Title

Link, transport, integrate: a Bayesian latent class mixture modeling framework for scalable algorithmic dementia classification in population-representative studies

Author

Shaw, Crystal Ruth Michelle

Publication Date

2023

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Link, transport, integrate: a Bayesian latent class mixture modeling framework for scalable algorithmic dementia classification in population-representative studies

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

by

Crystal Ruth Michelle Shaw

2023

ABSTRACT OF THE DISSERTATION

Link, transport, integrate: a Bayesian latent class mixture modeling framework for scalable algorithmic dementia classification in population-representative studies

by

Crystal Ruth Michelle Shaw

Doctor of Philosophy in Biostatistics

University of California, Los Angeles, 2023

Professor Thomas R. Belin, Co-Chair

Professor Elizabeth Rose Mayeda, Co-Chair

Gold-standard clinical dementia adjudication is resource intensive and infeasible in large, population-representative studies which are critical for public health research. Algorithmic dementia classification uses models to predict cognitive impairment and was developed to circumvent challenges of the gold-standard adjudication process. Several algorithms have been developed to classify dementia in the Health and Retirement Study (HRS) and rely on information in the Aging, Demographics, and Memory Study (ADAMS), a substudy of HRS initiated in 2001. Existing algorithms cannot incorporate neuropsychological measures as they are unavailable in HRS, and models cannot be adapted to include more comprehensive measures available in newer studies.

I propose a novel Bayesian latent class mixture modeling framework for algorithmic dementia classification that incorporates information from neuropsychological measures and can be adapted to include more comprehensive measures available in updated studies. The model uses latent class mixture models to create synthetic versions of datasets, incorporating information on relationships between sociodemographic, health, and cognitive measures and cognitive impairment classes through prior distributions based on studies with gold-standard adjudicated cases. This work involves three studies on aging: The Health and Retirement Study (HRS), The Harmonized Cognitive Assessment Protocol (HCAP, HRS substudy), and the Aging and Demographics Study (ADAMS, HRS substudy). Simulation studies were conducted to evaluate the role of study sample size and priors specified based on different data sources and sampling frames and their impact on algorithmic dementia classification results and inferences on racial/ethnic differences in dementia.

Analyses using priors from ADAMS accurately captured cognitive impairment classes preserved racial/ethnic differences in dementia for Black vs. White participants. Priors better calibrated to the analytic sample however improved estimates for Black and Hispanic participants and preserved racial/ethnic differences in dementia for Black vs. White and Hispanic vs. White participants. Applying the model to HCAP 2016 yielded reasonable estimates of cognitive impairment classes with proportions of impaired participants in line with findings published by HCAP investigators.

This dissertation lays important groundwork for strengthening algorithmic dementia classification in population-representative studies. Outcomes from this work are directly applicable to existing studies on AD/ADRD that are harmonizable with HRS/HCAP.

The dissertation of Crystal Ruth Michelle Shaw is approved.

Donatello Telesca

Ronald S. Brookmeyer

Jennifer Manly

Elizabeth Rose Mayeda, Committee Co-Chair

Thomas R. Belin, Committee Co-Chair

University of California, Los Angeles

2023

To my dad, Michael,

who believed I could do anything I set my mind to. I miss him every day.

To my son, JD,

who is sunshine personified. You make every day better and everything worth it.

TABLE OF CONTENTS

List of Abbreviations

Abbreviation	Definition
ADAMS	Aging, Demographics, and Memory Study
AD	Alzheimer's Disease
ADL	Activities of Daily Living
ADRD	Alzheimer's Disease-Related Dementias
BMI	Body Mass Index
FCS	Fully-conditional Specification
FDA	Food and Drug Administration
HCAP	Harmonized Cognitive Assessment Protocol
HRS	Health and Retirement Study
IADL	Instrumental Activities of Daily Living
LASSO	Least Absolute Shrinkage and Selection Operator
MCI	Mild Cognitive Impairment
MCMC	Markov Chain Monte Carlo
MI	Multiple Imputation
MMSE	Mini Mental State Exam
PMM	Predictive Mean Matching
RMSE	Root Mean Square Error
SRS	Simple Random Sample
US	United States

List of Figures

List of Tables

Acknowledgements

My time at UCLA has been rich, productive, and full of personal and professional growth. I could not have asked for a better PhD experience. I am so grateful for the opportunities I've had and the people I met along the way. I could not have done this without the extraordinary support of my colleagues, friends, and family.

