the primary outcome measures, the Alzheimer's Disease Assessment Scale-Cognitive Subscale (ADAS-Cog), for the Donepezil group. The left panel shows that patients who eventually dropped from the study had a much higher ADAS-Cog, indicating greater cognitive impairment than patients who completed the study, and that the gap increases over time. The right figure shows that the estimated mean from a DR approach is consistently higher than the mean of the observed data, indicating that a DR approach may help to improve estimated effects from this trial. These data are taken from the data archives at the ADCS; these and similar data are often used in simulation studies that inform the design of current AD trials.

3.2.2 Aims and organization of this chapter

The chapter is organized as follows: in section (3), we describe notation and give details of the IPW and sequential regression approaches which are the building blocks of our longitudinal DR methods. We briefly review existing DR methods for cross-sectional data in section (4) and longitudinal data in section (5). In section (6), we develop the DR imputation approach and show its equivalence with an AIPW approach. In section (7), we use the DR imputation approach to define two specific DR estimators. We use simulation to compare the two DR imputation estimators with an extension of the Bang and Robins estimator, as well as with maximum likelihood and GEE approaches, in section (8). Section (9) presents an application to a trial in Alzheimer's Disease, and section (10) presents discussions and conclusions.

3.3 Regression modeling and inverse probability weighting approaches to longitudinal dropout

3.3.1 Notation and data structure

Assume we have N iid subjects potentially observed at times $j = 1, \ldots, M$, and for individual i at time j there is data $L_{ij} = (Y_{ij}, X_{ij})$, where Y_{ij} is a univariate outcome and X_{ij} is a vector of covariates; there may also be a vector of always observed time independent baseline covariates X_{i0}. Let $\bar{L}_{ij} = (X_{i0}, L_{i1}^T, ..., L_{ij}^T)$ denote the historical data from time 1 to j. We will often drop the subscript i when the meaning is clear. We assume the distribution of $\bar{L}_M$ has finite second moments.

Each subject can potentially drop out from the study. Let $R_j \in \{0,1\}$ be a binary missing indicator, so that we observe R_j and $(R_j Y_j, R_j X_j)$. Under the assumption of monotone dropout, if $R_j = 0$ then for $t > j$, $R_t = 0$. Let C_j be a censoring indicator, where $C_j = 1$ indicates j is the last observed time for subject i, otherwise $C_j = 0$, and let J be the index of the last observed time point, so that $C_J = 1$. Under the MCAR assumption, R_j is independent of $\bar{L}_M$. Under the missing at random assumption (MAR), $P(R_j = 1|\bar{L}_M, R_{j-1} = 1) = P(R_j = 1|\bar{L}_{j-1}, R_{j-1} = 1)$ so that the probability of a missing outcome depends only on previously observed data. We also assume there is probability bounded away from zero of seeing full data over the whole support of L_M: $P(R_{iM} = 1|\bar{L}_M) > \varepsilon > 0$.

3.3.2 Estimating equations which define the estimand

We assume that there is a vector of parameters β, and a corresponding vector of sufficiently smooth estimating equations $U(\cdot, \beta)$ such that β^* is the unique solution to $E[U(\bar{L}_M, \beta)] = 0$. One or more of the parameters in β^* is the primary estimand of interest in the study. Then the solution $\hat{\beta}$ to the full data estimating equations $\sum_{i=1}^N U(\bar{L}_{iM}, \beta) = 0$ is consistent for β^* and asymptotically

normal, by standard arguments. For the convenience of notation, we will often suppress the dependence of U on β.

In particular we will often assume the data follow a generalized linear model for the mean $\mu_j = E(Y_j|X_j)$, with link function $g(\mu_j) = X_j\beta$. A common choice for longitudinal data is to assume a multivariate normal distribution with a specific form for the variance-covariance matrix of the random errors, such as a mixed effects model, which includes both fixed effects and random slopes and intercepts. Thus U might be taken to be the score equations from the likelihood, or alternatively, the generalized estimating equations (GEE) [25] applied to the full data:

$$\sum_{i=1}^{N} U_k(\bar{L}_{iM}) = \sum_{i=1}^{N} \frac{\partial \mu_i^T}{\partial \beta_k} V_i^{-1}(Y_i - \mu_i) = 0 \tag{3.1}$$

where V_i is an assumed working covariance matrix for Y_i. Here, the efficient choice for V_i^{-1} is the true covariance matrix of the data. However, under general regularity conditions, the solution $\hat{\beta}$ to the full data GEE's (3.1) is consistent for β^* and asymptotically normal, for arbitrary V.

With missing data, instead of (3.1) we observe

$$\sum_{i=1}^{N} U_k(\bar{L}_{iJ}) = \sum_{i=1}^{N} \frac{\partial \mu_i^T}{\partial \beta_k} V_i^{-1} D_i(Y_i - \mu_i) = 0 \tag{3.2}$$

where $D_i = \mathrm{diag}(R_{i1}, \ldots, R_{iM})$. The solution $\hat{\beta}$ to (3.2) remains consistent for the solution to (3.1) if the data are MCAR, since then $E[R_j Y_j] = E[Y_j]$ and so (3.2) is a consistent estimator of $E[U_k(\bar{L}_{iM})]$. However, when the dropout is MAR so that $E[R_j Y_j] \neq E[Y_j]$, then in general the solution $\hat{\beta}$ to (3.2) will not be consistent for β.

3.3.3 IPW estimating equations, for dropout that is MAR

Robins et al. (1995) [50] showed how to incorporate inverse-probability weights into U applied to observed data when the dropout is MAR. Let

$$\lambda_j = P(R_j = 0 | R_{j-1} = 1, \bar{L}_{j-1}) \tag{3.3}$$

be the discrete-time hazard of dropout at time j, let

$$\pi_j = \prod_{t=1}^{j} (1 - \lambda_t) = P(R_j = 1) \tag{3.4}$$

be the corresponding unconditional probability of being observed at time j, and let the weight matrix be $W_\pi = diag(R_1/\pi_1, ..., R_M/\pi_M)$. The λ's and thus π_j and W_π can be consistently estimated by logistic regression if the MAR assumption holds. Then the inverse probability weighted GEE (WGEE) has estimating equation

$$\sum_{i=1}^{N} U_k^W(\bar{L}_{iJ}) = \sum_{i=1}^{N} \frac{\partial \mu_i^T}{\partial \beta_k} V_i^{-1} W_i (Y_i - \mu_i) = 0. \tag{3.5}$$

For longitudinal data with dropout that is MAR, WGEE provides consistent estimates of the parameters β^* provided that the estimates of the π_j's are consistent, since then $E[R_j Y_j/\pi_j] = E[Y_j]$.

3.3.4 Regression-based sequential imputation, for dropout that is MAR

Alternatively, Paik (1997) [44] defined a sequential regression approach for imputing the missing outcome values. Consider a subject with $J = j$, so that Y_k is missing for $k > j$. The idea is to define a set of parametric imputation models $m_k^j(\bar{L}_j) = E[Y_k|\bar{L}_j]$. By the MAR assumption, $E[Y_k|R_k = 0, \bar{L}_j] = E[Y_k|R_{j+1} = 1, \bar{L}_j]$. Hence we may use observed data to construct a consistent estimate $\hat{m}_k^j$, and then use the estimate to impute missing values as $\hat{Y}_k = \hat{m}_k^j(L_j)$.

We give a formal version of Paik's sequential imputation algorithm, which will make the relation with Bang and Robins (2005) [2] more explicit :

Algorithm 1: Paik's sequential mean imputation

Result: For a given k, the completed data $\hat{Y}^I_{ik}$, in which all missing values at time k have been imputed.

1 **Initialize:** Identify all subjects i with $J_i \geq k-1$. Use these data to regress the observed values of Y_k on $\bar{L}_{k-1}$, to obtain a consistently estimated model $\hat{m}^{k-1}_k(\bar{L}_{k-1})$. For subjects with $J_i \geq k-1$ let

$$\hat{Y}^{k-1}_{ik} = \begin{cases} \hat{m}^{k-1}_k(\bar{L}_{i,k-1}) & \text{if } J = k-1 \\ Y_{ik} & \text{if } J > k-1 \end{cases}$$

2 **For $\mathbf{s = k-2, ..., 1}$ sequentially:** Regress the values of $\hat{Y}^{s+1}_{ik}$ on $\bar{L}_{is}$ to obtain a consistently estimated model $\hat{m}^s_k(\bar{L}_s)$. For all subjects with $J_i \geq s$ let

$$\hat{Y}^s_{ik} = \begin{cases} \hat{m}^s_k(\bar{L}_s) & \text{if } J_i = s \\ \hat{Y}^{s+1}_{ik} & \text{if } J_i > s \end{cases}$$

3 **Final step:** Output the completed data $\hat{Y}^I_{ik} = \hat{Y}^1_{ik}$

The above imputation is iterated through times $k = 2, \ldots, M$, requiring $M(M-1)/2$ estimated models, and then estimating equation (2.1) is solved using the completed data:

$$\sum_{i=1}^{N} U_k(\hat{L}^I_{iM}) = \sum_{i=1}^{N} \frac{\partial \mu_i^T}{\partial \beta_k} V_i^{-1}(\hat{Y}^I_i - \mu_i) = 0. \tag{3.6}$$

Under regularity conditions, this procedure gives consistent estimates for the parameters β^* whenever the mean models m^s_k are consistent and the data are MAR, as then the imputed estimating equation is a uniformly consistent estimator of the estimating equations $E[U(\bar{L}_M, \beta)]$.

3.4 Doubly robust estimators for cross-sectional data

Doubly robust estimators combine both an imputation model and estimated IPW's into an augmented inverse probability weighted (AIPW) estimator, in such a way that the estimator is consistent if either the imputation model or the IPW model is consistent. For expository reasons we give the cross-sectional form of these AIPW estimators here, as they are simpler and easier to understand.

AIPW doubly robust estimating equations

For cross sectional data (i.e. M=1), the augmented inverse probability weighted (AIPW) form [51] of an estimating equation U is

$$\sum_{i=1}^{N}\left(\frac{R_i}{\hat{\pi}_i}U_i+(1-\frac{R_i}{\hat{\pi}_i})\hat{H}(X_{i0})\right)=0 \tag{3.7}$$

where $\hat{H}(X_0)$ (the imputation model) is an estimate of $E[U|X_0]$.

It is easy to see that such an AIPW estimator is doubly robust if the data are MAR. If $\hat{\pi}$ is a consistent estimate of π, then $E[R]=\pi$, and it follows from the MAR assumption that $E[RU/\pi]=E[R]E[U]/\pi$. Similarly, $E[(1-R/\pi)H(X_0)]=0$, so that (3.7) is consistent for the estimating equation $E[U]=0$. If $\hat{H}(X_0)$ is consistent for $E[U|X_0]$, write (3.7) in the form

$$\sum_{i=1}^{N}\left(\hat{H}(X_{i0})+\frac{R_i}{\hat{\pi}_i}(U_i-\hat{H}(X_{i0})\right)=0 \tag{3.8}$$

which is again consistent for the estimating equation $E[U]=0$.

Regression-based doubly robust estimating equations

Scharfstein, Rotnitzky and Robins (1999) [58] showed that a doubly robust AIPW estimating equation (3.8) can sometimes be written in a regression-based form. Construct the estimate

$\hat{H}(X_0, \hat{\pi})$ by regressing outcome $R_i U_i$ on predictors $(R_i X_{i0}, R_i \pi_i^{-1})$ using a generalized linear model. Then the score equation for π^{-1} in this model is exactly

$$\sum_{i=1}^{N} \frac{R_i}{\hat{\pi}_i} \left(U_i - \hat{H}(X_{i0}) \right) = 0$$

and so the estimating equation $\sum_{i=1}^{N} \hat{H}(X_0, \hat{\pi}) = 0$ is implicitly of the form (3.8) and thus of the form (3.7).

3.5 Doubly robust estimators for longitudinal data

Optimal longitudinal AIPW estimating equations

In the longitudinal setting with MAR dropout, Tsiatis (2006) [68] showed that any consistent and asymptotically normal estimator of β^* using the observed data solves an AIPW estimating equation of the form

$$\sum_{i=1}^{N} \left(\frac{C_{i,M}}{\pi_{i,M}} U(\bar{L}_{iM}) + \sum_{j=1}^{M-1} \left(\frac{C_{i,j} - \lambda_{i,j+1} R_{i,j}}{\pi_{i,j+1}} \right) H^j(\bar{L}_{ij}) \right) = 0 \tag{3.9}$$

where H^j is an arbitrary function. The choice $H^j = E(U(\bar{L}_M)|\bar{L}_j)$ yields the estimator with the smallest variance. Tsiatis, Davidian and Cao (2011) [69] proposed an estimator that attains the locally smallest asymptotic variance, using a linear mixed effects model to obtain estimates $\hat{H}^j$, and introducing an additional parameter that solves a minimization problem. More generally, when estimates are substituted for the unknown quantities λ_j (and thus π_j) and $H^j = E(U(\bar{L}_M)|\bar{L}_j)$, the solution $\hat{\beta}^{AIPW}$ to (3.9) has the following properties [68] :

1. $\hat{\beta}^{AIPW}$ is consistent for β^* and asymptotically normal if either the missingness models $\hat{\lambda}_j$ or the imputation models $\hat{H}^j$ are consistent.

2. If the imputation models are consistent, then $\hat{\beta}^{AIPW}$ has smaller asymptotic variance than

the corresponding IPW estimator.

3. If both sets of models are consistent, $\hat{\beta}^{AIPW}$ has the smallest asymptotic variance among all doubly robust estimators of β.

4. For given estimators $\hat{H}^j$ and $\hat{\lambda}_j$, the asymptotic variance of $\hat{\beta}^{AIPW}$ depends on their probability limits (which may or may not be consistent), but not their asymptotic variance.

Seaman's doubly robust GEE

Seaman and Copas (2009) [63] proposed a two-step procedure for the case where U is a GEE of the form (3.1). In particular, for $j \leq J$, take

$$\hat{H}^j(\bar{L}_j) = \frac{\partial \mu_i^T}{\partial \beta_k} V_i^{-1} (\hat{Y}_k^j - \mu) \tag{3.10}$$

where $\hat{Y}_k^j$ is from Paik's sequential regression as in (1). For $j > J$, H^j can be taken to be 0, as the weights in (3.9) are then 0. Then, Newton-Raphson is used to solve the AIPW estimating equations

$$\sum_{i=1}^{N} \left(\frac{C_{i,M}}{\hat{\pi}_{i,M}} U(\bar{L}_{iM}) + \sum_{j=1}^{M-1} \left(\frac{C_{i,j} - \hat{\lambda}_{i,j+1} R_{i,j}}{\hat{\pi}_{i,j+1}} \right) \hat{H}^j \right) = 0. \tag{3.11}$$

Bang and Robins sequential doubly robust imputation for Y_M

The most commonly used approach was introduced by Bang and Robins (2005) [2], for the particular case where the estimand of interest is $E(Y_M)$. The algorithm uses sequential estimation to impute the values of Y_M. However, it includes $\hat{\pi}_j^{-1}$ as a covariate in each imputation model in order to achieve double robustness, similar to the ideas in section 3.4. The algorithm also differs from Paik's sequential imputation in that it uses imputed values $\hat{Y}_M$ as outcomes in each estimation step, even when observed values are available.

Algorithm 2: Bang & Robins DR sequential mean imputation for Y_M

Result: The completed data Y_{iM}^{BR}, in which all missing values at time M have been imputed, using a doubly robust method.

1 **Preliminary step:** Estimate $\hat{\pi}_2, ..., \hat{\pi}_M$ by maximum likelihood.

2 **Intialize:** Identify subjects with $J_i \geq M-1$. Regress the observed values of Y_{iM} on $\bar{L}_{i,M-1}, \hat{\pi}_M$, to obtain a consistently estimated model $\tilde{m}_M^{M-1}(\bar{L}_{M-1}, \hat{\pi}_M)$. For subjects with $J_i \geq M-1$ let $\tilde{Y}_{i,M}^{M-1} = \tilde{m}_M^{M-1}(\bar{L}_{i,M-1}, \hat{\pi}_{i,M})$.

3 **For s $= \mathbf{M-2}, ..., \mathbf{1}$ sequentially:** Regress the values of $\hat{Y}_M^{s+1}$ on $\bar{L}_s, \hat{\pi}_{s+1}$ to obtain a consistently estimated model $\tilde{m}_M^s(\bar{L}_s, \hat{\pi}_{s+1})$. For all subjects with $J_i \geq s$ let $\tilde{Y}_{i,M}^s = \tilde{m}_M^s(\bar{L}_{i,s}, \hat{\pi}_{i,s+1})$

4 **Final step:** Output the completed data $\hat{Y}_{iM}^{BR} = \tilde{Y}_{i,M}^1$.

For the special case $U(\bar{L}_M) = Y_M - \beta$, Bang and Robins (2005) [2] show that the estimator of $E(Y_M)$ given by $\sum_{j=1}^N \tilde{Y}_{i,M}^1$ satisfies an estimating equation of the form (3.9), and thus is doubly robust, using arguments as in section 3.4.

3.6 Longitudinal AIPW estimating equations in imputation form

We show that the AIPW estimating equations (3.9) can be written in a form that applies the full data estimating equations U to data which has been completed using a set of doubly-robust imputed observations, for the special case when $U(\bar{L}_M)$ is a GEE. This approach has the advantage that standard software can be used to solve for the estimates β^{AIPW}, and provides a framework for flexible construction of DR estimators.

3.6.1 The optimal AIPW equation in imputation form

Here we show that for any DR estimator written in AIPW form (3.9), there is an equivalent estimator in substitution (plug-in) form

$$\sum_{i=1}^{N} U(\hat{L}_{i,M}^{DR}) = \sum_{i=1}^{N} \frac{\partial \mu_i^T}{\partial \beta_k} V_i^{-1} (\hat{Y}_i^{DR} - \mu_i) = 0 \tag{3.12}$$

where $\hat{Y}_i^{DR}$ is a corresponding doubly robust estimator of the full data Y.

First, let $\pi_{M+1} = \pi_M$ and $\lambda_1 = \lambda_{M+1} = 0$. Then (3.9) can be rewritten as

$$\sum_{i=1}^{N} \sum_{j=1}^{M} \left(\frac{C_{ij} - \lambda_{j+1} R_{ij}}{\pi_{j+1}} \right) H^j(\bar{L}_{ij}) = 0. \tag{3.13}$$

Also, it is easy to see that $\sum_{j=1}^{M} (C_j - \hat{\lambda}_{j+1} R_j)/\hat{\pi}_{j+1} = 1$, since $\hat{\pi}$ satisfies (3.4).

Next, as in Seaman and Copas (2009) [63], for $U(\bar{L}_M)$ equals to a GEE of the form (3.1) we have

$$H^j = \frac{\partial \mu^T}{\partial \beta} V^{-1} (E[Y|\bar{L}_j] - \mu).$$

Then substituting into (3.13), switching the order of summation, and recognizing that $\sum_{j=1}^{M} (C_{ij} - \hat{\lambda}_{j+1} R_{ij})/\hat{\pi}_{j+1} = 1$ we obtain that the efficient doubly robust estimating equations can be written as:

$$\frac{\partial \mu^T}{\partial \beta} V^{-1} \left\{ \left(\sum_{j=1}^{M-1} \left(\frac{C_j - \lambda_{j+1} R_j}{\pi_{j+1}} \right) E[Y|\bar{L}_j] \right) - \mu \right\} \tag{3.14}$$

Finally, we can write

$$\hat{Y}^{DR} = \left\{ \left(\sum_{j=1}^{M-1} \left(\frac{C_j - \hat{\lambda}_{j+1} R_j}{\hat{\pi}_{j+1}} \right) \hat{E}[Y|\bar{L}_j] \right) \right\},$$

recognizing that the right hand side is representation (3.13) of a general DR estimator of $E[Y]$ in its AIPW form.

Form (3.12) has the advantage that, once the values of $\hat{Y}_i^{DR}$ are obtained using any preferred

method, standard software for the estimating equations U can be used to solve for the doubly robust estimates β^{AIPW}. From the theory in Tsiatis (2006), we can be assured that any efficient doubly robust estimator can be written in this form. Although we have given the argument in the special case of generalized estimating equations, the argument holds for any estimating equation that is linear in Y.

We may also use the considerations above to provide a simple demonstration that a general estimator $\hat{\beta}^{AIPW}$ is doubly robust from below equations.

$$E\left((\frac{C_{i,M}}{\pi_{i,M}})U(\bar{L}_{iM}) + \sum_{j=1}^{M-1}\left(\frac{C_{i,j}-\lambda_{i,j+1}R_{i,j}}{\pi_{i,j+1}}\right)H^j(\bar{L}_{ij})\right) =$$
$$E\left(U(\bar{L}_{iM}) + (\frac{C_{i,M}}{\pi_{i,M}}-1)U(\bar{L}_{iM}) + \sum_{j=1}^{M-1}\left(\frac{C_{i,j}-\lambda_{i,j+1}R_{i,j}}{\pi_{i,j+1}}\right)H^j(\bar{L}_{ij})\right) =$$
$$E\left[U(\bar{L}_{iM})\right] + E\left[\sum_{j=1}^{M-1}\left(\frac{C_{i,j}-\lambda_{i,j+1}R_{i,j}}{\pi_{i,j+1}}\right)\left(H^j(\bar{L}_{ij}) - U(\bar{L}_{iM})\right)\right] = 0$$

Noting that $R_j \perp H^{j-1}$ by the MAR assumption, on the one hand, if the probabilities in (3.9) are correct, then $E[C_j] = \lambda_{j+1}\pi_j$; on the other hand, if the $\hat{H}^j$ are consistent, then $E[\hat{H}^j(\bar{L}_j) - U(\bar{L}_M)] = 0$. Thus, in either case (3.9) is consistent for the complete data estimating equation $E[U(\bar{L}_M, \beta)] = 0$.

3.7 Imputation approaches to DR estimating equations

Here we apply the doubly robust imputation framework to construct two particular DR estimators for longitudinal data.

3.7.1 Doubly robust sequential imputation for longitudinal GEE's

An immediate consequence of (3.12) is that for any complete-data linear GEE $U(\bar{L}_M)$, a doubly robust estimator of β^* in the case of MAR longitudinal dropout is given by:

1. For each subject i, impute a doubly robust full data vector using the AIPW estimate of Y_{ik}, $k = 1, \ldots M$

$$\hat{Y}_{ik}^{AIPW} = \frac{C_{ik}Y_{ik}}{\hat{\pi}_{ik}} + \sum_{j=1}^{k-1}(\frac{C_{ij} - \hat{\lambda}_{ij+1}R_j}{\hat{\pi}_{ij+1}})\hat{m}_k^j(\bar{L}_{ij}) \tag{3.15}$$

   with the models $\hat{m}_k^j(\cdot)$ obtained by Paik sequential regression as in section 3.3.4.

2. Substitute $\hat{Y}_i^{AIPW}$ for Y_i^{DR} in (3.12), which may then be solved using standard software to obtain a doubly-robust estimate of β^*.

We denote the final estimator as $\hat{\beta}^{AIPW-I}$, to indicate AIPW estimating equations with sequential imputation. Note that this is equivalent to the doubly robust estimator in Seaman and Copas (2009) [63].

3.7.2 Computationally simpler baseline $\times$ time imputation for DR GEE's

Form (3.12) of the optimal AIPW estimating equations can be used to motivate a computationally simpler approach using only baseline covariates and time for the imputation models, at the price of potential loss in efficiency. We start with (3.14), and note that $E[Y_k|\bar{L}_j] = Y_k$ for $j \geq k$. Hence we can write the k^{th} component of $\hat{Y}^{DR}$ as

$$\hat{Y}_k^{DR} = \sum_{j=k}^{M}(\frac{C_j - \lambda_{j+1}R_j}{\pi_{j+1}})Y_k + \sum_{j=1}^{k-1}(\frac{C_j - \lambda_{j+1}R_t}{\pi_{j+1}})\hat{Y}_k^j.$$

We next take $\hat{Y}_k^j$ to be an estimate of $E[Y_k|X_0]$ independent of j, where X_0 contains baseline covariates and time. We are now outside the set of possible efficient estimators, except in the special case where $E[Y_k|\bar{L}_j] = E[Y_k|X_0]$. Then $\hat{Y}_k^{DR}$ can be written as

$$\hat{Y}_k^{AIPW-S} = \frac{R_k}{\hat{\pi}_k}Y_k + (1 - \frac{R_k}{\hat{\pi}_k})\hat{Y}_k, \tag{3.16}$$

since $\sum_{j=k}^{M}((C_j - \hat{\lambda}_{j+1}R_j)/\hat{\pi}_{j+1}) = R_k/\hat{\pi}_k$.

We may estimate, for example, a single mixed effects model $\hat{m}(X_0,t)$ using all the observed responses as outcomes, regressed on any always observed covariates, such as baseline covariates X_0 and time. Then $\hat{Y}_{ik} = \hat{m}(X_{i0},t_k)$.

To obtain the doubly robust estimator, substitute $\hat{Y}_i^{AIPW-S}$ for $\hat{Y}_i^{DR}$ in (3.12), and solve for $\hat{\beta}$ using standard software. We denote the final estimator $\hat{\beta}^{AIPW-S}$ as the simplified imputation model only uses baseline covariates and time.

If the covariates are sufficient to render the data MAR and the model for m is correct, then $\hat{\beta}^{AIPW-S}$ will be doubly robust. Note that only one mixed effects model is estimated for the imputation instead of $M(M-1)/2$ models as in the above two approaches. The cost is a potential loss of efficiency, and a stronger assumption regarding the MAR conditions.

3.7.3 Bang and Robin's imputation for longitudinal GEE's

We apply (3.12) to extend Bang and Robin's approach to longitudinal estimating equations. Using the algorithm in section 3.5, set $M = k$ and compute $\hat{Y}_{ik}^{BR}$ sequentially for $k = 2,...,M-1$. Substitute $\hat{Y}_{ik}^{BR}$ for Y_i^{DR} in (3.12). Equation (3.12) may then be solved using standard software to obtain a doubly-robust estimate of β^*. We denote the final estimator $\hat{\beta}^{BR}$, since this incorporates the Bang and Robins estimator in the AIPW equations.

3.8 Simulations

We use simulation to investigate the performance of three DR imputation-based methods for a normal longitudinal generalized estimating equation under MAR dropout: **AIPW-I,** based on augmented inverse probability weighting using sequential imputation (section 3.7.1); **AIPW-S**, a computationally simpler version of the sequential impuation approach using only baseline covariates (section 3.7.2); and **BR**, the extension of Bang and Robins' regression-based estimator (section 3.7.3). We study two different estimands: $E[Y_M]$, and a vector of regression coefficients $\hat{\beta}$.

For comparison, we include two traditional regression approaches, maximum likelihood by generalized least squares (**GLS**) using an unstructured covariance matrix and generalized estimating equations using an independence working covariance matrix (**GEE)**. We also include correction for dropout using Paik's sequential mean imputation (**PAIK**, section 3.3.4), and using inverse probability weighting (**WGEE**, section 3.3.3). We computed a point estimate and a bootstrap estimate of variance for each method with an associated 95% normal-theory confidence interval. We consider both a moderate dropout rate ($\approx 30\%$) and an extreme dropout rate ($\approx 50\%$). There are four situations, according to which of the missingness models and/or the imputation models are either correctly specified or are misspecified.

3.8.1 The data generating model, the primary estimand, and specification of correct and incorrect imputation models

Longitudinal responses Y_{ij} were generated from the mixed effects model

$$Y_{ij} = \quad \alpha_i + b_i t + \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_2 \times t + \varepsilon$$

with sample size $n = 500$, and three time points $t = 1,2,3$. The bootstrap sample size was 300. The primary estimand is either $E[Y_3]$ or the vector of regression coefficients.

Here, α_i and b_i were random intercepts and slopes from a bivariate normal with mean $\mu = \binom{1}{6}$ and covariance $\Sigma = \left(\begin{smallmatrix} 0.3 & 0.1 \\ 0.1 & 0.2 \end{smallmatrix}\right)$. The covariates were generated as $x_1 \sim$ N(5,1), $x_2 \sim$ Bernoulli(0.5), with $\varepsilon \sim$ N(0, 1). The data generating coefficients were $\beta_0 = 0.5$, $\beta_1 = 2$, $\beta_2 = -0.25$, $\beta_3 = -6$. Thus the expectation of Y_{ij} was $11.375, 14.375, 17.375$ for $j = 1,2,3$.

For all methods, the correctly specified mean model includes categorical time, (x_1, x_2), and the interaction of time and x_2. For sequential imputation methods, the correct models are of the form $Y_j \sim Y_1 + ... + Y_{j-1} + x_1 + x_2 + e$ for $j = (2,3)$. For all methods, the misspecified imputation model excludes x_2 and the corresponding interaction terms.

3.8.2 The dropout generating model and specification of correct and incorrect models for missingness

We generated dropouts according to the logistic regression models

$$logit(\lambda_2) = \gamma_2 + 0.5 \times y_1 - 2 \times x_2 + e$$
$$logit(\lambda_3) = \gamma_3 + 0.1 \times y_1 + 0.2 \times y_2 - 4 \times x_2 + e,$$

where λ_{ij} is the hazard of dropout as defined in formula (2.3). For the moderate dropout scenario, $\gamma_2 = -7.625, \gamma_3 = -5.225$ and the empirical mean dropout rates from 500 Monte Carlo repeats were 10.4% for Y_2 and 30.1% for Y_3. For the high dropout scenario, $\gamma_2 = -6.5, \gamma_3 = -3.5$ and the mean dropout rates were 22% and 51% for Y_2 and Y_3, respectively. Misspecified models for the missing mechanism omitted x_2.

3.8.3 Performance metrics and sample sizes

We report Monte Carlo estimates (sample size 500) for the bias, standard deviation, and root mean square error (RMSE) of the point estimate, and for the confidence interval, the coverage probability, and mean interval score of Gneiting and Raftery (2007) [23] given by:

$$S(\hat{l}, \hat{u}, \theta) = (\hat{u} - \hat{l}) + \frac{2}{\alpha}\left((\hat{l} - \theta)\mathbb{1}\{\theta < \hat{l}\} + (\theta - \hat{u})\mathbb{1}\{\hat{u} < \theta\}\right), \tag{3.17}$$

where θ is the true parameter of interest, and the interval limits are $(\hat{l}, \hat{u})$ and $\mathbb{1}\{\}$ denotes the indicator function. A GLS model with correct covariates, categorical time, and unstructured correlation matrix will serve as the gold standard in our comparisons.

Results for estimating $E(Y_3)$, under moderate dropout

The top left panel of Table 3.1 shows results for estimating $E(Y_3)$ under moderate dropout when both imputation and dropout models are correct. The "gold standard" GLS (correct maximum likelihood) model had a bias of -0.03. GEE had a worse bias of 0.09 as expected. All three doubly robust methods had bias smaller than 0.01, outperforming the gold standard. Paik's imputation performed well, and WGEE performed fairly with bias and less efficient. In all moderate dropout scenario simulations, around 1% of Monte Carlo repeats for WGEE failed to converge, and 4% of bootstrap estimates failed to converge, resulting in substantial standard errors in some cases for WGEE.

When the imputation model was correct, but the dropout model was misspecified (top right panel), all doubly robust methods had acceptable bias, ranging from 0.037 to -0.005, with RMSEs, coverage probabilities, interval scores, Monte Carlo standard deviations, and average estimated standard errors all similar to the gold standard. As expected, WGEE with an incorrect dropout model had a relatively larger bias and was inefficient.

When the dropout model was correct, but the imputation model was misspecified (bottom left panel), AIPW-I and AIPW-S still had acceptable bias and coverage probabilities, with some loss of efficiencies compared to prior scenarios. The B&R estimator had an unacceptably large bias and low coverage probability (85%), although it performed better than the non-doubly robust regression methods, which performed disastrously. WGEE had a bias of -0.22 and coverage probability of 68%; the incorrect GLS model had a coverage probability of almost 0, while a GEE model with wrong mean structure and wrong working correlation matrix performed the worst.

The bottom right panel showed results when both models were incorrect. The doubly robust methods performed better than the regression methods in every measure, although all methods had worse performance than in other scenarios.

Table 3.1: Comparisons of $E(Y_3)$ among methods in six evaluations: Bias, Root mean square error (RMSE), interval scores (Ints), coverage probability (Covp), Monte Carlo standard deviation (MCSD) and average standard errors (Ave SE), from 500 simulation runs and for $B = 300$ bootstrap.

	Bias	RMSE	Ints	Covp	MCSD	Ave SE	Bias	RMSE	Ints	Covp	MCSD	Ave SE
	Y correct P correct						Y correct P incorrect					
AIPW-I	-0.01	0.30	1.39	0.95	0.30	0.31	-0.00	0.30	1.37	0.95	0.30	0.30
B & R	-0.01	0.30	1.36	0.95	0.30	0.30	-0.01	0.30	1.36	0.95	0.30	0.30
AIPW-S	-0.01	0.31	1.39	0.95	0.31	0.31	0.04	0.31	1.39	0.95	0.31	0.31
Paik	-0.01	0.30	1.35	0.95	0.30	0.30	-0.01	0.30	1.35	0.95	0.30	0.30
WGEE	-0.03	0.36	1.84	0.92	0.36	0.31	0.09	0.37	1.81	0.90	0.36	0.31
GLS	-0.03	0.30	1.36	0.94	0.30	0.30	-0.03	0.30	1.36	0.94	0.30	0.30
GEE IND	-0.09	0.31	1.38	0.94	0.29	0.30	-0.09	0.31	1.38	0.94	0.29	0.30
	Y incorrect P correct						Y incorrect P incorrect					
AIPW-I	-0.01	0.31	1.44	0.95	0.31	0.31	-0.68	0.74	7.65	0.42	0.31	0.31
B & R	0.31	0.45	2.34	0.85	0.33	0.33	-0.62	0.69	6.42	0.50	0.31	0.31
AIPW-S	-0.04	0.38	1.72	0.95	0.38	0.36	-0.62	0.71	6.44	0.55	0.36	0.36
Paik	-0.65	0.72	6.95	0.45	0.31	0.32	-0.65	0.72	6.95	0.45	0.31	0.32
WGEE	-0.22	0.72	6.92	0.63	0.68	0.34	-1.41	1.56	33.98	0.11	0.66	0.34
GLS	-1.14	1.17	22.53	0.04	0.29	0.31	-1.14	1.17	22.53	0.04	0.29	0.31
GEE IND	-2.18	2.20	63.94	0.00	0.30	0.31	-2.18	2.20	63.94	0.00	0.30	0.31

Results for estimating regression coefficients β, under moderate dropout

Table 3.2 presents results for estimating the vector of six regression coefficients since the analysis model used a categorical time profile. Performance metrics, averaged across the coefficients, are : (1) average percent bias, (2) z-score (among estimators within simulation condition) of the relative RMSE's, (3) z-score of the interval score, (4) average coverage probability. Details are given in Appendix A.

With both models correctly specified, both the GLS and GEE estimating models gave an average 1% bias, AIPW-I and AIPW-S gave an average percent bias less than 0.05%, and B&R gave an average percent bias of 2%. The AIPW-I and AIPW-S methods had similar standardized RMSEs, standardized interval scores, and average coverage probabilities as the correct GLS and mean imputation methods, which were relatively better than other methods.

When the imputation model was correctly specified, but the dropout model was not, AIPW-

I and AIPW-S again performed as well as the gold standard on all metrics. The B&R method had an average 2.9% bias, which was not as good as other DR methods but not unacceptable. Notably, in these situations, the AIPW-S method had no actual loss in efficiency compared to more complex approaches.

When the dropout model was correctly specified, but the imputation model was not, the regression-based B&R method performed unacceptably, with a 23% bias and a large standardized RMSE and interval score. The AIPW-I and AIPW-S methods retained their good performance, with average percent bias less than 0.8% . Notably, the AIPW-I estimator had comparable efficiency to the correct regression methods.

When both models were misspecified, regression methods had average percent bias of

Table 3.2: Comparisons of $\hat{\beta}_p$ among methods in four evaluations: Average % absolute values of bias (*Bias**), Standardized RMSE in average (*RMSE**), Standardized interval scores in average (*Ints**) and average of coverage probabilities (*Covp**), from 500 simulation runs and for $B = 300$ bootstrap.

	Bias*	RMSE*	Ints*	Covp*	Bias*	RMSE*	Ints*	Covp*
	Y correct P correct				Y correct P incorrect			
AIPW-I	0.00	-0.45	-0.48	0.94	0.00	-0.45	-0.48	0.94
B & R	0.02	-0.20	-0.10	0.96	0.03	0.43	0.27	0.94
AIPW-S	0.00	-0.45	-0.48	0.94	0.01	-0.45	-0.48	0.94
Paik	0.00	-0.46	-0.48	0.94	0.00	-0.46	-0.48	0.94
WGEE	0.07	0.05	0.22	0.92	0.10	0.38	-0.26	0.92
GLS	0.01	-0.46	-0.48	0.94	0.01	-0.46	-0.48	0.94
GEE IND	0.01	-0.45	-0.48	0.92	0.01	-0.45	-0.48	0.92
	Y incorrect P correct				Y incorrect P incorrect			
AIPW-I	0.00	-0.44	-0.47	0.95	0.04	-0.35	-0.38	0.58
B & R	0.23	0.44	0.44	0.92	0.55	1.96	2.20	0.67
AIPW-S	0.01	-0.40	-0.45	0.92	0.04	-0.34	-0.40	0.67
Paik	0.10	-0.14	-0.18	0.24	0.10	-0.14	-0.18	0.24
WGEE	0.36	0.89	0.86	0.26	0.48	1.22	1.44	0.15
GLS	0.43	1.04	1.22	0.12	0.43	1.04	1.22	0.12
GEE IND	0.47	1.18	1.41	0.08	0.47	1.18	1.41	0.08

Details about how to derive the adjusted evaluations were in **??**
For *Bias**, *RMSE** and *Ints**, the lower the value, the better the performance.

For *Covp**, the higher the value, the better the performance.

more than 40% and average coverage probabilities lower than 15%. In contrast, the AIPW-I and AIPW-S methods had average percent bias less than 5% and average coverage probabilities higher than 55%. The B&R approach had the worst performance, indicating that it was not robust in this situation.

3.8.4 Extreme scenario

We conducted a similar comparison of these estimators in the extreme scenario, where the mean dropout rates for Y_2 was 22% and for Y_3 was 51%. Other simulation parameters were kept the same. Results were qualitatively similar to the moderate dropout scenario, with generally good performance of the doubly robust AIPW-I and AIPW-S estimators compared to the GLS estimator and much worse performance of the BR estimator. In some scenarios, the loss of efficiency of AIPW-S relative to AIPW-I became apparent, although its performance was still good. WGEE was very problematic due to convergence issues. Details are given in the Appendix A.

3.8.5 Summary of simulation results

The two doubly robust methods, AIPW-I and AIPW-S, demonstrated performance comparable to efficient MLE estimators whenever the imputation models were correctly specified in these simulations. When the imputation model was incorrect, but the dropout model was correct, the two AIPW doubly robust estimators still performed well, with low bias and good coverage probabilities, although with some loss of efficiencies. By contrast, the purely regression-based methods (GLS, GEE, Paik imputation) performed unacceptably, as expected. When both models were incorrectly specified, the doubly robust AIPW estimators outperformed the regression-based methods. WGEE did not perform competitively, perhaps due to convergence issues.

Among doubly robust estimators, the BR approach was not competitive with the two AIPW estimators in these simulations, either in its original version or as extended here to apply to general

estimating equations. This was especially true if the imputation model was misspecified, where the BR estimator was not acceptable. Comparing the AIPW methods, AIPW-I performed the best. The simplified AIPW-S estimator was comparable in the moderate (30%) dropout scenario but had increased bias and less efficiency in the extreme (50%) dropout scenario, although it was still acceptable.

3.9 Application to the MCI trial

The MCI trial was described in section 4.3.2. The primary outcome of the trial was time to progression to Alzheimer's disease (AD), and the main conclusion was that Vitamin E had no benefit, while Donepezil had some benefits over placebo at 12 months but not at 36 months. This is in accordance with the known symptomatic benefits of Donepezil. The trial showed no benefit in secondary analyses comparing within-patient change on the two cognitive measures Mini-Mental State Examination (MMSE) and the Alzheimer's Disease Assessment Scale-Cognitive Subscale 11 (ADAS-Cog 11) at 36 months.

Here, for simplicity, only the Donepezil group and placebo group are used. The two groups had similar demographic and clinical characteristics at baseline. MMSE and the ADAS-Cog 11 are used as our repeated measures outcomes. Higher MMSE and lower ADAS-Cog 11 indicate improved cognition. These measures were assessed at baseline, 12 months, 24 months, and 36 months. The missing rates for the Donepezil group and the placebo group were 26.1% and 16.6% at 12 months, 34.0% and 29.3% at 24 months, and 42.7% and 32.0% at 36 months.

We compared results from the two doubly robust estimators, AIPW-I and AIPW-s, and an MMRM model with an unstructured correlation matrix and discrete time, fitted by GLS. These models included categorical time, baseline outcome, arm, and the interaction between arm and time, similar to the standard analysis model for AD trials. The estimand of interest was the interaction between arm and time, which parameterizes the treatment effect at each time point.

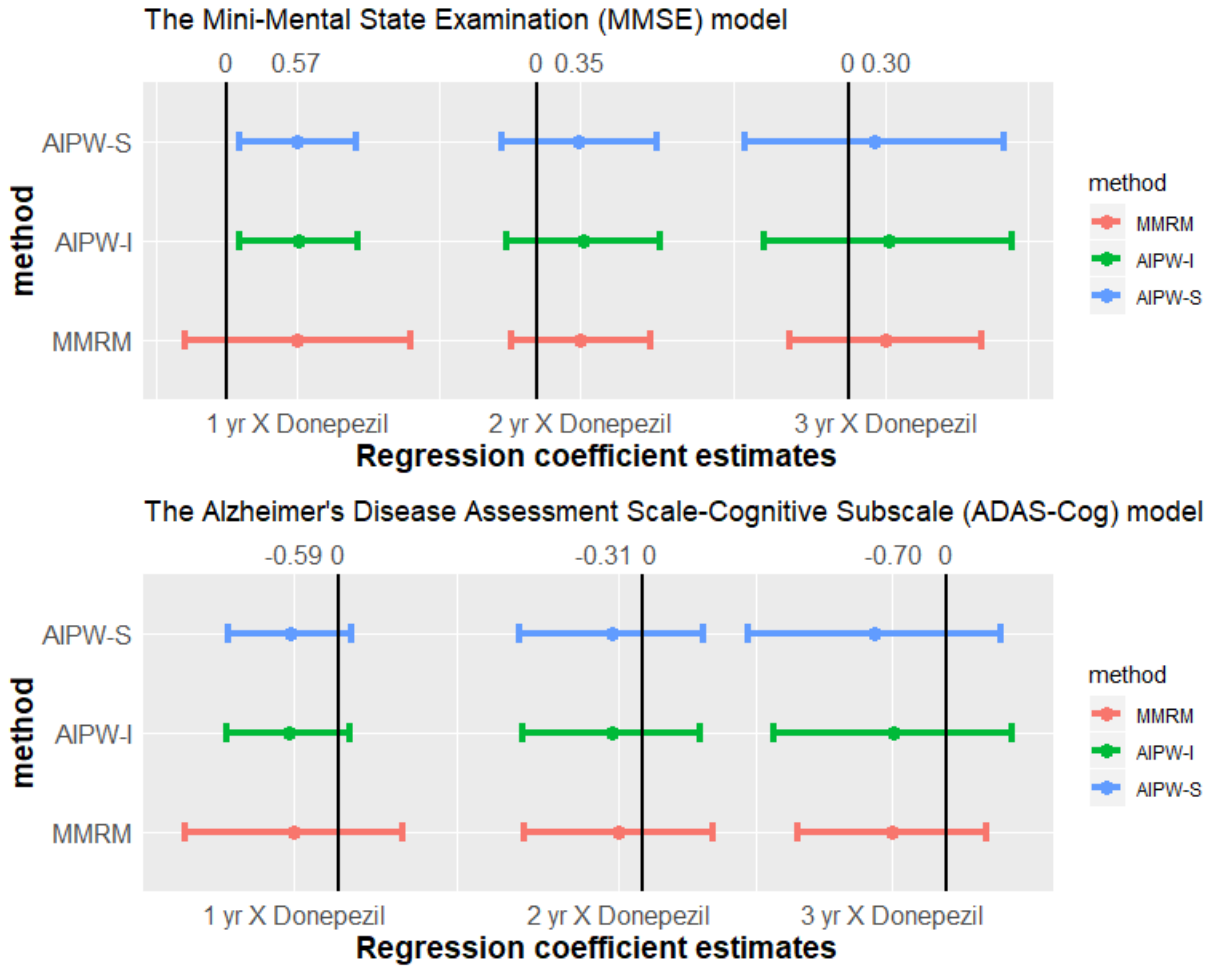

Figure 3.2: Regression coefficients of interaction terms

The dropout models for the doubly robust methods included arm and historical Y as covariates. Normal theory confidence intervals were constructed using bootstrap estimates of variance; the bootstrap sample size was 500.