The environment at UCLA biostatistics is warm and welcoming. I am thankful for a department with brilliant faculty who are genuinely invested in students, supportive staff who work tirelessly to make the department function, and bright students who value collaboration and helping each other succeed. Long hours and late nights are easier when you're laughing with your study group, and there are few bonds stronger than those forged in the fire of comps, quals, and a dissertation.

I'm extremely lucky to have been co-mentored by Dr. Tom Belin (UCLA Biostat) and Dr. Elizabeth Rose Mayeda (UCLA Epi). I could not have asked for co-mentors more dedicated to my personal and professional development or more suited to my personality. I've accomplished more and come farther than I could've imagined at the start of my program, and it's because of their excellence, dedication, and kindness as mentors. I am a better researcher and biostatistician because of them. I am also grateful to the rest of my dissertation committee "dream team," Dr. Donatello Telesca (UCLA Biostat), Dr. Ron Brookmeyer (UCLA Biostat), and Dr. Jen Manly (Columbia Neurology), all of whom are dedicated to their students and leaders in their respective disciplines. I got to learn from the best and the brightest.

The Mayeda Research Group was my research home at UCLA. I'm so thankful for this outstanding group of people and the tight community that we built together. The MRG was a special place to work; the culture of warmth, excellence, and support in this group is unmatched.

I was introduced to the MEthods in LOngitudinal DEMentia (MELODEM) group through my work with the MRG. I'm grateful to the MELODEM community for creating such a positive environment full of smart, critical thinking researchers on the cutting edge of aging and dementia research. I owe them a debt of gratitude for the initial spark of inspiration for this dissertation and for helping refine the ideas and direction of this project through consistently kind and constructive feedback on my MELODEM presentations.

This project has several moving parts, and it was easy at times to "miss the forest for the trees." This work is better for the thoughtful feedback of several excellent collaborators. I'm especially thankful to Eleanor Hayes-Larson who was the best simulation brainstorming buddy and co-author and has helped me get through several writing slumps, kept me focused on big-picture goals, provided feedback on early drafts of text and figures, and always had an encouraging word for me throughout the PhD process. There were several other MRG members who contributed to important details in this dissertation as well: Yingyan Wu contributed to dataset construction code, Joey Fong provided template code for automated tables, and Juliet Zhou code reviewed portions of the analyses. All MRG members contributed to my sanity and happiness during this process. This would've taken way longer and been less fun without you all.

I'm grateful to the National Institute on Aging for supporting my dissertation work through an individual predoctoral fellowship (F31 AG071191). I am also thankful to the HRS/HCAP/ADAMS co-investigators, participants, and their families, for contributing to important research on aging and dementia.

Lastly and most importantly, I am grateful to my mom and my son for their unconditional love and support. Graduate school would have been impossible without my supermom, who filled in every gap at home when I had early meetings or late-night work sessions and fed me

delicious homemade Thai food to “give me energy to study.” She is unequivocally dedicated to my success and has made countless sacrifices to ensure that me and my son have a bright future. My son doesn’t remember a time when I wasn’t in grad school, so I’m really excited to say that my “gigantic writing project” is done, buddy, and yes, I would love to spend all weekend cooking together and playing video games.