Figure 3.2 displays the point estimates and 95% confidence intervals at each time point for the three approaches. Using the MMRM approach, the Donepezil arm showed an estimated benefit of 0.574, 0.346, and 0.302 additional points on the MMSE over the placebo arm at years 1, 2, and 3, respectively (top panel). However, none of these differences were significant at the 5% level. The two doubly robust methods derived similar estimates for the regression coefficient of the interaction terms; however, the standard errors at earlier time points were smaller, achieving a

significant treatment effect at one year. The standard errors from DR methods became larger over time. However, we observed a similar phenomenon for the ADAS-Cog 11 in the lower panel, not attaining statistical significance. This is consistent with the known efficacy profile of Donepezil and the known poor performance of the ADAS-Cog 11 in an MCI population. When we mimic a two-year trial using only data from the first two years to estimate the year one and year two effects, the differences between methods are even more apparent (Table 3.3).

Table 3.3: Donepezil trial, MMSE outcome: estimated time and time x treatment effects at one and two years, by different estimators. Data from years 1 and 2.

	1-year		2-yr		1-yr:Donepezil		2-yr:Donepezil	
	Beta	SE	Beta	SE	Beta	SE	Beta	SE
MMRM	-0.618	0.235	-1.099	0.216	0.486	0.350	0.342	0.321
AIPW-I	-0.730	0.159	-1.250	0.215	0.583	0.245	0.368	0.314
AIPW-S	-0.725	0.160	-1.233	0.215	0.572	0.247	0.335	0.321

Standard errors for AIPW-I and AIPW-S were derived from 500 times bootstrap.

3.10 Discussion

In this chapter, we have shown that any doubly robust estimator that can be written in AIPW form can also be written in a substitution (plug-in) form, for estimating equations that are linear in the data. We use this framework to propose an imputation approach, AIPW-I, to construct a doubly robust estimator suitable for longitudinal estimating equations with dropout. We also propose a related but computationally simpler imputation estimator that uses only baseline data for the imputation model, AIPW-S. Finally, we use the imputation framework to extend the Bang and Robins approach for doubly robust estimates of a mean to more general estimands defined by estimating equations, in the setting of longitudinal data with dropout.

We compare the performance of these three DR imputation estimators through extensive simulations. The AIPW-I estimator performed well, equivalent to correct MLE models when the imputation model is correct, and showed good robustness and efficiency when the dropout model,

but not the imputation model, was correct. The simplified AIPW-S method performed as well as AIPW-I in all but the extreme simulation scenario. At each time point, AIPW-S has the form of a cross-sectional DR estimator, which is easy to interpret and compute. The original B&R estimator is also easy to compute, as is the extension to estimating equations presented here. The B&R estimator has been widely used in applications in its original formulation for estimating a mean. However, its performance in our simulations was the worst. Therefore, it is not recommended here.

These DR methods provide a good opportunity for sensitivity analysis in randomized trials and other settings. In our application to a trial of Donepezil for Alzheimer's disease, it was clear that significant dropout induced differential bias in sample means compared between study arms (Figure 3.1). Because the dropout rates depended on disease severity and were significantly different between the Donepezil group and the placebo group, it is unclear whether a standard regression model would produce consistent estimates. In such a situation, DR methods can be helpful as a sensitivity analysis for the primary analysis. Because of the imputation form of the proposed doubly robust estimators, the AIPW-I and AIPW-S estimators have promise in the clinical trials setting. Future work will compare these doubly robust estimators to the more usual multiple imputation approaches used in this setting.

3.11 Afterthoughts before next Chapter

In this chapter, we reviewed some widely used methods and a novel "Doubly Robust" approach to deal with monotone dropouts in longitudinal data, and compared their performance through simulation studies. We proposed a substitution procedure for a general form of the doubly robust method. We then developed a simplified form for the doubly robust method, which we believe may be easier to interpret and apply in practice under the appropriate circumstance. Since the substitution procedure has been demonstrated, how might investigators apply it in a randomized

trial?

The next chapter will focus more on the regulatory guidance for randomized trials and construct algorithms for doubly robust imputation, which may be useful in sensitivity analysis for such trials, under the missing not at random (MNAR) assumption. Multiple imputation and Paik's sequential mean imputation are reviewed and compared with doubly robust imputation, with algorithms incorporating the pattern mixture model framework.

This chapter, in full, has been submitted for publication as "Qiu, Yuqi; Messer, Karen S. *Doubly robust imputation in longitudinal studies, with an application to an Alzheimer's clinical trial*, submitted to Annals of Applied Statistics". The dissertation author was the primary author on this paper.

Chapter 4

Doubly Robust Imputation for Randomized Trials with Monotone Dropout under Missing not at Random: Applications in Alzheimer's Trials

4.1 Abstract

Sensitivity analysis has been an important area of development in methodology for randomized clinical trials. Regulatory guidelines emphasize the importance of conducting sensitivity analysis when there is loss to follow-up and when intercurrent events occur. Multiple imputation (MI) is one of the predominant missing at random (MAR) based methods used to deal with such missing data. Alternatively, when the assumption of missing not at random (MNAR) has to be incorporated into a sensitivity analysis, several extended MI methods such as δ-based and reference-based MI approaches have been developed. We propose using a substitution-based form of a doubly robust (DR) approach, building on our results from the previous chapter. In

this chapter, we extend DR imputation to δ-based and reference-based approaches to address the MNAR condition. Paik's sequential mean imputation has also been reviewed and extended. Then we compare the performance among these three imputation methods in a MAR scenario and their δ-based and reference-based extensions in an MNAR scenario, with simulation studies and application to two Alzheimer's trials in different stages of the disease. Considering that DR methods can deliver consistent estimation by correcting the bias through propensity scores even if the imputation model is not specified correctly, this study supports DR imputation as a competitive approach to performing sensitivity analysis for randomized trials.

4.2 Introduction

The International Council on Harmonization (ICH) and the Food and Drug Administration (FDA) have recently held several workshops and published guidance elucidating how to describe the *estimand* in clinical trials appropriately [66]. Since then, the word *estimand* in the context of randomized trials has been extensively discussed in the statistical literature [9, 11, 18, 19, 22, 37]. One of the fundamental components of specification of the *estimand* requires the treatment of any intercurrent events to be clarified specifically in the protocol. Discontinuation from the study is a typical class of intercurrent events, and the outcome-related or treatment-related dropout is prevalent in randomized trials for diseases such as Alzheimer's disease (AD) and other dementias. For example, in AD trials, the total attrition rate is often about 25% over two years. The treatment group usually has a higher attrition rate than the placebo group, even if the randomization performed well at baseline. Various methods to address dropout for randomized trials have been developed, depending on assumptions regarding the missing data mechanism. Rubin and Little (1976, 1992, 2002) [54, 29, 31] constructed and summarized the missing data mechanisms, defining three main types, namely, data missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). Little (1995) [30] further described similar

mechanisms in longitudinal data with dropouts. In randomized trials, the dropout mechanism is usually assumed to be MAR in the primary analysis, and the methods of handling it depend on the attrition rate and whether covariates are missing or not.

Recently, regulatory reports have emphasized the importance of conducting sensitivity analysis regarding the assumed dropout mechanism [17, 43], especially when severe dropout may potentially cause significant bias in the study. The motivation for running a sensitivity analysis is to examine the size of the treatment effect under several plausible assumptions regarding the dropout, supposing that dropout is not missing at random. Sensitivity analysis in randomized trials under the assumption of MNAR is a broad statistical topic of current interest. Consideration of various sensitivity analyses is also important, because it can assist investigators in specifying how to deal appropriately with intercurrent events, such as dropout, when defining the *estimand*. The fundamental assumption of MNAR is that the missingness is related to unobserved data. A straightforward idea to account for MNAR is to assume a specific model for the unobserved data, conditional on observed information and dropout status. The pattern mixture model (PMM) framework is a typical method that has been developed and used in this setting [39, 27, 28, 67]. The PMM framework defines several ”patterns” for the unobserved data, such as the time of dropout, and jointly model the distribution of observed data and missingness for each pattern. Δ-based adjustment and reference-based adjustment are two commonly used approaches under the PMM structure. Therefore, an approach that combines PMM with an imputation model becomes an attractive approach to handle MNAR data.

Multiple imputation is the most commonly used method to impute missing values for cross-sectional data or longitudinal data, whether the data is missing response or covariates, or both. Rubin [55] proposed the original idea of multiple imputation, and since then, it has been developed and explored in numerous studies in the context of randomized trials [35, 26, 53, 57, 7]. Multiple imputation is established based on the MAR assumption. By incorporating it with the PMM framework, it can be extended to sensitivity analysis under the MNAR assumption [49, 11].

Currently, most randomized trials use multiple imputation as the primary method for sensitivity analysis in the protocols. Paik (1977) [44] developed a sequential mean imputation for longitudinal data with missing response and proved that under MAR this approach is unbiased.

Doubly robust (DR) methods are an approach to modeling under the MAR assumption, which were initially proposed by Robins [50] as an improvement to inverse probability weighting (IPW) methods. The most common DR estimator, the augmented IPW (AIPW) estimator, has a form that combines IPW and an imputation model. Scharfstein and colleagues [58] pointed the "double robust" property of the AIPW estimator, namely that the AIPW model will be a consistent estimator if either the IPW model or the imputation model is consistently estimated. DR estimation has been explored for cross-sectional data with the missing outcome, and in the causal inference setting [58, 2, 6, 52]. It has also been extended to longitudinal data with dropouts [2, 6, 52, 63]. Tsiatis and colleagues first presented the estimating equation of DR estimator in AIPW form for longitudinal data, and then developed an improved approach to make it more efficient [68, 69].

In this study, we review the properties of the three proposed imputation methods under MAR, and then develop approaches for sensitivity analysis in randomized trials based on Paik's mean imputation and the DR substitution method proposed in Chapter 3. We show that under the condition that the imputation model is correctly specified, Paik's mean imputation has a somewhat heavier workflow than multiple imputation, but is competitive with multiple imputation under either MAR or MNAR scenarios. Doubly robust imputation involves extra steps compared to Paik's mean imputation, to deriving the IPW (also called the "propensity score"); however, it performs as well as multiple imputation and Paik's mean imputation. Considering that the imputation model can never be completely accurate in real-life data, the DR method is proposed as a competitive approach for sensitivity analysis, especially when there is comprehensive information about dropouts to construct precise missingness models.

4.2.1 Organization of this chapter

The chapter is organized as follows: in section (2), we introduce two AD trials in different stages of the disease and elucidate the primary estimand for different scenarios. In section (3), we describe notation and specifically review imputation methods for primary analysis under MAR in randomized trials. The PMM framework with δ-based and reference-based approaches is explored in section (4). Multiple imputation, Paik's mean imputation, and DR imputation are implemented using the PMM framework, and we demonstrate algorithms in detail for their use in sensitivity analyses. In sections (5) and (6), the three imputation methods are evaluated and compared by simulation studies, and then are applied to two AD trials under both the primary and sensitivity analysis scenarios. Section (7) presents discussions and conclusions.

4.3 Alzheimer's Trials

In this study, we acquired data from two randomized, double-blind, placebo-controlled clinical trials conducted by the Alzheimer's Disease Cooperative Study (ADCS) for patients at different stages of Alzheimer's disease.

4.3.1 DHA Trial

A randomized, double-blind, placebo-controlled trial of DHA supplementation in 402 individuals (238 in DHA and 164 in placebo) with mild to moderate Alzheimer's disease (Mini-Mental State Examination scores, 14-26) was conducted between November 2007 and May 2009 at 51 US clinical research sites of the ADCS [48]. The dropout rates by month 18 were 25.6% and 22% for the DHA and placebo groups, respectively. The Alzheimer's disease assessment scale-cognitive subscale (ADAS-Cog) and clinical dementia rating sum of boxes (CDR-SOB), two continuous scores widely used for cognition and function, were the two primary outcomes for this study.

4.3.2 Donepezil Trial

The ADCS conducted this trial between March 1999 and January 2004 ([46]). This study compared the time to progression to possible or probable AD among 769 subjects with mild cognitive impairment (MCI) randomized to treatment for 36 months with either Donepezil (n=253), Vitamin E (n=257), or placebo (n=259). Dropout rates by month 36 were 42.7%, 38.5%, and 33.2% for the Donepezil, Vitamin E, and placebo groups, respectively. This trial also assessed the ADAS-Cog and CDR-SOB as secondary outcomes. We compare the Donepezil and placebo groups in our example.

4.3.3 Primary Estimand

In November 2019, the International Council on Harmonization (ICH) released the final version of an addendum (R1) to ICH E9 guidance [66] called "Estimands and Sensitivity Analysis in Clinical Trials" addressing statistical methods for use in clinical trials. In this context, the estimand is specifically referring to the treatment effect associated with a clinical trial objective, rather than its more general statistical meaning.

Four essential attributes need to be explicitly specified to describe the estimand, namely: (1) defining the targeted study population; (2) defining the endpoint of interest; (3) describing any intercurrent events and how to account for them in detail; (4) summarizing the variable of interest at the population level. The estimand should be clearly specified in the protocol before conducting the study.

Here, we describe the primary estimands for both AD trials as examples. The targeted populations were elucidated in sections 4.3.1 and 4.3.2. For both estimands in this paper, we are most interested in the change from baseline of ADAS-Cog 11 total score at the last time point designed in the protocols (18 months for DHA trial and 36 months for Donepezil trial). We mainly focus on the monotone dropout during the randomized trials, which is a typical intercurrent event.

To summarize the variable of interest at the population level, we plan to report the modeled least squares means of the difference in ADAS-Cog change between the control arm and active arm at the last time point, as our unbiased estimators for our estimands. Under the two assumptions regarding the dropout mechanism, MAR and MNAR, the specifications regarding dropouts and how to handle them are different. For MAR, the dropouts are assumed to continue the effect of treatment in their originally randomized arms. Therefore, a mixed effects model with repeated measures (MMRM) would be appropriate to deal with the dropouts. Under the assumption of MNAR, for example, if the dropouts are assumed to stop the treatment effect after withdrawing from the study, multiple imputation with jump-to-reference adjustment would be appropriate. Thus, the description of the estimand depends on the scientific question of interest, and determines the appropriate statistical approach.

4.4 Summary of Imputation Methods in Primary Analysis

In a longitudinal data structure, we have N iid subjects potentially observed at times $j = 1, \ldots, M$, and for individual i at time j there is data $L_{ij} = (Y_{ij}, X_{ij})$, where Y_{ij} is an univariate outcome and X_{ij} is a vector of covariates; there may also be a vector of always observed time independent baseline covariates X_{i0}. Let $\bar{L}_{ij} = (X_{i0}, L_{i1}^T, ..., L_{ij}^T)$ denote the historical data from time 1 to j. We will often drop the subscript i when the meaning is clear. We assume the distribution of $\bar{L}_M$ has finite second moments.

Dropout can potentially happen on any individuals at time $t > 1$. Let $R_j \in \{0, 1\}$ be a dichotomous variable as missing indicator, so that R_j and $(R_j Y_j, R_j X_j)$ are observed. Under the assumption of monotone dropout, if $R_j = 0$ then for $t > j$, $R_t = 0$. Let C_j be another dichotomous variable as censoring indicator, where $C_{ij} = 1$ indicates j is the last observed time for subject i, otherwise $C_{ij} = 0$, and let J be the index of the last observed time point, so that $C_J = 1$. Under the missing completely at random assumption (MCAR) assumption, R_j and C_j are independent

of $\bar{L}_M$. Under the missing at random assumption (MAR), $P(R_j = 1|\bar{L}_M, R_{j-1} = 1) = P(R_j = 1|\bar{L}_{j-1}, R_{j-1} = 1)$ so that the probability of a missing outcome depends only on previously observed data. We also assume there is probability bounded away from zero of seeing full data over the whole support of L_M: $P(R_{iM} = 1|\bar{L}_M) > \varepsilon > 0$.

In most randomized trials, MAR is assumed to hold for the primary analysis. The analysis population usually is an intent-to-treat (ITT) population, and the covariates usually are baseline characteristics, which are fully observed. When dropouts happened, there are two widely used ways to adjust the analysis for the attrition rate. If the attrition rate is low, a complete case analysis would be used. Complete case analysis deletes subjects who dropped from the study and performs the analysis on completers only. This method could make a consistent mean estimate when the missing is at random and all useful covariates are measured and controlled in the analysis model. However, because less information is used in the model, this method may be inefficient. If the attrition rate is moderate or even worse, imputation methods would be more helpful. Imputation methods file the missing responses in a variety of ways and make use of all observed information. In the past, investigators commonly use last observation carried forward (LOCF). This method replaces the missing responses by the last observed response, which is $(Y_j|R_j = 0) = (Y_t|C_t = 1)$ where $t < j$. Worst case analysis is a more conservative alternative that investigators sometimes use. This method imputes the worst observed response among the active arm for dropouts from the active arm, and imputes the best response among the control arm for dropouts from the control arm. It can be expressed as $(Y_j|R_j = 0, arm) = \text{Worst}(Y|arm)$. These two methods are each problematic in some situations; for example, LOCF would make biased estimates if the dropout is related to the arm, while the worst case analysis is more appropriate for a sensitivity analysis instead of the primary analysis. Regression-based imputation using maximum likelihood estimators and multiple imputation are most commonly used at present. We will introduce these two methods and two additional model-based imputation methods in this section.

4.4.1 Maximum Likelihood Estimator

Under the assumptions of MAR and monotone dropout, a correctly specified regression model such as a mixed effects model would make a consistent mean estimate for the primary estimand. Comparing to the complete case analysis, this approach utilizes all the available information thus obtains more power. A standard mixed effects model $Y_{ij} = Z_{ij}^T B_i + X_{ij}^T \beta + \varepsilon_{ij}$ includes both random effects $Z_{ij}B_i = b_{i1} + b_{i2}t$ and fixed effects $X_{ij}\beta_i$, where we usually assume (b_{i1}, b_{i2}) are random variables distributed from bivariate normal distribution with $\mu = c(0,0)$ and $G = \left(\begin{smallmatrix} g_1 & g_{12} \\ g_{12} & g_2 \end{smallmatrix}\right)$, and random errors ε_{ij} are from normal distribution with $\mu = 0$ and variance σ_0^2. $Z_i = \left(1_M, (t_{i,1}, ..., t_{i,M})^T\right)$ is the design matrix for random effects. This form can be rewritten as $Y_{ij} = X_{ij}^T \beta + e_{ij}$ where the new error term $e_i = Z_i^T B_i + \varepsilon_i$ follows a multivariate normal distribution $N_M\left(\mu = 0_M, \Sigma = \left(\begin{smallmatrix} \sigma_{1,1}^2 & \cdots & \sigma_{1,M}^2 \\ \vdots & \ddots & \vdots \\ \sigma_{M,1}^2 & \cdots & \sigma_{M,M}^2 \end{smallmatrix}\right)\right)$. Specifically, $\sigma_{j,j}^2 = Z_j G Z_j^T + \sigma_0^2$ and $\sigma_{j,k}^2 = Z_j G Z_k^T$. Transfer the model to matrix form $Y = X\beta + error$, here $error$ is a vector of $n \times m$ elements and its distribution is $N_{n\times m}(0, D)$, thus the maximum likelihood estimator is $\hat{\beta}_{MLE} = (X^T D^{-1} X)^{-1} X^T V^{-1} Y$. This model can be generalized to similar form when each individual has different times of repeats. Also, the generalized least squares (GLS) model, which only includes fixed effects but correlated ε_{ij} within one subject i, can also be expressed in similar form as the maximum likelihood estimator.

Many statistical packages in different platforms can estimate the mixed effect model. It is a commonly used analysis model for randomized trials. This method will consistently estimate the primary estimand under the assumption that dropouts are still obtaining the treatment effect. However, note that with MAR dropouts, both the mean structure and the covariance structure must be correctly specified to get consistent estimates.

4.4.2 Multiple Imputation

When the attrition rate is high, the mixed effects model may have convergence issues, especially when assuming a fully parameterized covariance structure. Multiple imputation was

developed under MAR and there is a large literature using this method to fill in the missing outcomes first, then performing the analysis model with complete imputed data. Rubin [55] first introduced the multiple imputation idea by imputing a missing item with multiple values; since then, many studies have explored the multiple imputation under different situations. Van Buuren and colleagues [70] proposed a specific implementation of multiple imputation called "multiple imputation with chained equations" (MICE), which is frequently used in applications. MICE has an iterative process to construct each imputed data set. A proper multiple imputation with chained equation process is as below:

- Start from a variable X_p with missing items, regress X_p on other variables using all observed values with linear regression or logistic regression depends on X_p to derive $\hat{\beta}$ and residual standard error $\hat{\sigma}$.
- Draw $\hat{\beta}_b$ and $\hat{\sigma}_b$ from the Bayesian posterior of $(\hat{\beta}, \hat{\sigma})$ and impute missing values with $\hat{\beta_b}$ plus a random error with variance of $\hat{\sigma}_b^2$.
- Apply the above steps for every variable with missing values, and treat the already imputed values as observed.
- Repeat the above steps Q times, which each Q is an iteration of the chain so that the imputed values can be updated. At the end of Q times, one complete imputed data set is created.
- Repeat the above steps B times, and then we finally obtain B fully imputed data sets.

Many other imputation approaches have been developed in the multiple imputation literature to replace the Bayesian regression. However, most of them were proved to be "improper" multiple imputation. Rubin [55] demonstrated that multiple imputation with Bayesian regression is proper; that is, it obtains with valid variance estimates by Rubin's rule. Given the context of the data, these approaches can be customized by setting the order of imputing X_p, choosing which variables should be included as predictors for X_p, and so on.

In the context of a standard randomized trial, we note that when the dropout is monotone and covariates are fully observed, MICE seems unnecessary. Under the monotone dropout setting, Y_j should be predicted by previous $Y_1, ..., Y_{j-1}$ and X but not any Y_k with $k > j$. Thus, this sequential regression process would start from regressing the first Y_j with missing items, usually Y_2, on Y_1 and X. Since Y_1 and X are fully observed, Y_2 should be fully imputed after this step. Along with the sequential regression process, each following Y_j should be fully imputed, and iteration makes no changes to the imputed values. This allows for a more efficient implementation of multiple imputation.

4.4.3 Paik's Imputation

Since MICE is not necessary for monotone dropout with covariates fully observed, other imputation methods may perform as well as the sequential imputation described above. Here we would like to briefly introduce an alternative sequential regression process called Paik's mean imputation [44] for handling longitudinal data with monotone dropouts. Paik's mean imputation forms the foundation for the doubly robust estimators for longitudinal data, which appear in the literature [2, 63].

Paik defined a sequential regression approach for imputing the missing outcome values under MAR. Instead of imputing Y_2 at first, Paik's mean imputation starts with imputing any Y_j regarding the dropout patterns. In other words, defining $J_i = j|C_{ij} = 1$ is the last observed time point for subject i, we can categorize subjects into M dropout patterns from $J_i = 1$ represents Y_1 observed only to $J_i = M$ refers to completers. Then for a given time point $k, 1 < k \leq M$, we impute the missing Y_k by:

- First regress Y_k on $\bar{L}_{k-1}$ for subjects with dropout pattern $J_i \geq k$ to obtain a consistently imputation model $\hat{m}_k^{k-1}(\bar{L}_{k-1})$.
- Then impute Y_k using model $\hat{m}_k^{k-1}(\bar{L}_{k-1})$ for subjects with dropout pattern $J_i = k-1$, and

treat the imputed Y_k as observed values.

- Sequentially from $J = k-1$ to 2, regress Y_k on $\bar{L}_{J-1}$ for subjects with dropout pattern $J_i \geq J$ to obtain consistently imputation model $\hat{m}_k^{J-1}(\bar{L}_{J-1})$. Impute Y_k using the model for subjects with dropout pattern $J_i = J-1$, and treat the imputed values as observed for next sequence.

Both multiple imputation and Paik's imputation use sequential regressions to obtain the same number of imputation models. The main difference between multiple imputation process and Paik's mean imputation process is that multiple imputation starts imputation from the first Y_j with missing items and uses imputed Y_j as a covariate in the following sequences to impute Y_k where $k > j$. In contrast, Paik's imputation starts imputation simultaneously with Y_2 to Y_M and uses imputed Y_j as outcomes in the subsequent regression models. Under the monotone dropout assumption, the sample size of the sequential regression models decreases for multiple imputation but increases for Paik's mean imputation, while the number of covariates increases for multiple imputation but decreases for Paik's mean imputation. The total computational load remains the same between the two methods.

4.4.4 Doubly Robust Imputation

A novel method called "doubly robust" (DR) estimator based on semi-parametric theory has been developed for missing data issue or causal inference problems. The DR estimator combines both an imputation model and estimated inverse probability weights (IPW) into an augmented inverse probability weighted (AIPW) estimator, in such a way that the estimator is consistent if either the imputation model or the IPW model is consistent under MAR. For cross sectional data (i.e. M=1), the AIPW form ([51]) of a DR estimating equation U is

$$\sum_{i=1}^{N}\left(\frac{R_i}{\hat{\pi}_i}U_i + (1-\frac{R_i}{\hat{\pi}_i})\hat{H}(X_{i0})\right) = 0 \tag{4.1}$$

where $\hat{H}(X_0)$ (the imputation model) is an estimate of $E[U|X_0]$.

Many studies have explored DR estimators in the cross-sectional data setting, while for longitudinal data, there are fewer methodological studies, and it is hard to find examples of use in applications, perhaps due to the complexity in calculation and interpretation. We extended the DR method as an imputation method and proposed a specific simplified form for longitudinal data in Chapter 3; here, we would like only to introduce the general form and idea of the DR method in AIPW form.

Tsiatis [68] proved that the estimating equation of a DR in longitudinal form is:

$$\sum_{i=1}^{N}\left(\frac{C_{i,M}}{\pi_{i,M}}U(\bar{L}_{iM})+\sum_{j=1}^{M-1}\left(\frac{C_{i,j}-\lambda_{i,j+1}R_{i,j}}{\pi_{i,j+1}}\right)H^{j}(\bar{L}_{ij})\right)=0 \tag{4.2}$$

where $\lambda_j = Pr(C_j = 1|\bar{L}_{j-1})$ is the hazard function and $\pi_{i,j}=\prod_{t=1}^{j}(1-\lambda_t) = Pr(R_j = 1)$ is the probability of observed or not.

In practice, the imputation model for $H^j(\bar{L}_{ij})$ can be any consistently estimated model. Seaman [63] came up with using Paik's mean imputation model for $H^j(\bar{L}_{ij})$. We showed that if we set the target estimating equation $U_j = Y_j - E(Y_j)$ and assume $\pi_{M+1} = \pi_M$ and $\lambda_1 = \lambda_{M+1} = 0$, then the DR process can be used as an imputation method with

$$\hat{Y}^{DR}=\left\{\Big(\sum_{j=1}^{M-1}\big(\frac{C_j-\hat{\lambda}_{j+1}R_j}{\hat{\pi}_{j+1}}\big)\hat{E}[Y|\bar{L}_j]\Big)\right\}. \tag{4.3}$$

Like multiple imputation and Paik's mean imputation, after filling in the missing items, any standard statistical analysis can be performed on the fully imputed data set. When applying the DR imputation, there are two parts of models to be estimated, and they can be calculated separately. The first part is the IPW or propensity scores at each time point, these models are usually fitted by logistic regression. The second part is the imputation model for $E(Y_j)$, and the process is same as section 4.4.3 if we choose to use Paik's mean imputation to estimate $H^j(\bar{L}_{ij})$.

As we introduced at the beginning of section 4.4.4, DR estimators double the chance of

making a consistent estimate by combining both the missing model and imputation model. DR imputation inherits this property, thus the fully imputed data set supports consistent estimation when either of the two models is correctly specified. In section 4.6 we showed that when the imputation model is correct, DR imputation performs as well as multiple imputation and Paik's imputation.

4.5 Algorithms for Imputation Methods in Sensitivity Analysis

Regulatory reports [17, 43] emphasize the importance of presenting a sensitivity analysis, because of the concerns about the assumption of MAR in the primary analysis. MAR requires the imputation model to be correctly specified, which not only asks for necessary covariates to be controlled, but also requires a correct covariance structure to be specified for repeated measures data. However, there is no test to demonstrate whether a model is accurately specified or a data set is MAR in practice. Sensitivity analysis assists in constructing a more conservative framework when estimating the primary estimand, which tests the reliance of the study's conclusions on the MAR assumption. More specifically, sensitivity analysis considers the condition that the missing mechanism is MNAR, in which the unobserved data after discontinuation from the study has a different distribution than the observed data, particularly in the treatment group.

One of the most favored and most straightforward ways to model the different distributions of observed data and unobserved data is the pattern mixture model (PMM). Formula 4.4 is the expression of PMM framework, where Y_{obs} and Y_{mis} represent observed and missing Y, respectively. This model jointly models the distributions between the outcome Y and the missingness R, given observed covariates X. In our context, pattern refers to the dropout time J_i as elucidated in section 4.4.3. The PMM can be further elaborated according to multiple meanings of the word "pattern". For example, subjects can also be categorized by dropout reasons if the clinical data collects this

information.

$$\begin{aligned} f(Y_{obs}, Y_{mis}, R|X) &= f(R|X) f(Y_{obs}, Y_{mis}|R, X) \\ &= f(R|X) f(Y_{obs}|R, X) f(Y_{mis}|Y_{obs}, R, X) \end{aligned} \tag{4.4}$$

Since Y_{mis} is unobserved, we cannot determine the function of $f(Y_{mis}|R,X)$ from the data, which is denoted as the "identifying restrictions" in the literature [67]. Although we cannot identify the "true" distribution of Y_{mis}, an assumed or modeled distribution for $f(Y_{mis}|Y_{obs},R,X)$ can be constructed, incorporating the distribution of Y_{obs}, by means of a series of link functions which connect Y_{mis} and Y_{obs}.

In this section, we focus on the algorithms of two widely used methods under the PMM approach, namely δ-based adjustment and reference-based adjustment, incorporating the three imputation methods of multiple imputation, Paik's mean imputation, and DR imputation.

4.5.1 Δ-based Adjustment

Δ-based adjustment builds a link function $f(Y_{mis}|X) = g\big(f(Y_{obs}|X)\big) = f(Y_{obs}|X) + \delta$ for active arm, where δ is a reasonable number selected under the context of the clinical outcome Y. This link function assumes the distribution of $Y|X, arm = Active$ shifts by an amount δ from the original distribution, after a subject withdraws from the study. This adjustment lessens the treatment effect consistently; for example, if the continuous outcome Y becomes worse with a greater value, δ would be a positive number to make the adjusted Y worse than before; on the contrary, if a smaller value of Y infers more severe of the disease, then δ should be a negative number. Δ-based adjustment can be used in three main approaches:

- The first approach is applying δ just once. For active arm's subjects with dropout pattern J_i, only shift Y_{J_i+1} by δ as the imputed $\hat{Y}_{J_i+1}$, and the subsequent missing outcomes $Y_{J_i+2}, ..., Y_M$

follow the original distribution of $Y|arm = Active$ given X and $Y_1, ..., Y_{J_i}, \hat{Y}_{J_i+1}$.

- The second approach, which we call it δ-post-adjustment, is shifting every Y_{mis} by δ after all $Y_{mis}|arm = Active$ have been imputed under MAR.

- The last approach is a sequential process that imputes Y_{J_i+1} under MAR and adjusts it by δ, then imputes Y_{J_i+2} under MAR given observed X, Y and the δ-adjusted Y_{J_i+1}, and so on till the last missing outcome Y_M is imputed and adjusted.

In the following sections, we mainly focus on the third approach. Δ-based adjustment helps investigators test the critical value of the treatment effect, which is called the tipping point of the δ, which precisely brings the estimate of our primary estimand to non-significant.

Δ-based Multiple Imputation

As the principal method of handling missing data, multiple imputation in sensitivity analysis has been developed in many studies [11, 49]. Here we write down the algorithms for δ-based multiple imputation algorithm, and in section 4.5.2 the reference-based multiple imputation algorithm, along with some special issues that need to be addressed when practically applying these methods to data.

Algorithm 3 describes the δ-based multiple imputation process. As we mentioned in section 4.4.2, under the assumption of monotone dropout, the chained equation does not make additional contribution. The algorithm combines with δ adjustment after each unobserved Y_j has been imputed sequentially. The key of this approach is that each $\hat{Y}_{j+1}|arm = Active$ should inherit the distribution of observed $\bar{L}_{j-1}|arm = Active$ and then be δ-adjusted $\hat{Y}_j$ with an additional δ shift. In some statistical software like SAS and STATA, there are built-in packages to perform this process. However, in R, it is not straightforward to obtain this estimate. We drafted some code in R based on the "MICE" package [72] for this process that can be shared by request.

Algorithm 3 describes the procedure for one of the multiple data sets created by multiple

imputation. To obtain B imputed data sets, the algorithm is repeated B times. Furthermore, in order to obtain the estimate of our primary estimand, a mixed effects model is fitted to every fully imputed data set with the least squares mean estimator and its standard error of the estimand being derived, followed by applying Rubin's rule to attain the final estimate of the primary estimand and its variance.

Algorithm 3: Sequential regression based MI with δ-based adjustment

Result: A fully imputed response data set $\hat{Y}^{MI}$, in which all missing outcomes have been imputed, using a δ-based sequential regression MI.

1 **for** j = *2,...,M sequentially* **do**

2 **MI step 1:** Regress the values of Y_{ij} on $\bar{L}_{j-1}$ using complete data to obtain estimated model $m_j(\bar{L}_{j-1})$ and then for imputation b using Bayesian paradigm to draw model $m_j^b(\bar{L}_{i,k-1})$.

3 **δ-adjustment step:** Add a δ to $Y_{i,j}^{Miss}$ in active arm. Usually the add-on δ indicates active arm at time point j worse by a value of δ. Numerically add or minus the δ should be decided by the context.

$$\hat{Y}_{i,j}^{Miss} = \begin{cases} \hat{m}_j^b(\bar{L}_{i,k-1}) & \text{if } arm = Control \\ \hat{m}_j^b(\bar{L}_{i,k-1}) + \delta & \text{if } arm = Active \end{cases}$$

4 **MI step 2:** Merge $\hat{Y}_{i,j}^{Miss}$ into $Y_{i,j}^{obs}$ as observed data, then go to the next sequence $j = j+1$.

5 **Note:** In MI step 1, later $\bar{L}_j$ would include the imputed $\hat{Y}_{i,j}^{Miss}$ from previous round.

Δ-based Paik's Imputation

Paik's mean imputation is widely used in the construction of longitudinal doubly robust estimators.[2, 63] However, few studies discuss Paik's mean imputation in the sensitivity analysis unlike multiple imputation. We extend the procedure of Paik's mean imputation combining with δ-based approach in algorithm 4. As discussed in section 4.4.3, under the monotone dropout assumption, we do not see multiple imputation outperform Paik's mean imputation. Since Paik's imputation could start imputing Y_j at any time point j, algorithm 4 presents the process for a given time k. Consistent with the procedure in primary analysis 4.4.3, Paik's mean imputation imputes

Y_k by the sequence of subjects' dropout patterns, and uses the imputed values as outcomes in the sequential regression models. From $J_i = k-1$ to $J_i = 1$, δ is added to the model predicted $\hat{m}_k^{J_i}(\bar{L}_{i,J_i})$ before fitting next model $\hat{m}_k^{J_i-1}(\bar{L}_{i,J_i-1})$ so that the subsequent models could account for the previous δ-adjusted $\hat{Y}_k$. We programmed Paik's mean imputation in both primary analysis and sensitivity analysis in R, and the codes can be shared as well.

By assigning $k = 2, ..., M$ and applying algorithm 4, a fully imputed data set is produced by Paik's mean imputation with δ adjustment. Analogous to section 4.5.1, least squares mean estimator of the primary estimand would be obtained by fitting a mixed effects model on the fully imputed data set. The variance could be estimated by the Bootstrap resampling procedure. The analytical variance could also be attained with a complicated form, which is not shown here.

Algorithm 4: Paik's imputation with δ-based adjustment

Result: For a given time k, the fully imputed response data $\hat{Y}_k^{Paik}$, in which all missing outcomes at time k have been imputed, using a δ-based Paik's mean imputation.

1 **Paik step:** Identify all subjects i with $J_i \geq k$. Use these data to regress the observed values of Y_k on $\bar{L}_{k-1}$ to obtain a consistently estimated model $\hat{m}_k^{k-1}(\bar{L}_{k-1})$.

2 **δ-adjustment step:** Add a δ to $Y_{i,j}^{Miss}$ in active arm. Usually the add-on δ indicates active arm at time point j worse by a value of δ. Numerically add or minus the δ should be decided by the context.

$$\hat{Y}_{ik}^{(k-1)} = \begin{cases} \hat{m}_k^{k-1}(\bar{L}_{i,k-1}) & \text{if } J_i = k-1 \text{ and } arm = Control \\ \hat{m}_k^{k-1}(\bar{L}_{i,k-1}) + \delta & \text{if } J_i = k-1 \text{ and } arm = Active \\ Y_{ik} & \text{if } J_i > k-1 \end{cases}$$

3 **For $\mathbf{s = k-2, ..., 1}$ sequentially:** Regress the values of $\hat{Y}_{ik}^{s+1}$ on $\bar{L}_{is}$ to obtain a consistently estimated model $\hat{m}_k^s(\bar{L}_s)$. For all subjects let

$$\hat{Y}_{ik}^{Paik} = \hat{Y}_{ik}^{(s)} = \begin{cases} \hat{m}_k^s(\bar{L}_s) & \text{if } J_i = s \text{ and } arm = Control \\ \hat{m}_k^s(\bar{L}_s) + \delta & \text{if } J_i = s \text{ and } arm = Active \\ \hat{Y}_{ik}^{s+1} & \text{if } J_i > s \end{cases}$$

Δ-based Doubly Robust Imputation

DR imputation is a novel method for dealing with the missing data, and its properties have been introduced in section 4.4.4. No literature that we are aware of has explored the DR imputation in sensitivity analysis for randomized trials so far.

To estimate the estimand under MAR, DR imputation combines propensity scores and an imputation model such as Paik's mean imputation, thus if Paik's mean imputation model is correctly specified, the estimate should be consistent for Paik's imputation as well as DR imputation. In contrast, if Paik's model is not correctly specified, Paik's imputation would fail to make a consistent estimate, but DR imputation still has the chance to obtain a consistent estimate by using propensity scores to fix the bias. In the context of the sensitivity analysis, DR imputation should also inherit this robustness property.

Algorithm 5 demonstrates the procedure of DR imputation with δ adjustment, with Paik's model as the imputation model. For a given time k, the algorithm is similar to Algorithm 4, with additional steps to derive propensity scores by MLE and to obtain the final DR imputation by incorporating propensity scores and Paik's imputed values. Also, by conducting the Algorithm 5 $M-1$ times, we can obtain a fully imputed data set by DR imputation with δ adjustment. After that, the procedure is identical to section 5, where we calculated the least squares mean estimator and its variance for the primary estimand.

4.5.2 Reference-based Adjustment

Reference-based adjustment has been an active area of exploration recently. The original idea was by Little and Yau [26], and extended to several similar approaches. Reference-based adjustment creates a link function $f(Y_{mis}|X, arm = Active) = g\big(f(Y_{obs}|X, arm = Control)\big) = f(Y_{obs}|X, arm = Control)$, indicating that the essential idea assumes those subjects in the active arm who dropped out from the study would follow the same distribution as the subjects in control

Algorithm 5: DR imputation with δ-based adjustment

Result: For a given time $k > 1$, the fully imputed data $\hat{Y}_{ik}^{AIPW}$, in which all missing values at time k have been imputed, using a δ-based AIPW imputation.

1 **Preliminary step:** Estimate $\hat{\lambda}_2, ..., \hat{\lambda}_k$ and their corresponding $\hat{\pi}_j$ by maximum likelihood.

2 **Paik step:** Identify all subjects i with $J_i \geq k$. Use these data to regress the observed values of Y_k on $\bar{L}_{k-1}$, to obtain a consistently estimated model $\hat{m}_k^{k-1}(\bar{L}_{k-1})$. For subjects with $J_i > k-1$, let $\hat{Y}_{ik}^{(k-1)} = Y_{ik}$.

3 **δ-adjustment step:** For subjects with $J_i = k-1$ let

$$\hat{Y}_{ik}^{(k-1)} = \begin{cases} \hat{m}_k^{k-1}(\bar{L}_{i,k-1}) & \text{if } arm = Control \\ \hat{m}_k^{k-1}(\bar{L}_{i,k-1}) + \delta & \text{if } arm = Active \end{cases}$$

4 **For s = k − 2, ..., 1 sequentially:** Regress the values of $\hat{Y}_{ik}^{s+1}$ on $\bar{L}_{is}$ to obtain a consistently estimated model $\hat{m}_k^s(\bar{L}_s)$. For all subjects with $J_i > s$ let $\hat{Y}_{ik}^s = \hat{Y}_{ik}^{s+1}$; for subjects with $J_i = s$ let

$$\hat{Y}_{ik}^{(s)} = \begin{cases} \hat{m}_k^s(\bar{L}_s) & \text{if } arm = Control \\ \hat{m}_k^s(\bar{L}_s) + \delta & \text{if } arm = Active \end{cases}$$

5 **Calculating AIPW imputed values by:**

$$\hat{Y}_{ik}^{AIPW} = \frac{C_{ik} Y_{ik}}{\hat{\pi}_{ik}} + \sum_{j=1}^{k-1} \left(\frac{C_{ij} - \hat{\lambda}_{ij+1} R_j}{\hat{\pi}_{ij+1}}\right) \hat{Y}_{ik}^{(j)} \tag{4.5}$$

arm, because they stop taking treatment and return to standard of care.

Currently, many investigators use jump-to-reference (J2R) adjustment, one of the most straightforward and most frequently used reference-based approaches, for designing their sensitivity analysis in clinical trials. Jump-to-reference adjustment reflects the scheme of the reference-based framework, which models the distribution of $Y_{mis}|arm = Active$ following the distribution of $Y|arm = Control$, given observed X and Y. Carpenter [8] defined and placed three other options into the reference-based family, which are "copy increments in reference (CIR)" approach, "copy reference (CR)" approach, and "last mean carried forward (LMCF)" approach. CIR means the link function $f(Y_{i,k}|X, arm = Active) = g\big(f(Y_{i,J_i}|X, arm = Active)\big) = f(Y_{i,J_i}|X, arm = Active) + w_j$, where $k > J_i$ indicating Y_k is unobserved, and w_j is the mean increment observed from the control (reference) arm. CR refers to an approach that identifying all subjects i who are dropout and in the active arm, then allow their distribution of $Y|X$ follow the distribution of $Y|X$ from the control arm, for all Y regardless of observed or not. LMCF may be more familiar, which imputes the missing outcomes by $E(Y_{J_i}|arm)$. LMCF is a transformation of LOCF described in section 4.4 and not related to the reference-based framework as other approaches. We focus on the jump-to-reference approach in later sections when describing the reference-based adjustment.

Comparing δ-based and reference-based adjustments, we found that they are equivalent after transformation under some conditions. For example, if $Y_{ij}|arm = Active$ and $Y_{ij}|arm = Control$ are independent random variables from two multivariate normal distributions with the same variance but different means, and if their mean difference is a linear function of time, the two models could be specified as a linear mixed effects model presented in section 4.4.1:

$$y_{ij} \quad = \quad X\beta + \alpha_1 \times j + \alpha_2 \times j \times I(arm = Active) + e_{ij} \tag{4.6}$$

where X are fully observed baseline covariates. In addition, if the randomization is well done at baseline between arms (which is a common assumption in randomized trials), CIR approach

would be identical to the jump-to-reference approach, and δ-based adjustment would be equivalent to these two reference-based approaches as well, with $\delta = -\alpha_2$. However, we also found that if missingness is related to historical Y, and if there are random slopes (even if they have a mean of 0), this equivalence would be broken between reference-based approach and δ-based approach. We discuss more about the issue with comparing these methods in the simulation study section 4.6.

Reference-based Multiple Imputation

Multiple imputation with reference-based method also has been extensively studied [11, 49]. The whole procedure is similar to that in the δ-based adjustment section 4.5.1. Algorithm 6 presents this procedure for a single imputed data set. Sequentially from $j = 2$ to M, missing outcome Y_j^{Miss} is imputed by distribution of control arm's Y_j^{obs}, given covariates X except *arm*, and previous observed or imputed $Y_1, ..., Y_{j-1}$.

By repeating Algorithm 6 B times, we can obtain B fully imputed data sets and then perform statistical analysis such as a mixed effects model on each data set to calculate the least squares mean estimator and its standard error for our primary estimand. Finally, we use Rubin's rule to merge B estimates into one ultimate reference-based multiple imputation estimate.

Algorithm 6: Sequential regression based MI with J2R adjustment

Result: A fully imputed response data set $\hat{Y}^{MI}$, in which all missing outcomes have been imputed, using a reference-based (J2R) sequential regression MI.

for *j = 2,...,M sequentially* **do**

- **MI step 1:** Regress the values of Y_{ij} on $\bar{L}_{j-1}$ using $Y, X | arm = Control, R_j = 1$, where $\bar{L}_{j-1}$ doesn't include *arm*, to obtain estimated model $m_j(\bar{L}_{j-1})$ and then for imputation b using Bayesian paradigm to draw model $m_j^b(\bar{L}_{j-1})$.
- **J2R-adjustment step:** Let $\hat{Y}_{i,j}^{Miss} = m_j^b(\bar{L}_{j-1})$ regardless of arms.
- **MI step 2:** Combine $\hat{Y}_{i,j}^{Miss}$ and $Y_{i,j}^{obs}$ then go to the next sequence $j = j + 1$.

Note: In MI step 1, later $\bar{L}_j$ would include the imputed $\hat{Y}_{i,j}^{Miss}$ from previous round.

Reference-based Paik's Imputation

We describe the procedure of jump-to-reference with Paik's mean imputation in Algorithm 7. Similar to the primary analysis and δ-based approaches, Paik's mean imputation with reference-based adjustment is a comparable method to the popular multiple imputation with a similar workflow.

Algorithm 7 shows how to impute Y_j at a given time point through Paik's mean imputation. To attain a fully imputed data set, one should repeat this procedure $M-1$ times till all time points have been imputed. As with the δ-based adjustment, the reference-based adjustment happens within each sequence in Paik's mean imputation. The adjusted values would be merged with observed values and used as the outcome for the subsequent imputation model. After obtaining a fully imputed data set, we collect the least squares mean estimator and its standard error of our primary estimand through the same procedure described before.

Unlike multiple imputation, which starts from $j = 2$ to M sequentially and all Y_j can be imputed in one cycle, Paik's imputation can only impute Y_j at a given time k in a cycle, and starts with dropout pattern $J_i \geq k$ to $J_i = 1$ sequentially, that is from subjects with longer duration to minimum duration. Meanwhile, multiple imputation has to create B data sets to account for the imputation variability, but Paik's mean imputation only needs a single imputed data set. However, the total workload is the same between the two methods.

Reference-based Doubly Robust Imputation

Last but not least, DR imputation with reference-based adjustment is summarized in this section with the algorithm in Algorithm 8. Again we use Paik's mean imputation as the imputation model within the DR process. The procedure of Algorithm 7 is nested in Algorithm 8, while DR imputation additionally derives propensity scores at each time point and combines them with Paik's imputed values in the end.

Comparing to Paik's mean imputation, DR imputation adds a missingness model to doubly

Algorithm 7: Paik's imputation with J2R adjustment

Result: For a given time k, the fully imputed response data $\hat{Y}_k^{Paik}$, in which all missing outcomes at time k have been imputed, using a reference-based (J2R) Paik's mean imputation.

1 **Paik step:** Identify all subjects i in control arm with $J_i \geq k$. Use these data to regress the observed values of Y_k on $\bar{L}_{k-1}$, where $\bar{L}$ doesn't include *arm*, to obtain a consistently estimated model $\hat{m}_k^{k-1}(\bar{L}_{k-1})$.

2 **J2R-adjustment step:** For all subjects let

$$\hat{Y}_{ik}^{k-1} = \begin{cases} \hat{m}_k^{k-1}(\bar{L}_{i,k-1}) & \text{if } J_i = k-1 \\ Y_{ik} & \text{if } J_i > k-1 \end{cases}$$

3 **For $\mathbf{s = k-2, ..., 1}$ sequentially:** Regress the values of $\hat{Y}_{ik}^{s+1}$ from subjects in control arm on $\bar{L}_{is}$ to obtain a consistently estimated model $\hat{m}_k^s(\bar{L}_s)$. For all subjects let

$$\hat{Y}_{ik}^{Paik} = \hat{Y}_{ik}^{s} = \begin{cases} \hat{m}_k^s(\bar{L}_s) & \text{if } J_i = s \\ \hat{Y}_{ik}^{s+1} & \text{if } J_i > s \end{cases}$$

ensure that the estimation is consistent, which is an improvement over Paik. In contrast, the penalty is that DR imputation has the extra steps of obtaining the propensity scores, as well as potentially reduces efficiency. Like Algorithm 7, Algorithm 8 also derives the imputed value of Y_j at a given time k. Thus the complete Y_j should be imputed by reproducing this procedure $M-1$ times. After attaining the fully imputed data set, one can do the same before estimating the primary estimand.

4.6 Simulation Study

Simulation studies are performed to evaluate multiple imputation (MI), Paik's mean imputation, and AIPW-form DR imputation in both primary analysis assuming dropout at random and sensitivity analysis assuming dropout not at random with 500 Monte Carlo repeats. Monotone dropout is presumed, and baseline covariates are assumed to be completely observed.

The complete data is generated from a linear mixed effects model with random intercepts and random slopes. Specifically, longitudinal outcome Y_{ij} is generated from model (4.8) with

Algorithm 8: DR imputation with J2R adjustment

Result: For a given time k, the fully imputed response data $\hat{Y}_{ik}^{AIPW}$, in which all missing outcomes at time k have been imputed, using a reference-based (J2R) DR imputation.

1 **Preliminary step:** Estimate $\hat{\lambda}_2, ..., \hat{\lambda}_k$ and their corresponding $\hat{\pi}_j$ by maximum likelihood.

2 **Paik step:** Identify all subjects i in control arm with $J_i \geq k$. Use these data to regress the observed values of Y_k on $\bar{L}_{k-1}$, where $\bar{L}$ doesn't include *arm*, to obtain a consistently estimated model $\hat{m}_k^{k-1}(\bar{L}_{k-1})$.

3 **J2R-adjustment step:** For all subjects let

$$\hat{Y}_{ik}^{k-1} = \begin{cases} \hat{m}_k^{k-1}(\bar{L}_{i,k-1}) & \text{if } J_i = k-1 \\ Y_{ik} & \text{if } J_i > k-1 \end{cases}$$

4 **For $\mathbf{s = k-2, ..., 1}$ sequentially:** Regress the values of $\hat{Y}_{ik}^{s+1}$ from subjects in control arm on $\bar{L}_{is}$ to obtain a consistently estimated model $\hat{m}_k^s(\bar{L}_s)$. For all subjects let

$$\hat{Y}_{ik}^{s} = \begin{cases} \hat{m}_k^{s}(\bar{L}_s) & \text{if } J_i = s \\ \hat{Y}_{ik}^{s+1} & \text{if } J_i > s \end{cases}$$

5 **Calculating AIPW imputed values by:**

$$\hat{Y}_{ik}^{AIPW} = \frac{C_{ik} Y_{ik}}{\hat{\pi}_{ik}} + \sum_{j=1}^{k-1} \left(\frac{C_{ij} - \hat{\lambda}_{ij+1} R_j}{\hat{\pi}_{ij+1}}\right) \hat{Y}_{ik}^{j} \tag{4.7}$$

$i = 1, ..., 500$ and $j = 1, 2, 3$. Covariates are simulated with mean and standard deviation derived from DHA trial, where continuous covariates are random variables from normal distributions and dichotomous covariates are distributed from Bernoulli distributions. Covariance of random effects is taken from the trial as well, where b_{i1} and b_{i2} came from a bivariate normal distribution with mean $\mu = \binom{0}{0}$ and covariance $G = \begin{pmatrix} 9.60 & -1.08 \\ -1.08 & 5.90 \end{pmatrix}$. Two arms are randomly assigned to the subjects in each repeat with same likelihood.

$$\begin{aligned} Y_{ij} &= b_{i1} + b_{i2}\text{Visit} + \beta_0 + \beta_1\text{Baseline Value} + \beta_2\text{MMSE} + \beta_3\text{Gender} \\ &+ \beta_4\text{Arm} + \beta_5\text{Visit} + \beta_6\text{Arm*Visit} + \varepsilon_{ij} \end{aligned} \tag{4.8}$$

Dropouts are simulated from two logistic regression models (4.9 and 4.10) for $Pr(\lambda_1 = 1)$ and $Pr(\lambda_2 = 1)$. In these two models we assumed that whether a subject withdraws from the study or not depends on its historical Y and *arm*. The attrition rates at $j = 2$ is around 18% and at $j = 3$ is around 35%.

$$logit(Pr(\lambda_1 = 1)) = -27 + 0.5Y_1 + 0.75\text{Baseline Value} + \text{Arm} \tag{4.9}$$

$$logit(Pr(\lambda_2 = 1)) = -31 + 0.3Y_1 + 0.5Y_2 + 0.9\text{Baseline Value} + \text{Arm} \tag{4.10}$$

We focus on estimating the *primary estimand*, which is the difference based on the least squares mean estimator in $E(Y_M|X)$ between arms, as described in section 4.3.3. In addition, based on the least squares mean estimators of $E(Y_M|X)$ for each arm, we define an analogous secondary estimand to be estimated. Bias, average standard errors, root mean square errors, Monte Carlo standard deviation, and coverage probability are reported. Multiple imputation makes use of Rubin's rule to calculate the standard error, while Paik's mean imputation and DR imputation use the bootstrap to derive the standard errors. Sizes of both multiple imputation and the bootstrap

replicates are set to $B = 200$.

4.6.1 Primary Analysis

In the primary analysis, we assume missing at random for monotone dropouts. Under this assumption, the imputation model can be correctly specified when all useful covariates are measured and controlled in the model. The truth of the secondary estimand is 12.06 and 8.46 for control arm and active arm respectively, and the truth of the *primary estimand* is -3.6.

As reported in Table 4.1, all three imputation methods show biases smaller than 0.03 for every estimation. They also had consistent results of the averaging standard errors, where AIPW was always approximately 0.01 higher than the other two methods. The three methods indicate similar root mean square errors and coverage probabilities as well, demonstrating the three methods are consistent with each other in both bias and efficiency. AIPW imputation performs as well as multiple imputation in our simulation. Considering it is a robust method with the ability to fix the bias even if the imputation model is not specified correctly, AIPW imputation seems to be a competitive approach for handling longitudinal data with monotone dropouts.

Table 4.1: Simulation Results for the *primary estimand* and the secondary estimand

least squares Means		Primary Analysis					Sensitivity Analysis				
	Method	Bias	SE	RMSE	MCSD	Covp	Bias	SE	RMSE	MCSD	Covp
Control	MI	-0.023	0.512	0.545	0.544	0.938	-0.023	0.528	0.555	0.555	0.948
	Paik	-0.011	0.513	0.545	0.545	0.940	-0.012	0.524	0.557	0.557	0.938
	AIPW	-0.016	0.518	0.554	0.553	0.936	-0.005	0.528	0.562	0.562	0.934
Active	MI	-0.019	0.512	0.553	0.552	0.934	0.039	0.525	0.532	0.531	0.950
	Paik	-0.009	0.515	0.555	0.555	0.926	0.051	0.499	0.535	0.532	0.932
	AIPW	-0.012	0.520	0.553	0.553	0.922	0.011	0.501	0.534	0.534	0.940
Active $-$ *Control*	MI	0.013	0.708	0.705	0.705	0.948	0.073	0.710	0.658	0.654	0.966
	Paik	0.012	0.711	0.709	0.709	0.942	0.073	0.662	0.661	0.657	0.940
	AIPW	0.013	0.721	0.714	0.714	0.954	0.026	0.675	0.667	0.667	0.950

Abbreviations: MI, multiple imputatoin; Paik, Paik's mean imputation; AIPW, AIPW-form DR imputation; SE, standard error; RMSE, root mean square error; MCSD, Monte Carlo standard deviation; Covp, coverage probability.
MI constructed 200 data sets within each repeat; 200 Bootstrap resampling was done within each repeat for Paik and AIPW.
SE: Averaging standard error across 500 Monte Carlo repeats.

4.6.2 Sensitivity Analysis

Simulation studies of sensitivity analysis have been explored very little in the previous literature. Here we design a jump-to-reference-based simulation study to evaluate the performance among the three imputation methods. δ-based sensitivity analysis is not performed, but the idea would be similar. An essential point in the sensitivity analysis simulation is how to quantify the "truth". In this simulation, the true value of mean $Y|arm = Control$ would be identical to the primary analysis because, under jump-to-reference definition, participants in the control arm who dropped from the study should still follow their original distribution. However, the true value of mean $Y|arm = Active$ would differ from the primary analysis because the "truth" of $Y_{mis}|arm = Active$ will follow the control arm's distribution.

Table 4.1 presents the results of this sensitivity analysis. Although all the biases are very small, AIPW imputation attains the smallest bias in each estimation across the three imputation methods. Comparing with multiple imputation, AIPW imputation always has a bias about 3 to 4 times smaller than multiple imputation. Consistent results of average standard errors, root mean square errors, and Monte Carlo standard deviation are reported among the three methods. All three methods obtain acceptable coverage probabilities, while multiple imputation repeatedly reports the highest coverage probabilities among all methods.

4.6.3 Summary and Discussion

Our analyses indicate that the three imputation methods perform very well, with small biases and acceptable efficiencies and coverage probabilities. Since the imputation models are correctly specified for all methods, they all obtain consistent estimates as we expected. In this simulation study, AIPW-form DR imputation has similar efficiency as multiple imputation and Paik's imputation, even when the attrition rate is modest.

In the sensitivity analysis, we are happy to observe that the AIPW-form of DR imputation

outperforms multiple imputation and Paik's imputation in bias. Its efficiency is as good as the other two methods. Thus, we anticipate that the double robustness of this method may provide a substantial advantage without substantial loss of efficiency.

As we discussed in section 4.5.2, if subjects in both arms have distributions that only differ in means, and this difference is a linear function of time, then the jump-to-reference approach can be seen to be equivalent to a δ-based approach. An interesting finding from this simulation study is that, under a model similar to the generating model 4.8, the jump-to-reference approach cannot be transferred to a δ-based approach, even though the model is linear and both arms are following multivariate normal distributions with non-identical mean. That is because when $Pr(R_j = 1)$ is a function of $Y_1, ..., Y_{j-1}$, the distributions $f(y_{ij}|R_{iM} = 0)$ for subjects who dropped from the study, and $f(y_{ij}|R_{iM} = 1)$ for the completers can be distinct.

Imagine an elementary setting with only two time points, subjects who will drop later may have a lower mean of Y_1 ($E(Y_{i1}|R_{iM} = 0) < E(Y_{i1}|R_{iM} = 1)$), but an average steeper slope (higher increment), due to a higher mean random slope than that of completers. For dropouts i in the active arm, imputing its $\hat{Y}_{i2}$ through a jump-to-reference approach should add the averaging increment from the control arm's distribution to its Y_{i1}, which accounts for its random intercept b_{i1}, but not its random slope b_{i2}, from time 1 to time 2. By contrast, the δ-based approach adds an amount δ to its $\hat{Y}_{i2}$, which the $\hat{Y}_{i2}$ is drawn from $f(y_{ij}|R_{iM} = 0)$, thus accounts for its b_{i1} and b_{i2}.

In conclusion, the AIPW-form of the DR imputation method is a competitive approach under either MAR or MNAR, as supported by our simulation studies. Furthermore, we suggest that in sensitivity analysis, whether performing a jump-to-reference adjustment or a δ-based adjustment depends on the desired assumptions regarding distribution of the data. These choices may be discussed explicitly in the context of defining the corresponding estimands. An arbitrary choice of the method may lead to unexpected results.

4.7 Analyses in the Alzheimer's Trials

In this section, the three imputation methods are applied to the two Alzheimer's disease trials described in section 4.3, one for mild cognitive impairment patients, and one for mild-to-moderate AD patients. For both applications, a subgroup of the population is taken for the analysis stratified by Apolipoprotein E (APOE) e4 status. APOE is a gene that is highly correlated to AD progression. Studies show that individuals with APOE e4 positive are more likely to progress to AD than individuals without.

As in the simulation study, we test estimation of both the secondary estimand and the *primary estimand*, and compare results among different imputation methods. Standard errors and 95% confidence intervals are also presented. Standard errors of multiple imputation are calculated following Rubin's rule with 200 data sets, while other approaches use 200 times Bootstrap to derive the standard errors.

4.7.1 DHA Trial

Mild to moderate AD is a stage prior to severe AD. In the DHA trial [48], participants were all clinically diagnosed as AD or probable AD at the time of enrollment. Note that, the false positive rate of AD among APOE e4 negative individuals in the trial is likely lower than in the general MCI population, since they have not been diagnosed with AD. Under this consideration, we take the subgroup of APOE e4 negative subjects for the analysis, because we believe that the drug may bring more benefits to them than to APOE e4 positive individuals. Within this subgroup, the dropout rates are 27.4% and 18.6% for DHA and Placebo groups, respectively.

The change of ADAS-Cog total score was used as the primary outcome in the original publication. Following their analysis methods, a mixed effects model was applied, and fixed effects include baseline ADAS-Cog score, baseline Mini-Mental State Examination (MMSE) score, gender, arm, categorical visits, and the interaction of arm by visits.

Primary Analysis

The primary analysis is under the assumption of MAR. The three imputation methods are compared to each other, along with the mixed model with repeated measures (MMRM) as shown in Table 4.2. For the secondary estimand, all three imputation models provide similar mean estimates of ADAS-Cog change at 18 months of around 12.7 for the placebo group and 8.8 for the DHA group, while MMRM reports slightly smaller estimates for both arms. This phenomenon happens at months 6 and 12 as well, and the detailed results can be found in Appendix B. Standard errors are similar across all approaches. In particular, Paik's imputation and AIPW imputation are almost identical in standard errors for both arms, while multiple imputation has a close but smaller standard error in the placebo group, and a greater standard error in the DHA group.

For estimating the *primary estimand*, the three imputation methods obtained consistent results, and mean estimates ranged from -3.86 to -3.91. At the same time, the MMRM attains a smaller estimate of the treatment effect of -3.77. Multiple imputation attains a similar standard error to MMRM, while Paik's mean imputation and AIPW attain smaller standard errors.

In summary, the three imputation methods derive consistent estimates among each other, while MMRM shows minor differences compared with the imputation methods. Focusing on the least squares mean estimate of the difference between arms, we found all imputation methods enlarge the mean estimates compared to the original MMRM estimates, and remain the statistically significant results, indicating that the DHA does have benefits in this particular APOE e4 negative group.

Sensitivity analysis

Both δ-based and reference-based methods are performed in the sensitivity analysis to evaluate the estimation of the *primary estimand* and the secondary estimand when the dropout is not at random. In our analyses, jump-to-reference (J2R) represents the reference-based approach, and three δ's are selected for the δ-based approach. Results are displayed in Table 4.3.

Table 4.2: Primary Analyses for DHA trial Subgroup

ADAS-Cog score 18 month	Est	SE	Lower	Upper	Est	SE	Lower	Upper
	MMRM				MI			
Placebo	12.43	1.16	10.13	14.73	12.72	1.18	10.41	15.03
DHA	8.66	1.01	6.66	10.66	8.85	1.02	6.85	10.85
DHA - Placebo	-3.77	1.55	-6.80	-0.74	-3.87	1.56	-6.93	-0.80
	Paik				AIPW			
Placebo	12.69	1.22	10.29	15.09	12.69	1.22	10.31	15.08
DHA	8.83	0.95	6.96	10.70	8.78	0.95	6.91	10.65
DHA - Placebo	-3.86	1.48	-6.77	-0.96	-3.91	1.47	-6.80	-1.03

Data Source: The Alzheimer's Disease Cooperative Study (https://www.adcs.org/).
Abbreviations: Est, estimate; SE, standard error; Lower, lower bound of 95% confidence interval; Upper, upper bound of 95% confidence interval.

Overall, across the different imputation methods, in the placebo arm they obtain similar mean estimates and standard errors for the secondary estimand. In the DHA arm, mean estimates and the corresponding standard errors differ between multiple imputation and the other two imputation methods for the secondary estimand, leading to a difference in the estimates for the *primary estimand* as well. Paik's imputation and AIPW imputation repeatedly show smaller mean estimates for the DAH arm than does multiple imputation, which leads to an increase by about 10% ($\delta = 3$) in the estimated difference in ADAS-Cog score between arms.

Furthermore, multiple imputation obtains greater standard errors estimates for both estimands, compared with Paik and AIPW imputation. Specifically, in the jump-to-reference condition, multiple imputation has an approximately 17% and 20% increase in standard errors than the other two imputation procedures, respectively. These differences are smaller in the δ-based approaches.

Comparing the sensitivity analysis results with the primary analysis can be of interest as well. Either the jump-to-reference or δ-based method provides a conservative way to test the drug effect when dropouts happen. Regarding the definition of the jump-to-reference adjustment, the mean estimate of the secondary estimand in the placebo group remains as what it is in the primary analysis, while the mean estimate of the secondary estimand in the DHA group increases from about 8.82 to 9.36 across all three imputation methods. This change brings a reduction

in the estimate of the *primary estimand* from about -3.87 to -3.34. However, even assuming a conservative effect by using jump-to-reference estimate, in this trial, the differences in change of ADAS-Cog score between arms are still significant across the three imputation methods.

Δ-based approaches show the same direction of a higher mean ADAS-Cog score in the DHA arm and reduce the mean difference between arms. As presented in Table 4.3, the differences between δ-adjusted estimates and primary analysis become larger when δ increases, and the values of the jump-to-reference estimate are between $\delta = 1$ and $\delta = 2$ for all three imputation methods. Multiple imputation indicates that the difference between arms turns to be non-significant after assuming a shift of $\delta = 2$ among study dropouts in the DHA arm (95% confidence interval upper bound = 0.05). By comparison, AIPW imputation indicates that it requires a shift of $\delta = 3$ to have this boundary effect. Paik's imputation reports that this critical value of shift would be even larger than 3. The exact critical value (upper bound = 0) of δ could be detected by fitting this model for more accurate numbers of δ. For example, δ_0 could be a number near but smaller than 2 and 3 for multiple imputation and AIPW, respectively, and near but greater than 3 for Paik's imputation. This implies that the sensitivity analysis using DR imputation indicates a more robust treatment effect than the other methods in this trial.

Table 4.3: Sensitivity Analyses for DHA trial Subgroup

ADAS-Cog score 18 month	Est	SE	Lower	Upper	Est	SE	Lower	Upper	Est	SE	Lower	Upper
jump-to-reference			MI				Paik				AIPW	
Placebo	12.65	1.27	10.16	15.14	12.76	1.24	10.33	15.18	12.68	1.23	10.26	15.09
DHA	9.40	1.10	7.25	11.55	9.27	0.94	7.43	11.10	9.34	0.93	7.51	11.17
DHA - Placebo	-3.25	1.64	-6.47	-0.03	-3.49	1.36	-6.16	-0.82	-3.34	1.36	-6.01	-0.67
$\delta = 1$												
Placebo	12.72	1.19	10.38	15.06	12.72	1.22	10.32	15.12	12.69	1.22	10.31	15.07
DHA	9.24	1.03	7.21	11.26	9.06	0.97	7.16	10.95	9.09	0.96	7.21	10.98
DHA - Placebo	-3.48	1.58	-6.58	-0.38	-3.66	1.50	-6.60	-0.72	-3.60	1.49	-6.51	-0.68
$\delta = 2$												
Placebo	12.72	1.21	10.34	15.09	12.72	1.22	10.32	15.12	12.69	1.22	10.30	15.07
DHA	9.63	1.05	7.57	11.68	9.35	0.98	7.43	11.28	9.41	0.98	7.50	11.32
DHA - Placebo	-3.09	1.60	-6.23	0.05	-3.36	1.52	-6.34	-0.39	-3.28	1.51	-6.23	-0.33
$\delta = 3$												
Placebo	12.71	1.23	10.30	15.13	12.71	1.22	10.31	15.11	12.68	1.22	10.30	15.06
DHA	10.01	1.07	7.92	12.10	9.65	1.00	7.69	11.61	9.72	0.99	7.77	11.67
DHA - Placebo	-2.70	1.63	-5.90	0.50	-3.06	1.54	-6.08	-0.05	-2.96	1.53	-5.95	0.03

4.7.2 Donepezil Trial

MCI is an early stage of AD, and in this trial, which had progression to AD as the primary outcome, the overall rate of progression from MCI to AD was 16% per year. There was no significant difference in the probability of progression to AD in the Donepezil group compared with the placebo group. Since the cohort was not clinically diagnosed with probable AD at the onset of the study, many subjects with an APOE e4 negative genotype may actually have been false positives, and may not progress to AD eventually. Thus in our application, we took the APOE e4 positive subgroup for the Donepezil trial so that a drug effect could be more easily detected.

As in our analysis of the DHA trial, the change of ADAS-Cog score is used here as the primary outcome. The fixed effects include baseline score of ADAS-Cog, age, gender, education, arm, categorical visit, and interaction of arm by visit. Again we estimate both the secondary estimand and the *primary estimand*: the average change from baseline of ADAS-Cog total score at month 36 for each arm; and the difference in the change of ADAS-Cog total score between arms at month 36. The average change score at other time points is also estimated, and details can be found in Appendix B.

Primary Analysis

Similar to the DHA trial analysis, we compare the three imputation methods among each other and with MMRM. As shown in Table 4.4, the mean estimates of both the secondary estimand and the *primary estimand* are similar among the three imputation methods. Compared with MMRM, the three imputation methods increase the estimates of ADAS-Cog change in the placebo arm and also the differences in ADAS-Cog change between the Donepezil arm and placebo arm.

For the standard errors, we observe comparable results as in the DHA trial. Multiple imputation has standard errors similar to MMRM, while Paik's imputation and AIPW imputation are almost identical. Multiple imputation and MMRM have about 14% and 17% higher standard errors than Paik's and AIPW imputations in Donepezil arm, for the secondary estimand and the

primary estimand respectively.

Although there is still no significant effect seen for Donepezil in slowing the change in ADAS-Cog, the three imputation methods consistently attain larger differences between arms, indicating that the primary analysis model may not be specified correctly.

Table 4.4: Primary Analyses for Donepezil Trial Subgroup

ADAS-Cog score 3 year	Est	StdErr	Lower	Upper	Est	StdErr	Lower	Upper
	MMRM				MI			
Placebo	6.24	0.82	4.62	7.87	6.40	0.84	4.76	8.04
Donepezil	4.85	0.84	3.20	6.50	4.82	0.85	3.15	6.49
Donepezil - Placebo	-1.39	1.18	-3.70	0.91	-1.58	1.20	-3.93	0.77
	Paik				AIPW			
Placebo	6.42	0.89	4.68	8.16	6.43	0.89	4.69	8.18
Donepezil	4.82	0.73	3.40	6.25	4.87	0.73	3.43	6.30
Donepezil - Placebo	-1.60	1.02	-3.59	0.39	-1.57	1.02	-3.58	0.44

Sensitivity analysis

As described in the DHA trial analysis, the two sensitivity approaches, jump-to-reference and δ-based analysis, are combined with the three imputation methods and then applied to the data to test the impact on conclusions if the dropout is not at random. The results are reported in Table 4.5. Among the three imputation methods, they all have similar mean estimates by jump-to-reference adjustment, while multiple imputation has 27% larger standard errors than Paik's and AIPW imputations when estimating the *primary estimand*. With the δ-based adjustment, multiple imputation always has the smallest mean estimate of the difference between arms, while Paik's imputation has the biggest estimate. The difference in estimates among the imputation methods tends to increase along with the δ increases, and when $\delta = 3$, Paik's imputation obtains a 40% larger estimated effect than multiple imputation. The mean estimates of AIPW imputation remain in the intermediate among the three imputation methods. In contrast to the δ-based method, the jump-to-reference method reports a small estimate of the secondary estimand in the Donepezil

arm and a large estimate of the *primary estimand*, indicating that the reference group's distribution and drug group's distribution is similar. The secondary estimand for the placebo group has similar estimates in all conditions, which is as expected.

The statement of no difference between the placebo group and Donepezil group distributions is also supported by comparing the primary analysis with the sensitivity analysis. In the primary analysis, the three imputation methods agree with the estimates of 6.4 for the secondary estimand in the placebo group, 4.85 for the secondary estimand in the Donepezil group, and -1.6 for estimating the *primary estimand*. There are almost no changes of those estimates after jump-to-reference sensitivity adjustment, demonstrating that even if those patients in the Donepezil group follow the placebo group's distribution after dropping out, their mean ADAS-Cog change does not differ from the mean assuming they continue to take the Donepezil. This result is consistent with the original finding of the trial that participants in the Donepezil group do not show a significant difference in ADAS-Cog change at months 36 compared with participants in the placebo group. Δ-based adjustment is an approach that artificially adds a difference to the distributions between the reference group and active group. Furthermore, as we observed in the primary analysis, MMRM obtains an estimate of the *primary estimand* at -1.39, which is smaller than that of imputation methods. After comparing with δ-based adjustments, if we took a value of δ between 0.5 to 1, the estimate of the *primary estimand* from the three imputation methods would be comparable with MMRM's estimate in the primary analysis.

4.7.3 Summary

Figure 4.1 illustrates the estimates of the *primary estimand* at each time point for both jump-to-reference approach and three δ-based approaches, comparing with the estimate of MMRM in the primary analysis (red line). In summary, the conservative methods in sensitivity analysis clearly make changes in the estimates for the *primary estimand*. They all narrow the treatment effect and reduce the difference between arms (the *primary estimand*) to some degree. Δ-based

Table 4.5: Sensitivity Analyses for Donepezil trial Subgroup

ADAS-Cog score 3 year	Est	StdErr	Lower	Upper	Est	StdErr	Lower	Upper	Est	StdErr	Lower	Upper
jump-to-reference			MI				Paik				AIPW	
Placebo	6.36	0.86	4.69	8.04	6.39	0.89	4.64	8.14	6.40	0.90	4.64	8.16
Donepezil	4.80	0.88	3.08	6.53	4.83	0.70	3.46	6.20	4.87	0.71	3.49	6.26
Donepezil - Placebo	-1.56	1.22	-3.95	0.83	-1.56	0.95	-3.43	0.30	-1.53	0.96	-3.41	0.36
$\delta = 1$												
Placebo	6.40	0.84	4.75	8.04	6.42	0.89	4.68	8.16	6.43	0.89	4.69	8.18
Donepezil	5.13	0.86	3.46	6.81	5.06	0.73	3.62	6.50	5.12	0.74	3.67	6.57
Donepezil - Placebo	-1.26	1.20	-3.62	1.10	-1.36	1.02	-3.36	0.64	-1.31	1.03	-3.33	0.70
$\delta = 2$												
Placebo	6.39	0.84	4.74	8.04	6.42	0.89	4.67	8.16	6.43	0.89	4.69	8.18
Donepezil	5.45	0.86	3.76	7.13	5.30	0.74	3.85	6.76	5.37	0.75	3.90	6.84
Donepezil - Placebo	-0.95	1.21	-3.31	1.42	-1.12	1.02	-3.12	0.89	-1.06	1.03	-3.08	0.96
$\delta = 3$												
Placebo	6.39	0.85	4.73	8.05	6.42	0.89	4.68	8.16	6.43	0.89	4.68	8.18
Donepezil	5.76	0.87	4.07	7.46	5.54	0.76	4.06	7.02	5.62	0.76	4.13	7.12
Donepezil - Placebo	-0.63	1.22	-3.01	1.75	-0.88	1.03	-2.90	1.14	-0.81	1.04	-2.85	1.23

adjustment attains estimates strictly in order with the values of δ, and the difference of estimates among δ-based approaches increases with time. It is straightforward to expect a change between primary analysis and δ-based sensitivity analysis, no matter what the two distributions of the control arm and active arm are. Jump-to-reference adjustment, however, tells another story because it does consider the difference of distributions between arms. If the two distributions are approximately identical, the jump-to-reference approach will make no change to the estimates in the primary analysis, as in the Donepezil trial example. An essential aspect of the δ-based method is that it can be used to identify the critical value of δ that makes the significant effect of treatment disappear, by tracking the trend of estimates from a sequence of ordinal values of δ. This exercise identifies the magnitude of the MNAR effect required to alter the conclusions of the trial.

4.8 Discussion

In this study, we have shown that any consistently estimated imputation model can be used for joint modeling within a PMM framework to obtain consistent estimates under the assumption of MNAR. We also provide explicit algorithms for these joint models, which are constructed in a straightforward manner by incorporating the PMM adjustment into the imputation model

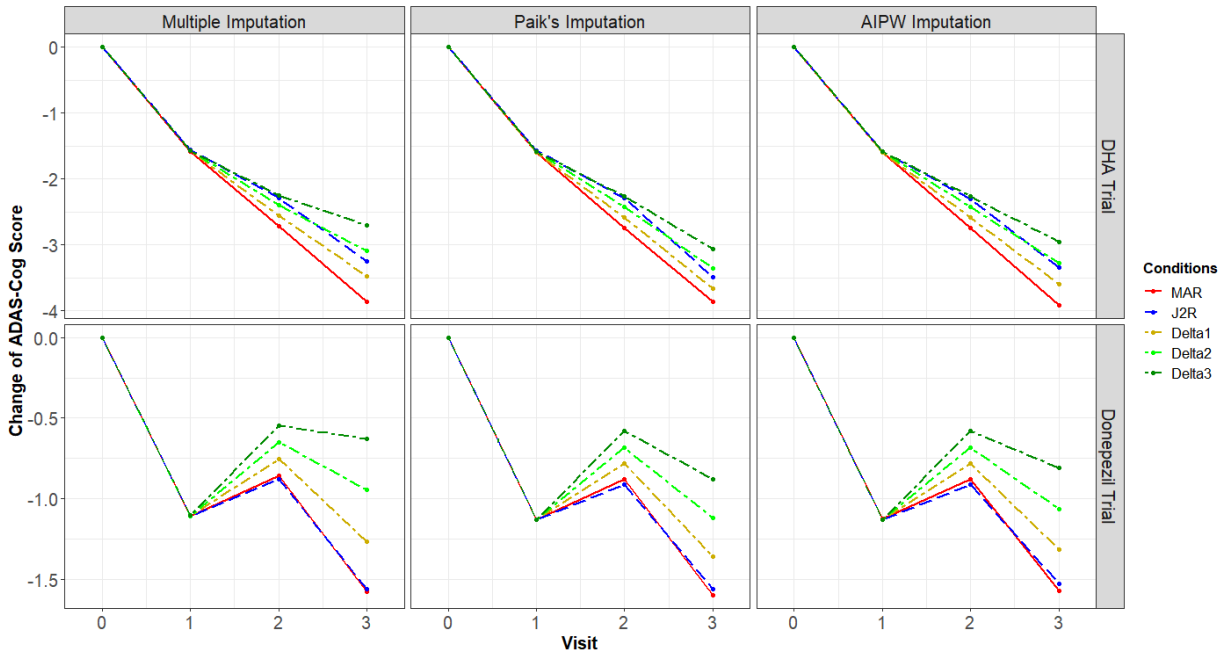

Figure 4.1: Sensitivity Analyses by methods in the two AD trials

procedure. We discuss the possibility that reference-based PMM can be restated as a δ-based PMM under different conditions. Furthermore, the AIPW-form DR imputation obtains estimates that perform as well as multiple imputation and Paik's mean imputation for the primary analysis assuming MAR. At the same time, it outperforms the other two imputation methods in bias for sensitivity analysis assuming MNAR, based on the simulation study. This finding supports the viewpoint that DR imputation is a safer and more robust approach for sensitivity analysis in randomized trials. Finally, we capture that in the primary analyses of the two AD trials, MMRM repeatedly obtains smaller treatment effects than the three imputation methods, which may lead to further discussions about the appropriate statistical modeling for primary analysis in randomized trials with dropouts.

There are several possibilities to explain why MMRM obtains different estimates from the other imputation methods. Under MAR, MMRM will derive consistent estimates only if both the mean and the covariance structures are correctly specified. Since the other three imputation models control identical covariates as in MMRM, one possibility would be that the covariance

structure is wrong. Another possibility is that the data is MNAR. As we described in section 4.7, the three imputation methods can eventually achieve similar estimates to MMRM by selecting a particular δ with δ-based adjustment, which quantifies the degree of effect under MNAR needed to explain these differences. If the difference is due to model misspecification rather than MNAR missingness, DR imputation would be more favored because it can overcome any misspecification in the outcome model by using propensity scores.

Workflow is an alternative evaluation metric to compare methods. In order to construct a single completely imputed data set in longitudinal data with monotone dropout and fully observed covariates, multiple imputation turns out to have a light workflow, with only $M-1$ linear regression models needed since chained equations are not needed anymore. Paik's mean imputation has a somewhat heavier workflow with $M(1+M)/2-1$ models to be fitted. In our example, DR imputation uses Paik's mean imputation as the outcome model; thus, the workflow of DR imputation is even heavier than Paik's mean imputation with an additional $M-1$ logistic regression models to be fitted to estimate the propensity score. Due to the complicated form of the analytical variance of Paik's imputation, if the bootstrap resampling procedure is chosen to estimate standard errors for Paik's and DR imputation methods, and if the size B of Bootstrap is identical to the size B of data sets created by multiple imputation, the total workflow would stay in this ranking (MI, Paik's imputation, DR imputation) among the three methods. In our DHA trial application with $B = 200$, however, the three methods all caused around 110 seconds for obtaining the final least squares mean of estimates. Although the time is similar here, workflow is still an aspect that should be considered.

Continuing with the above topic, DR imputation may be simplified in some circumstances by choosing other outcome models instead of Paik's mean imputation. In chapter 3, we proposed a simplified DR imputation method with a mixed effects model, and the workflow would be $M-1$ logistic regression models plus one mixed effects model, which would help save time.

Another point we would like to make is that sensitivity analysis may be more meaningful

when the primary analysis shows a significant treatment effect. If the treatment effect is very small, in other words, if the distributions of the control arm and active arm are similar, conducting a sensitivity analysis using the PMM framework is not helpful. The reference-based adjustment would not make any change due to the similarity between the control arm and active arm; δ-based adjustment artificially makes differences between the arms but in the wrong direction.

Finally, we also would like to make some discussions specifically for AD trials. First, as we observed in the Donepezil trial, the ADAS-Cog total score may not be the best measurement for people in the early stage of the disease, such as the MCI patients in this trial. Also, these subjects may decline slowly in the first months of the trial, and they may decline faster later; thus, a non-linear trend may be more suitable to model. Secondly, although we obtained encouraging results from the subgroup analysis in the DHA trial, investigators should be cautious about performing and reporting such ad hoc subgroup analysis. Lastly, we encourage and recommend investigators to collect specific reasons for dropouts and also to continue to follow the participants even if they have discontinued study medication. The additional information could significantly help statisticians to construct more reliable and valid models in the future.

4.9 Afterthoughts before next Chapter

So far, we have reviewed popular imputation methods for dropouts in randomized trials, and constructed algorithms for investigators to use in the sensitivity analysis. The doubly robust method has been shown to be a competitive approach to address monotone dropouts for longitudinal data.

In the last chapter, we will discuss the findings from Chapters 2 to 4, and make overall conclusions about the three projects. We will also explain our future work and continue our research work in this area.

This chapter, in full, has been prepared for submission for publication as ”Qiu, Yuqi;

Feldman, Howard H.; Messer, Karen S. *Doubly Robust Imputations for Randomized Trials with Monotone Dropout under Missing not at Random: Applications in Alzheimer's Trials*". The dissertation author was the primary author on this paper.

Chapter 5

Conclusions and Future Work

This dissertation illustrates novel statistical approaches using *Doubly Robust* estimation methods to address dropout-related bias in longitudinal data, with applications and discussion in Alzheimer's disease clinical trials. Starting with a study exploring cognitive heterogeneity in probable AD, we investigated the common question in AD trials about correcting for potential bias due to discontinuation, evaluating both classic, widely used models and novel approaches, either under MAR or under MNAR assumptions regarding the dropout. We reviewed regulatory preferred methods such as MMRM, GEE, and multiple imputation; comparable methods such as WGEE, IPW, Paik's mean imputation; and novel alternatives such as doubly robust methods, including our proposed novel form of longitudinal doubly robust imputation estimators.

Specifically, in chapter two, we identified heterogeneity in cognitive profiles of probable AD patients by principal component analysis and Gaussian model-based clustering across independent cohorts. We proposed a decision rule to classify patients as typical or atypical AD using neuropsychological tests. The identified atypical group was associated with younger age, male sex, lower probability of APOE e4, less severe global dementia, higher depression scores, lower Braak stage at autopsy, and slower cognitive decline. We demonstrated that distinct cognitive profiles among clinically diagnosed probable AD patients could be consistently identified. This

heterogeneity is associated with tangle pathology and with different rates of decline.

In chapter three, we proved that most doubly robust estimators for longitudinal data could be written in a more straightforward substitution form, which may be easier to understand and apply under MAR. Moreover, a simplified AIPW form was constructed. We illustrated the substitution approach with the simplified AIPW form using an AD trial with comparison to MMRM. Simulation studies were also performed comparing classic methods, alternative methods, and doubly robust methods under four different situations regarding the model specifications. Based on the theoretical properties and results supported by simulation studies, we confirmed that the doubly robust method performs well in bias and efficiency comparing other outcome-model-based approaches in dealing with longitudinal data with monotone dropouts when the outcome model is correctly specified. On the other hand, the doubly robust estimator obtains unbiased estimators with acceptable efficiency when the outcome model is misspecified, but the missingness model is correct. Furthermore, the imputation approach we presented has the advantage of computational simplicity and transparency compared to existing doubly robust approaches.

In chapter four, we reviewed several imputation methods and the PMM framework, and then constructed algorithms that combine the imputation approaches and PMM framework under MNAR. Regulatory requirements were discussed in this chapter, and we compared the doubly robust imputation to multiple imputation, which is the predominant method used in practice for handling missing data. Two AD trials with different stages of disease conducted by ADCS were used as example applications. We identified that imputation methods all increased the estimated treatment effects compared to the widely used MMRM. Simulation studies under either MAR or MNAR were performed. We described the connections between the MNAR statistical framework and the new framework on estimands proposed by the regulatory agencies. Furthermore, we again confirmed that doubly robust imputation performs as well as other imputation methods when the outcome model is correctly specified. We suggest that the doubly robust imputation method is

competitive due to its robustness properties.

In conclusion, the doubly robust imputation method is recommended for longitudinal data with monotone dropout because (1) through both theoretical and practical proofs, when assuming the outcome model is correctly specified (both mean structure and covariance structure for MMRM), the DR imputation method can obtain unbiased estimators, with acceptable standard errors; (2) once the outcome model is misspecified (including insufficient covariate adjustment), which is a usual case in real-life clinical trials settings, DR imputation can still obtain unbiased estimators by specifying a correct missingness model, while other outcome-only- model-based methods would have biased estimators due to dropouts; (3) when both outcome model and missingness model are not able to be specified correctly, our simulation studies support that DR imputation method obtains better estimators than other methods with less bias and better RMSE.

The main limitation of the DR imputation method is the increased workflow. Our proposed AIPW-S form is simpler and easier to calculate than existing longitudinal DR estimators in the literature, thus alleviating some workflow burden. The trade-off is that the AIPW-S estimator may lose some efficiency. Another obstacle for practical use is the difficulty of obtaining the analytical variance for the DR imputation method. Considering the DR estimator belongs to the M-estimator class, a sandwich estimator would be appropriate for estimating the variance of the DR method. Our next step would be finding a more straightforward and interpretable way to derive the variance for the DR imputation method and program it in commonly used statistical software.

Some points in the context of AD clinical trials are also worth discussion. First, investigators would be eager to determine the particular circumstances in clinical trials in which the DR imputation method should be applied. As discussed above, when the outcome model is not convincing, the DR imputation method would be more favorable. For example, sometimes the number of covariates (including demographics, clinical characteristics, biomarkers, etc.) that needs to be controlled in the model is large, which can cause the MMRM to fail to converge; or the unstructured covariance matrix may not converge due to a large number of visits. In these

scenarios, the DR imputation method may utilize the missingness model to control the covariates and keep the outcome model simple and clean.

Secondly, in chapter 4, we suggest that investigators continue collecting the assessments even if the participants dropped out from study treatment. This is a budget issue in most regulated trials, and whether it is worth continuing follow-up may be debated. As statisticians, we suppose that increasing information (data) would better assist us in building more precise models, obtaining unbiased estimators, and enhancing the power. We also believe that additional data can help us determine whether the missing at random assumption is correct. Due to the limited budget, collecting all post-dropout data may be unnecessary. It may be more feasible to collect the final visit or just a few subsequent visits.

Thirdly, many trials have been paused or remotely assessed due to the COVID-19 pandemic, leading to a surge of missing data. The DR imputation method is proposed to deal with monotone dropout. It may be extended using the PMM framework to help in this situation with a lot of intermediate missing data. This would be an important topic for future research.

Several other topics are also valuable and can continue to be investigated in the future. Firstly, we would like to discuss the performance of doubly robust imputation methods incorporating the PMM framework under MNAR in different model specifications. It would be interesting to evaluate it when there are misspecifications in either the outcome model or missingness model, or both, compared to other imputation approaches. Secondly, the variance is worthy of further study. We obtained variance estimates through the bootstrap resampling procedure in this dissertation. It is worth investigating the analytical variance and the corresponding sandwich estimator for our proposed doubly robust estimator in the future. Thirdly, as we recommend recording more information about dropouts and other biomarkers in clinical trials, the approach to utilize the additional information with our proposed methods would be an interesting topic.

Appendix A

Additional simulation results for Chapter 3

A.1 Simulation performance metrics for regression coefficients

For the estimand of the vector of regression coefficients, we report the following performance metrics, averaged across the six coefficients except intercept: (1) Average percent bias for each method. For each method, we took the absolute value of the percentage bias for each coefficient β_p and then calculated the average across the six coefficients. (2) Z-score (among estimators by conditions) of the relative RMSE's. For each estimator in each of 22 conditions (7 estimators by 4 "correctness" conditions, minus duplicate conditions), the RMSE was calculated and divided by β_p. Then we computed the z-score across the 23 conditions, and finally averaged the z-scores across the 6 coefficients. (3) Z-score (among estimators by conditions) of the interval score. We standardized and averaged the interval scores with the same process as for the RMSE. (4) Average coverage probability. The coverage probability of a 95% confidence interval for each $\beta'_p s$ was averaged over the 6 coefficients, for each method.

Table A.1: Comparisons of $E(Y_3)$ among methods in six evaluations: Extreme Scenario

	Bias	RMSE	Ints	Covp	MCSD	Ave SE	Bias	RMSE	Ints	Covp	MCSD	Ave SE
	Y correct P correct						Y correct P incorrect					
AIPW-I	0.01	0.36	1.70	0.95	0.36	0.35	0.00	0.34	1.66	0.95	0.34	0.34
B & R	0.00	0.32	1.51	0.95	0.32	0.32	-0.00	0.32	1.53	0.95	0.32	0.32
AIPW-S	-0.00	0.40	1.80	0.94	0.40	0.36	0.10	0.37	1.71	0.95	0.36	0.35
Paik	-0.00	0.32	1.51	0.95	0.32	0.32	-0.00	0.32	1.51	0.95	0.32	0.32
WGEE	-0.10	0.46	2.87	0.85	0.45	0.44	0.12	1.50	5.12	0.90	1.50	0.51
GLS	-0.05	0.32	1.53	0.95	0.31	0.31	-0.05	0.32	1.53	0.95	0.31	0.31
GEE IND	-0.15	0.35	1.65	0.93	0.31	0.32	-0.15	0.35	1.65	0.93	0.31	0.32
	Y incorrect P correct						Y incorrect P incorrect					
AIPW-I	-0.01	0.48	2.19	0.93	0.48	0.44	-1.51	1.56	33.84	0.05	0.39	0.37
B & R	0.33	0.49	2.50	0.82	0.36	0.36	-1.56	1.60	36.85	0.01	0.33	0.34
AIPW-S	-0.09	1.32	5.53	0.84	1.31	0.90	-1.33	1.57	21.53	0.41	0.84	0.74
Paik	-1.70	1.73	41.43	0.00	0.35	0.36	-1.70	1.73	41.43	0.00	0.35	0.36
WGEE	-1.26	1.87	30.49	0.29	1.39	0.55	-2.99	3.17	79.20	0.06	1.03	0.62
GLS	-3.18	3.19	103.45	0.00	0.31	0.32	-3.18	3.19	103.45	0.00	0.31	0.32
GEE IND	-4.42	4.43	155.83	0.00	0.28	0.28	-4.42	4.43	155.83	0.00	0.28	0.28

Table A.2: Comparisons of $\hat{\beta}_p$ among methods in four evaluations: Extreme Scenario

	Bias	RMSE	Ints	Covp	Bias	RMSE	Ints	Covp
	Y correct P correct				Y correct P incorrect			
AIPW-I	0.01	-0.52	-0.49	0.95	0.01	-0.52	-0.49	0.95
B & R	0.05	0.42	0.05	0.96	0.01	-0.47	-0.34	0.94
AIPW-S	0.01	-0.50	-0.49	0.95	0.01	-0.51	-0.49	0.94
Paik	0.00	-0.54	-0.49	0.95	0.00	-0.54	-0.49	0.95
WGEE	0.14	0.87	0.19	0.87	0.16	0.34	-0.06	0.88
GLS	0.01	-0.54	-0.49	0.94	0.01	-0.54	-0.49	0.94
GEE IND	0.01	-0.53	-0.49	0.92	0.01	-0.53	-0.49	0.92
	Y incorrect P correct				Y incorrect P incorrect			
AIPW-I	0.01	-0.46	-0.47	0.94	0.10	-0.27	-0.21	0.39
B & R	0.11	-0.02	-0.01	0.86	0.58	1.86	2.42	0.56
AIPW-S	0.02	-0.20	-0.39	0.86	0.10	-0.20	-0.22	0.56
Paik	0.22	0.20	0.38	0.29	0.22	0.20	0.38	0.29
WGEE	0.45	0.73	1.01	0.15	0.61	1.15	2.00	0.03
GLS	0.50	0.95	1.40	0.05	0.50	0.95	1.40	0.05
GEE IND	0.55	1.08	1.58	0.04	0.55	1.08	1.58	0.04

Appendix B

Additional application results for Chapter 4

B.1 DHA Trial

Table B.1: Primary Analyses for DHA Trial Subgroup, full results for treatment effects

	Est	StdErr	Lower	Upper	Est	StdErr	Lower	Upper
	MMRM				MI			
DHA - Placebo 1 year	-1.57	0.85	-3.23	0.09	-1.58	0.85	-3.24	0.08
DHA - Placebo 2 year	-2.85	1.19	-5.19	-0.52	-2.72	1.20	-5.08	-0.36
DHA - Placebo 3 year	-3.77	1.55	-6.80	-0.74	-3.87	1.56	-6.93	-0.80
	Paik				AIPW			
DHA - Placebo 1 year	-1.60	0.81	-3.18	-0.02	-1.60	0.81	-3.18	-0.02
DHA - Placebo 2 year	-2.75	1.21	-5.11	-0.38	-2.74	1.20	-5.10	-0.38
DHA - Placebo 3 year	-3.86	1.48	-6.77	-0.96	-3.91	1.47	-6.80	-1.03

B.2 Donepezil Trial

Table B.2: Primary Analyses for DHA Trial Subgroup, full results for $\hat{Y}$

	Est	StdErr	Lower	Upper	Est	StdErr	Lower	Upper
	MMRM				MI			
Placebo 1 year	5.77	0.65	4.50	7.05	5.81	0.65	4.55	7.08
Placebo 2 year	6.53	0.90	4.75	8.32	6.66	0.91	4.87	8.45
Placebo 3 year	12.43	1.16	10.13	14.73	12.72	1.18	10.41	15.03
DHA 1 year	4.20	0.54	3.14	5.26	4.24	0.54	3.19	5.28
DHA 2 year	3.68	0.77	2.15	5.21	3.94	0.78	2.41	5.47
DHA 3 year	8.66	1.01	6.66	10.66	8.85	1.02	6.85	10.85
	Paik				AIPW			
Placebo 1 year	5.83	0.66	4.54	7.12	5.83	0.66	4.54	7.12
Placebo 2 year	6.65	0.92	4.84	8.47	6.65	0.92	4.84	8.46
Placebo 3 year	12.69	1.22	10.29	15.09	12.69	1.22	10.31	15.08
DHA 1 year	4.23	0.50	3.24	5.21	4.23	0.50	3.24	5.22
DHA 2 year	3.91	0.78	2.38	5.44	3.91	0.78	2.38	5.44
DHA 3 year	8.83	0.95	6.96	10.70	8.78	0.95	6.91	10.65

Table B.3: Sensitivity Analyses for DHA Trial Subgroup, full results for treatment effects

	Est	StdErr	Lower	Upper	Est	StdErr	Lower	Upper	Est	StdErr	Lower	Upper
J2R	MI				Paik				AIPW			
DHA - Placebo 6 months	-1.56	0.84	-3.21	0.10	-1.57	0.81	-3.15	0.02	-1.58	0.81	-3.16	0.00
DHA - Placebo 12 months	-2.29	1.24	-4.73	0.15	-2.28	1.17	-4.58	0.01	-2.30	1.17	-4.59	-0.01
DHA - Placebo 18 months	-3.25	1.64	-6.47	-0.03	-3.49	1.36	-6.16	-0.82	-3.34	1.36	-6.01	-0.67
$\delta = 1$												
DHA - Placebo 6 months	-1.58	0.85	-3.23	0.08	-1.59	0.81	-3.17	-0.01	-1.59	0.81	-3.17	-0.01
DHA - Placebo 12 months	-2.56	1.22	-4.95	-0.17	-2.58	1.22	-4.97	-0.19	-2.58	1.22	-4.97	-0.20
DHA - Placebo 18 months	-3.48	1.58	-6.58	-0.38	-3.66	1.50	-6.60	-0.72	-3.60	1.49	-6.51	-0.68
$\delta = 2$												
DHA - Placebo 6 months	-1.57	0.84	-3.23	0.08	-1.58	0.81	-3.16	-0.01	-1.59	0.80	-3.16	-0.01
DHA - Placebo 12 months	-2.40	1.23	-4.82	0.01	-2.42	1.23	-4.84	-0.00	-2.42	1.23	-4.84	-0.01
DHA - Placebo 18 months	-3.09	1.60	-6.23	0.05	-3.36	1.52	-6.34	-0.39	-3.28	1.51	-6.23	-0.33
$\delta = 3$												
DHA - Placebo 6 months	-1.57	0.84	-3.22	0.08	-1.58	0.80	-3.15	-0.00	-1.58	0.80	-3.15	-0.01
DHA - Placebo 12 months	-2.24	1.25	-4.69	0.20	-2.26	1.25	-4.71	0.19	-2.26	1.25	-4.70	0.19
DHA - Placebo 18 months	-2.70	1.63	-5.90	0.50	-3.06	1.54	-6.08	-0.05	-2.96	1.53	-5.95	0.03

Table B.4: Sensitivity Analyses for DHA Trial Subgroup, full results for $\hat{Y}$

	Est	StdErr	Lower	Upper	Est	StdErr	Lower	Upper	Est	StdErr	Lower	Upper
J2R			MI				Paik				AIPW	
Placebo 1	5.80	0.65	4.54	7.07	5.81	0.66	4.51	7.10	5.82	0.66	4.52	7.11
Placebo 2	6.61	0.95	4.75	8.48	6.60	0.94	4.77	8.44	6.61	0.93	4.78	8.44
Placebo 3	12.65	1.27	10.16	15.14	12.76	1.24	10.33	15.18	12.68	1.23	10.26	15.09
DHA 1	4.25	0.54	3.20	5.29	4.24	0.50	3.25	5.23	4.24	0.50	3.25	5.23
DHA 2	4.32	0.81	2.74	5.91	4.32	0.76	2.82	5.82	4.32	0.76	2.82	5.81
DHA 3	9.40	1.10	7.25	11.55	9.27	0.94	7.43	11.10	9.34	0.93	7.51	11.17
$\delta = 1$												
Placebo 6 months	5.81	0.65	4.55	7.08	5.82	0.66	4.53	7.12	5.82	0.66	4.53	7.12
Placebo 12 months	6.66	0.92	4.84	8.47	6.65	0.92	4.83	8.46	6.65	0.92	4.84	8.46
Placebo 18 months	12.72	1.19	10.38	15.06	12.72	1.22	10.32	15.12	12.69	1.22	10.31	15.07
DHA 6 months	4.24	0.54	3.19	5.29	4.23	0.50	3.24	5.22	4.23	0.50	3.24	5.22
DHA 12 months	4.10	0.79	2.55	5.64	4.07	0.79	2.51	5.62	4.07	0.79	2.51	5.62
DHA 18 months	9.24	1.03	7.21	11.26	9.06	0.97	7.16	10.95	9.09	0.96	7.21	10.98
$\delta = 2$												
Placebo 6 months	5.81	0.65	4.55	7.08	5.82	0.66	4.53	7.11	5.82	0.66	4.53	7.11
Placebo 12 months	6.66	0.94	4.82	8.49	6.64	0.93	4.83	8.46	6.64	0.92	4.83	8.46
Placebo 18 months	12.72	1.21	10.34	15.09	12.72	1.22	10.32	15.12	12.69	1.22	10.30	15.07
DHA 6 months	4.24	0.54	3.19	5.29	4.23	0.50	3.25	5.22	4.23	0.50	3.25	5.22
DHA 12 months	4.25	0.80	2.69	5.82	4.22	0.81	2.64	5.81	4.22	0.81	2.64	5.81
DHA 18 months	9.63	1.05	7.57	11.68	9.35	0.98	7.43	11.28	9.41	0.98	7.50	11.32
$\delta = 3$												
Placebo 6 months	5.81	0.64	4.55	7.07	5.81	0.66	4.52	7.11	5.82	0.66	4.53	7.11
Placebo 12 months	6.65	0.95	4.79	8.51	6.64	0.93	4.82	8.46	6.64	0.93	4.83	8.46
Placebo 18 months	12.71	1.23	10.30	15.13	12.71	1.22	10.31	15.11	12.68	1.22	10.30	15.06
DHA 6 months	4.24	0.53	3.19	5.29	4.24	0.50	3.25	5.22	4.24	0.50	3.25	5.22
DHA 12 months	4.41	0.81	2.82	6.00	4.38	0.83	2.76	6.00	4.38	0.83	2.76	6.00
DHA 18 months	10.01	1.07	7.92	12.10	9.65	1.00	7.69	11.61	9.72	0.99	7.77	11.67

Table B.5: Primary Analyses for Donepezil Trial Subgroup, full results for treatment effects

	Est	StdErr	Lower	Upper	Est	StdErr	Lower	Upper
		MMRM				MI		
Donepezil - Placebo 1 year	-1.10	0.57	-2.22	0.03	-1.11	0.57	-2.23	0.02
Donepezil - Placebo 2 year	-0.81	0.82	-2.41	0.79	-0.85	0.82	-2.46	0.76
Donepezil - Placebo 3 year	-1.39	1.18	-3.70	0.91	-1.58	1.20	-3.93	0.77
		Paik				AIPW		
Donepezil - Placebo 1 year	-1.13	0.49	-2.08	-0.17	-1.13	0.49	-2.09	-0.17
Donepezil - Placebo 2 year	-0.88	0.77	-2.40	0.64	-0.88	0.77	-2.40	0.64
Donepezil - Placebo 3 year	-1.60	1.02	-3.59	0.39	-1.57	1.02	-3.58	0.44

Table B.6: Primary Analyses for Donepezil Trial Subgroup, full results for $\hat{Y}$

	Est	StdErr	Lower	Upper	Est	StdErr	Lower	Upper
	MMRM				MI			
Placebo 1 year	1.25	0.40	0.46	2.04	1.26	0.40	0.48	2.05
Placebo 2 year	2.92	0.58	1.79	4.06	2.97	0.58	1.84	4.10
Placebo 3 year	6.24	0.82	4.62	7.87	6.40	0.84	4.76	8.04
Donepezil 1 year	0.15	0.41	-0.66	0.95	0.15	0.41	-0.64	0.95
Donepezil 2 year	2.11	0.58	0.98	3.25	2.11	0.58	0.97	3.26
Donepezil 3 year	4.85	0.84	3.20	6.50	4.82	0.85	3.15	6.49
	Paik				AIPW			
Placebo 1 year	1.27	0.36	0.56	1.98	1.27	0.36	0.56	1.98
Placebo 2 year	2.99	0.57	1.88	4.11	2.99	0.57	1.87	4.11
Placebo 3 year	6.42	0.89	4.68	8.16	6.43	0.89	4.69	8.18
Donepezil 1 year	0.15	0.37	-0.59	0.88	0.15	0.37	-0.59	0.88
Donepezil 2 year	2.12	0.57	1.00	3.23	2.12	0.57	1.00	3.23
Donepezil 3 year	4.82	0.73	3.40	6.25	4.87	0.73	3.43	6.30

Table B.7: Sensitivity Analyses for Donepezil Trial Subgroup, full results for treatment effects

	Est	StdErr	Lower	Upper	Est	StdErr	Lower	Upper	Est	StdErr	Lower	Upper
J2R	MI				Paik				AIPW			
Donepezil - Placebo 1	-1.11	0.57	-2.23	0.02	-1.13	0.49	-2.09	-0.17	-1.13	0.49	-2.09	-0.17
Donepezil - Placebo 2	-0.88	0.82	-2.49	0.73	-0.91	0.73	-2.35	0.53	-0.91	0.73	-2.35	0.53
Donepezil - Placebo 3	-1.56	1.22	-3.95	0.83	-1.56	0.95	-3.43	0.30	-1.53	0.96	-3.41	0.36
$\delta = 1$												
Donepezil - Placebo 1 year	-1.11	0.57	-2.23	0.02	-1.13	0.49	-2.09	-0.17	-1.13	0.49	-2.09	-0.17
Donepezil - Placebo 2 year	-0.75	0.82	-2.36	0.86	-0.78	0.77	-2.29	0.73	-0.78	0.77	-2.29	0.73
Donepezil - Placebo 3 year	-1.26	1.20	-3.62	1.10	-1.36	1.02	-3.36	0.64	-1.31	1.03	-3.33	0.70
$\delta = 2$												
Donepezil - Placebo 1 year	-1.10	0.57	-2.23	0.02	-1.13	0.49	-2.09	-0.17	-1.13	0.49	-2.08	-0.17
Donepezil - Placebo 2 year	-0.65	0.82	-2.26	0.96	-0.68	0.77	-2.19	0.83	-0.68	0.77	-2.19	0.83
Donepezil - Placebo 3 year	-0.95	1.21	-3.31	1.42	-1.12	1.02	-3.12	0.89	-1.06	1.03	-3.08	0.96
$\delta = 3$												
Donepezil - Placebo 1 year	-1.10	0.57	-2.23	0.02	-1.13	0.49	-2.09	-0.17	-1.13	0.49	-2.09	-0.17
Donepezil - Placebo 2 year	-0.55	0.82	-2.16	1.07	-0.58	0.77	-2.10	0.94	-0.58	0.77	-2.09	0.93
Donepezil - Placebo 3 year	-0.63	1.22	-3.01	1.75	-0.88	1.03	-2.90	1.14	-0.81	1.04	-2.85	1.23

Table B.8: Sensitivity Analyses for Donepezil Trial Subgroup, full results for $\hat{Y}$

	Est	StdErr	Lower	Upper	Est	StdErr	Lower	Upper	Est	StdErr	Lower	Upper
J2R			MI				Paik				AIPW	
Placebo 1	1.26	0.40	0.48	2.05	1.27	0.36	0.56	1.98	1.27	0.36	0.56	1.98
Placebo 2	2.99	0.58	1.85	4.13	3.01	0.57	1.89	4.13	3.01	0.57	1.89	4.13
Placebo 3	6.36	0.86	4.69	8.04	6.39	0.89	4.64	8.14	6.40	0.90	4.64	8.16
Donepezil 1	0.15	0.41	-0.65	0.95	0.14	0.37	-0.59	0.88	0.14	0.37	-0.59	0.88
Donepezil 2	2.11	0.58	0.97	3.25	2.10	0.53	1.05	3.14	2.10	0.53	1.05	3.14
Donepezil 3	4.80	0.88	3.08	6.53	4.83	0.70	3.46	6.20	4.87	0.71	3.49	6.26
$\delta = 1$												
Placebo 1 year	1.26	0.40	0.48	2.05	1.27	0.36	0.56	1.98	1.27	0.36	0.56	1.98
Placebo 2 year	2.97	0.58	1.84	4.10	2.99	0.57	1.87	4.11	2.99	0.57	1.87	4.11
Placebo 3 year	6.40	0.84	4.75	8.04	6.42	0.89	4.68	8.16	6.43	0.89	4.69	8.18
Donepezil 1 year	0.16	0.41	-0.64	0.96	0.15	0.37	-0.59	0.88	0.15	0.37	-0.59	0.88
Donepezil 2 year	2.21	0.58	1.07	3.36	2.21	0.57	1.11	3.32	2.21	0.57	1.11	3.32
Donepezil 3 year	5.13	0.86	3.46	6.81	5.06	0.73	3.62	6.50	5.12	0.74	3.67	6.57
$\delta = 2$												
Placebo 1 year	1.26	0.40	0.48	2.04	1.27	0.36	0.56	1.98	1.27	0.36	0.56	1.98
Placebo 2 year	2.97	0.58	1.84	4.10	2.99	0.57	1.87	4.11	2.99	0.57	1.87	4.11
Placebo 3 year	6.39	0.84	4.74	8.04	6.42	0.89	4.67	8.16	6.43	0.89	4.69	8.18
Donepezil 1 year	0.16	0.41	-0.64	0.96	0.14	0.37	-0.59	0.88	0.14	0.37	-0.59	0.88
Donepezil 2 year	2.32	0.58	1.17	3.46	2.31	0.57	1.21	3.42	2.31	0.57	1.20	3.42
Donepezil 3 year	5.45	0.86	3.76	7.13	5.30	0.74	3.85	6.76	5.37	0.75	3.90	6.84
$\delta = 3$												
Placebo 1 year	1.26	0.40	0.47	2.04	1.27	0.36	0.56	1.98	1.27	0.36	0.56	1.98
Placebo 2 year	2.96	0.58	1.83	4.10	2.99	0.57	1.87	4.11	2.99	0.57	1.87	4.11
Placebo 3 year	6.39	0.85	4.73	8.05	6.42	0.89	4.68	8.16	6.43	0.89	4.68	8.18
Donepezil 1 year	0.16	0.41	-0.64	0.96	0.14	0.38	-0.59	0.88	0.14	0.37	-0.59	0.88
Donepezil 2 year	2.42	0.59	1.27	3.57	2.41	0.57	1.30	3.52	2.41	0.57	1.30	3.52
Donepezil 3 year	5.76	0.87	4.07	7.46	5.54	0.76	4.06	7.02	5.62	0.76	4.13	7.12

Bibliography

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

[2] Heejung Bang and James M. Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, dec 2005.

[3] Josephine Barnes, Bradford C Dickerson, Chris Frost, Lize C Jiskoot, David Wolk, and Wiesje M van der Flier. Alzheimer's disease first symptoms are age dependent: Evidence from the NACC dataset. *Alzheimers & Dementia*, 11(11):1349–1357, nov 2015.

[4] Thomas G Beach, Sarah E Monsell, Leslie E Phillips, and Walter Kukull. Accuracy of the Clinical Diagnosis of Alzheimer Disease at National Institute on Aging Alzheimer Disease Centers, 2005-2010. *Journal of Neuropathology and Experimental Neurology*, 71(4):266–273, apr 2012.

[5] H Braak and E Braak. Neuropathological Staging of Alzheimer-Related Changes. *Acta Neuropathologica*, 82(4):239–259, 1991.

[6] Weihua Cao, Anastasios A. Tsiatis, and Marie Davidian. Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika*, 96(3):723–734, sep 2009.

[7] James R. Carpenter and Michael G. Kenward. *Multiple Imputation and its Application*. John Wiley & Sons, Inc., 2013.

[8] James R. Carpenter, James H. Roger, and Michael G. Kenward. Analysis of Longitudinal Trials with Protocol Deviation: a Framework for Relevant, Accessible Assumptions, and Inference Via Multiple Imputation. *Journal of Biopharmaceutical Statistics*, 23(6):1352–1371, NOV 2 2013.

[9] Michelle Casey, Evgeny Degtyarev, Maria Jose Lechuga, Paola Aimone, Alain Ravaud, Robert J Motzer, Feng Liu, Viktoriya Stalbovskaya, Rui Tang, Emily Butler, Oliver Sailer, Susan Halabi, and Daniel George. Estimand framework: Are we asking the right questions? A case study in the solid tumor setting. *Pharmaceutical Statistics*, 20(2):324–334, mar 2021.

[10] Paul K Crane, Emily Trittschuh, Shubhabrata Mukherjee, Andrew J Saykin, R Elizabeth Sanders, Eric B Larson, Susan M McCurry, Wayne McCormick, James D Bowen, Thomas Grabowski, Mackenzie Moore, Julianna Bauman, Alden L Gross, C Dirk Keene, Thomas D Bird, Laura E Gibbons, Jesse Mez, and Executive Prominent Alzheimer's. Incidence of cognitively defined late-onset Alzheimer's dementia subgroups from a prospective cohort study. *Alzheimers & Dementia*, 13(12):1307–1316, dec 2017.

[11] Suzie Cro, Tim P Morris, Michael G Kenward, and James R Carpenter. Sensitivity analysis for clinical trials with missing continuous outcome data using controlled multiple imputation: A practical guide. *Statistics in Medicine*, 39(21):2815–2842, sep 2020.

[12] Julie E Davidson, Michael C Irizarry, Bethany C Bray, Sally Wetten, Nicholas Galwey, Rachel Gibson, Michael Borrie, Richard Delisle, Howard H Feldman, Ging-Yuek Hsiung, Luis Fornazzari, Serge Gauthier, Danilo Guzman, Inge Loy-English, Ron Keren, Andrew Kertesz, Peter St George-Hyslop, John Wherrett, and Andreas U Monsch. An exploration of cognitive subgroups in Alzheimer's disease. *Journal of the International Neuropsychological Society*, 16(2):233–243, mar 2010.

[13] Bradford C Dickerson, David A Wolk, and Alzheimer's Dis Neuroimaging Initi. Dysexecutive versus amnesic phenotypes of very mild Alzheimer's disease are associated with distinct clinical, genetic and cortical thinning characteristics. *Journal of Neurology Neurosurgery And Psychiatry*, 82(1):45–51, jan 2011.

[14] B Dubois, H H Feldman, and C Jacova. Advancing research diagnostic criteria for Alzheimer's disease: the IWG-2 criteria (vol 13, pg 614, 2014). *Lancet Neurology*, 13(8):757, aug 2014.

[15] Bruno Dubois, Howard H. Feldman, Claudia Jacova, Jeffrey L. Cummings, Steven T. DeKosky, Pascale Barberger-Gateau, Andre Delacourte, Giovanni Frisoni, Nick C. Fox, Douglas Galasko, Serge Gauthier, Harald Hampel, Gregory A. Jicha, Kenichi Meguro, John O'Brien, Florence Pasquier, Philippe Robert, Martin Rossor, Steven Salloway, Marie Sarazin, Leonardo C. de Souza, Yaakov Stern, Pieter J. Visser, and Philip Scheltens. Revising the definition of Alzheimer's disease: a new lexicon. *Lancet Neurology*, 9(11):1118–1127, NOV 2010.

[16] Bruno Dubois, Howard H Feldman, Claudia Jacova, Steven T Dekosky, Pascale Barberger-Gateau, Jeffrey Cummings, Andre Delocourte, Douglas Galasko, Serge Gauthier, Gregory Jicha, Kenichi Meguro, John O'Brien, Florence Pasquier, Philippe Robert, Martin Rossor, Steven Solloway, Yaakov Stern, Pieter J Visser, and Philip Scheltens. Research criteria for the diagnosis of Alzheimer's disease: revising the NINCDS-ADRDA criteria. *Lancet Neurology*, 6(8):734–746, aug 2007.

[17] European Medicines Agency. *Guideline on missing data in confirmatory clinical trials*. https://database.ich.org/sites/default/files/E9-R1_Step4_Guideline_2019_1203.pdf, July 2010.

[18] Mallorie H Fiero, Madeline Pe, Chana Weinstock, Bellinda L King-Kallimanis, Scott Komo, Heidi D Klepin, Stacy W Gray, Andrew Bottomley, Paul G Kluetz, and Rajeshwari Sridhara. Demystifying the estimand framework: a case study using patient-reported outcomes in oncology. *Lancet Oncology*, 21(10):E488–E494, oct 2020.

[19] Lysbeth Floden, Stacie Hudgens, Hailin Yu, and Melanie Bell. Imputation strategies within the estimand framework to evaluate the overall likelihood of patient improvement in longitudinal trails. *Quality of Life Research*, 29(SUPPL 1, 1, SI):S26, oct 2020.

[20] C Fraley and A E Raftery. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97(458):611–631, jun 2002.

[21] C Fraley, A E Raftery, T B Murphy, and L Scrucca. *mclust Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation. Technical Report*. No. 597, Department of Statistics, University of Washington, 2012.

[22] Alberto Garcia-Hernandez, Teresa Perez, Maria del Carmen Pardo, and Dimitris Rizopoulos. MMRMvs joint modeling of longitudinal responses and time to study drug discontinuation in clinical trials using a "de jure" estimand. *Pharmaceutical Statistics*, 19(6):909–927, nov 2020.

[23] Tilmann Gneiting and Adrian E. Raftery. Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

[24] D S Knopman, S T DeKosky, J L Cummings, H Chui, J Corey-Bloom, N Relkin, G W Small, B Miller, and J C Stevens. Practice parameter: Diagnosis of dementia (an evidence-based review) - Report of the Quality Standards Subcommittee of the American Academy of Neurology. *Neurology*, 56(9):1143–1153, may 2001.

[25] Kung-Yee Liang and Scott L Zeger. Longitudinal Data Analysis Using Generalized Linear Models. *Biometrika*, 73(1):13–22, 1986.

[26] R Little and L Yau. Intent-to-treat analysis for longitudinal studies with drop-outs. *Biometrics*, 52(4):1324–1333, dec 1996.

[27] R J A Little. Pattern-Mixture Models for Multivariate Incomplete Data. *Journal of the American Statistical Association*, 88(421):125–134, mar 1993.

[28] RJA Little, Garrett Fitzmaurice, Davidian Marie, Geert Verbeke, and Geert Molenberghs. *Longitudinal Data Analysis: Selection and pattern-mixture models*. Chapman and Hall/CRC, 2008.

[29] Roderick J A Little. Regression With Missing X's: A Review. *Journal of the American Statistical Association*, 87(420):1227–1237, mar 1992.

[30] Roderick J A Little. Modeling the Drop-Out Mechanism in Repeated-Measures Studies. *Journal of the American Statistical Association*, 90(431):1112–1121, mar 1995.

[31] Roderick J. A. Little and Donald B. Rubin. *Statistical Analysis with Missing Data*. John Wiley & Sons, Inc., 2 edition, 2002.

[32] Guy McKhann, David Drachman, Marshall Folstein, Robert Katzman, Donald Price, and Emanuel M. Stadlan. Clinical diagnosis of alzheimer's disease. *Neurology*, 34(7):939–939, 1984.

[33] Guy M. McKhann, David S. Knopman, Howard Chertkow, Bradley T. Hyman, Clifford R. Jack, Jr., Claudia H. Kawas, William E. Klunk, Walter J. Koroshetz, Jennifer J. Manly, Richard Mayeux, Richard C. Mohs, John C. Morris, Martin N. Rossor, Philip Scheltens, Maria C. Carrillo, Bill Thies, Sandra Weintraub, and Creighton H. Phelps. 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. *Alzheimers & Dementia*, 7(3):263–269, MAY 2011.

[34] Guy M McKhann, David S Knopman, Howard Chertkow, Bradley T Hyman, Clifford R Jack Jr., Claudia H Kawas, William E Klunk, Walter J Koroshetz, Jennifer J Manly, Richard Mayeux, Richard C Mohs, John C Morris, Martin N Rossor, Philip Scheltens, Maria C Carrillo, Bill Thies, Sandra Weintraub, and Creighton H Phelps. 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. *Alzheimers & Dementia*, 7(3):263–269, may 2011.

[35] X L Meng. Multiple-Imputation Inferences with Uncongenial Sources of Input. *Statistical Science*, 9(4):538–558, nov 1994.

[36] Jesse Mez, Stephanie Cosentino, Adam M Brickman, Edward D Huey, Jennifer J Manly, and Richard Mayeux. Dysexecutive Versus Amnestic Alzheimer Disease Subgroups: Analysis of Demographic, Genetic, and Vascular Factors. *Alzheimer Disease & Associated Disorders*, 27(3):218–225, 2013.

[37] Hege Michiels, Cristina Sotto, An Vandebosch, and Stijn Vansteelandt. A novel estimand to adjust for rescue treatment in randomized clinical trials. *Statistics in Medicine*.

[38] S S Mirra, A Heyman, D McKeel, S M Sumi, B J Crain, L M Brownlee, F S Vogel, J P Hughes, G van Belle, and L and Berg. The Consortium to Establish a Registry for Alzheimer{\textquoteright}s Disease (CERAD). *Neurology*, 41(4):479, 1991.

[39] Geert Molenberghs and Michael G. Kenward. *Missing Data in Clinical Studies*. John Wiley & Sons, Inc., 2007.

[40] Sarah E Monsell, Hiroko H Dodge, Xiao-Hua Zhou, Yunqi Bu, Lilah M Besser, Charles Mock, Stephen E Hawes, Walter A Kukull, Sandra Weintraub, and Neuropsychology Work Grp Advisory. Results From the NACC Uniform Data Set Neuropsychological Battery Crosswalk Study. *Alzheimer Disease & Associated Disorders*, 30(2):134–139, 2016.

[41] J C Morris. The Clinical Dementia Rating (CDR) - Current Version And Scoring Rules. *Neurology*, 43(11):2412–2414, nov 1993.

[42] John C Morris, Sandra Weintraub, Helena C Chui, Jeffrey Cummings, Charles DeCarli, Steven Ferris, Norman L Foster, Douglas Galasko, Neill Graff-Radford, Elaine R Peskind, Duane Beekly, Erin M Ramos, and Walter A Kukull. The uniform data set (UDS): Clinical and cognitive variables and descriptive data from Alzheimer disease centers. *Alzheimer Disease & Associated Disorders*, 20(4):210–216, 2006.

[43] National Research Council (US) Panel on Handling Missing Data in Clinical Trials. *The Prevention and Treatment of Missing Data in Clinical Trials*. Washington (DC): National Academies Press (US), 2010.

[44] Myunghee Cho Paik. The Generalized Estimating Equation Approach When Data are Not Missing Completely at Random. *Journal of the American Statistical Association*, 92(440):1320–1329, dec 1997.

[45] Jessica Peter, Ahmed Abdulkadir, Christoph Kaller, Dorothee Kuemmerer, Michael Huell, Werner Vach, Stefan Kloeppel, and Alzheimer's Dis Neuroimaging Initi. Subgroups of Alzheimer's Disease: Stability of Empirical Clusters Over Time. *Journal of Alzheimers Disease*, 42(2):651–661, 2014.

[46] Ronald C. Petersen, Ronald G. Thomas, Michael Grundman, David Bennett, Rachelle Doody, Steven Ferris, Douglas Galasko, Shelia Jin, Jeffrey Kaye, Allan Levey, Eric Pfeiffer, Mary Sano, Christopher H. van Dyck, and Leon J. Thal. Vitamin E and Donepezil for the Treatment of Mild Cognitive Impairment. *New England Journal of Medicine*, 352(23):2379–2388, jun 2005.

[47] R I Pfeffer, T T Kurosaki, C H Harrah, J M Chance, and S Filos. Measurement of Functional Activities in Older Adults in the Community. *Journals of Gerontology*, 37(3):323–329, 1982.

[48] 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, Jr., Michael Weiner, Lynne Shinto, and Paul S. Aisen. Docosahexaenoic Acid Supplementation and Cognitive Decline in Alzheimer Disease A Randomized Trial. *Jama-Journal of the American Medical Association*, 304(17):1903–1911, NOV 3 2010.

[49] Bohdana Ratitch, Michael O'Kelly, and Robert Tosiello. Missing data in clinical trials: from clinical assumptions to statistical analysis using patternmixture models. *Pharmaceutical Statistics*, 12(6):337–347, NOV 2013.

[50] James M Robins, Andrea Rotnitzky, Lue Ping Zhao, and Lue Ping ZHAO. Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data. *Journal of the American Statistical Association*, 90(429):106–121, 1995.

[51] James M. Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of Regression Coefficients When Some Regressors Are Not Always Observed. *Journal of the American Statistical Association*, 89(427):846, sep 1994.

[52] Andrea Rotnitzky, Quanhong Lei, Mariela Sued, and James M. Robins. Improved double-robust estimation in missing data and causal inference models. *Biometrika*, 99(2):439–456, JUN 2012.

[53] D B Rubin. Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91(434):473–489, jun 1996.

[54] Donald B Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, dec 1976.

[55] Donald B Rubin. *Multiple imputation for nonresponse in surveys*. John Wiley & Sons, Inc., 2008.

[56] E Scarpini, P Scheltens, and H Feldman. Treatment of Alzheimer's disease: current status and new perspectives. *Lancet Neurology*, 2(9):539–547, SEP 2003.

[57] J. L. Schafer. *Analysis of Incomplete Multivariate Data*. Chapman and Hall/CRC, 1997.

[58] Daniel O. Scharfstein, Andrea Rotnitzky, and James M. Robins. Adjusting for Nonignorable Drop-Out Using Semiparametric Nonresponse Models. *Journal of the American Statistical Association*, 94(448):1096–1120, dec 1999.

[59] Nienke M E Scheltens, Francisca Galindo-Garre, Yolande A L Pijnenburg, Annelies E van der Vlies, Lieke L Smits, Teddy Koene, Charlotte E Teunissen, Frederik Barkhof, Mike P Wattjes, Philip Scheltens, and Wiesje M van der Flier. The identification of cognitive subtypes in Alzheimer's disease dementia using latent class analysis. *Journal of Neurology Neurosurgery and Psychiatry*, 87(3):235–243, mar 2016.

[60] Nienke M E Scheltens, Betty M Tijms, Teddy Koene, Frederik Barkhof, Charlotte E Teunissen, Steffen Wolfsgruber, Michael Wagner, Johannes Kornhuber, Oliver Peters, Brendan I Cohn-Sheehy, Gil D Rabinovici, Bruce L Miller, Joel H Kramer, Philip Scheltens, Wiesje M van der Flier, and Alzheimer's Dis Neuroimaging. Cognitive subtypes of probable Alzheimer's disease robustly identified in four cohorts. *Alzheimers & Dementia*, 13(11):1226–1236, nov 2017.

[61] Philip Scheltens, Kaj Blennow, Monique M. B. Breteler, Bart de Strooper, Giovanni B. Frisoni, Stephen Salloway, and Wiesje Maria Van der Flier. Alzheimer's disease. *Lancet*, 388(10043):505–517, JUL 30 2016.

[62] Gideon Schwarz. Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2):461–464, mar 1978.

[63] Shaun Seaman and Andrew Copas. Doubly robust generalized estimating equations for longitudinal data. *Statistics in Medicine*, 28(6):937–955, mar 2009.

[64] Shaun R. Seaman and Stijn Vansteelandt. Introduction to Double Robust Methods for Incomplete Data. *Statistical Science*, 33(2):184–197, 2018.

[65] Cheryl L Stopford, Julie S Snowden, Jennifer C Thompson, and David Neary. Variability in cognitive presentation of Alzheimer's disease. *Cortex*, 44(2):185–195, feb 2008.

[66] The International Council on Harmonization (ICH). *Addendum on Estimands and Sensitivity Analysis in Clinical Trials to the Guideline on Statistical Principles for Clinical Trials E9(R1) Final version (Step 4)*. https://database.ich.org/sites/default/files/E9-R1_Step4_Guideline_2019_1203.pdf, 2019.

[67] H Thijs, G Molenberghs, B Michiels, G Verbeke, and D Curran. Strategies to fit pattern-mixture models. *Biostatistics*, 3(2):245–265, jun 2002.

[68] Anastasios A. Tsiatis. *Semiparametric Theory and Missing Data*. Springer New York, 2006.

[69] Anastasios A. Tsiatis, Marie Davidian, and Weihua Cao. Improved Doubly Robust Estimation When Data Are Monotonely Coarsened, with Application to Longitudinal Studies with Dropout. *Biometrics*, 67(2):536–545, jun 2011.

[70] S. Van Buuren, J. P. L. Brand, C. G. M. Groothuis-Oudshoorn, and D. B. Rubin. Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*, 76(12):1049–1064, DEC 2006.

[71] Stef van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, 45(3):1–67, DEC 2011.

[72] Stef van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, 45(3):1–67, DEC 2011.

[73] Emma R L C Vardy, Andrew H Ford, Peter Gallagher, Rosie Watson, Ian G McKeith, Andrew Blamire, and John T O'Brien. Distinct cognitive phenotypes in Alzheimer's disease in older people. *International Psychogeriatrics*, 25(10):1659–1666, oct 2013.

[74] Sandra Weintraub, Lilah Besser, Hiroko H Dodge, Merilee Teylan, Steven Ferris, Felicia C Goldstein, Bruno Giordani, Joel Kramer, David Loewenstein, Dan Marson, Dan Mungas, David Salmon, Kathleen Welsh-Bohmer, Xiao-Hua Zhou, Steven D Shirk, Alireza Atri, Walter A Kukull, Creighton Phelps, and John C Morris. Version 3 of the Alzheimer Disease Centers' Neuropsychological Test Battery in the Uniform Data Set (UDS). *Alzheimer Disease & Associated Disorders*, 32(1):10–17, 2018.

[75] Sandra Weintraub, David Salmon, Nathaniel Mercaldo, Steven Ferris, Neill R Graff-Radford, Helena Chui, Jeffrey Cummings, Charles DeCarli, Norman L Foster, Douglas Galasko, Elaine Peskind, Woodrow Dietrich, Duane L Beekly, Walter A Kukull, and John C Morris. The Alzheimer's Disease Centers' Uniform Data Set (UDS) The Neuropsychologic Test Battery. *Alzheimer Disease & Associated Disorders*, 23(2):91–101, 2009.

[76] J A Yesavage, T L Brink, T L Rose, O Lum, V Huang, M Adey, and V O Leirer. Development and Validation of a Geriatric Depression Screening Scale - a Preliminary-Report. *Journal of Psychiatric Research*, 17(1):37–49, 1983.