Vita

Education

MS Applied Mathematics; California State University, Northridge	2017
BA Mathematics; California State University, Northridge	2011

Employment

External Collaborator; Pfizer, Inc.	Sept 2018 – Present
Graduate Student Researcher; UCLA FSPH Department of Epidemiology	Apr 2018 – Present
Statistical Programmer; UCLA Health	Mar 2018 – Sept 2022
Computational Biology Summer Student; Pfizer, Inc.	Jun 2018 – Sept 2018
Teaching Assistant; UCLA FSPH Department of Biostatistics	Sept 2017 – Mar 2018
Research Assistant; California State University, Northridge Department of Mathematics	Sept 2016 – Jun 2017
Teaching Assistant; California State University, Northridge Department of Mathematics	Aug 2014 – Dec 2016
Statistician Associate; Social Bluebook, LLC	May 2015 – May 2016
Math Teacher and Department Chair; CHAMPS Charter High School	Aug 2012 – May 2014

Publications

Mayeda ER and **Shaw C**. Invited commentary: Algorithmic dementia classification: promises and challenges. 2023; *In Press.*

Shaw C, Wu Y, Zimmerman SC, Hayes-Larson E, Belin TR, Power MC, Glymour MM, Mayeda ER. Comparison of imputation strategies for incomplete longitudinal late in lifecourse epidemiology. *American Journal of Epidemiology*. 2023; *In Press.*

Mobley TM, **Shaw C**, Hayes-Larson E, Fong J, Gilsanz P, Gee GC, Brookmeyer R, Whitmer RA, Casey JA, Mayeda ER. Neighborhood disadvantage and dementia incidence in a cohort of Asian American and non-Latino White older adults in Northern California. *Alzheimer's & Dementia*. 2022; *Online ahead of print. https://doi.org/10.1002/alz.12660*.

Mungas D, **Shaw C**, Hayes-Larson E, DeCarli C, Tomaszewski Farias S, Olichney J, Hernandez Saucedo H, Gilsanz P, Glymour MM, Whitmer R, Mayeda ER. Cognitive impairment in racially/ethnically diverse older adults: Accounting for sources of diagnostic bias. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*. 2021; 13(1):e12265. doi:10.1002/dad2.12265

Hayes-Larson E, **Shaw C**, Ackley SF, Zimmerman SC, Glymour MM, Graff RE, Witte JS, Kobayashi LC, Mayeda ER. The role of dementia diagnostic delay in understanding the inverse cancer-dementia link. J Gerontol A Biol Sci Med Sci. 2022; 77(6): 1254-1260. doi: 10.1093/gerona/glab341. PMID: 34788817; PMCID: PMC9159667

Shaw C, Mayeda ER, Telesca D, Belin TR. 2021. A Bayesian latent class mixture modeling framework for algorithmic dementia classification in population-representative studies. In JSM Proceedings, Statistics in Epidemiology Section. Virtual Conference: American Statistical Associate. 1302-1319.

Shaw C, Hayes-Larson E, Glymour MM, Dufouil C, Hohman TJ, Whitmer RA, Kobayashi LC, Brookmeyer R, Mayeda ER. Evaluation of selective survival and sex/gender differences in dementia incidence using a simulation model. *JAMA Network* Open. 2021; 4(3):e211001. doi:10.1001/jamanetworkopen.2021.1001

Hayes-Larson E, Mobley TM, Gilmore-Bykovskyi A, **Shaw C**, Karlamanga A, Manly JJ, Mayeda ER. Racial/ethnic differences in health-related quality of life in persons with and without dementia. *Journal of the American Geriatrics Society.* 2020; 69(3):629-636. doi: 10.1111/jgs.16908

Corlier FW, **Shaw C**, Hayes-Larson E, Mungas D, Tomaszewski Farias S, Glymour MM, Whitmer RA, Mayeda ER. Association Between Cognitive Test Performance and Subjective Cognitive Decline in a Diverse Cohort of Older Adults. *Alzheimer Disease & Associated Disorders*. 2020; 34(3):198-205. doi: 10.1097/WAD.0000000000000383

Mojirsheibani M and **Shaw C**. Classification with incomplete functional covariates. *Statistics & Probability Letters*. 2018; 139:40-46.

Selected Presentations

Shaw C, Mayeda ER, Telesca D, Belin TR. (2022, June). Link, transport, integrate: A Bayesian latent class mixture modeling framework for scalable algorithmic dementia classification in population-representative studies. Paper presented at the MEthods for LOngitudinal Studies in DEMentia progress meeting, Bordeaux, France.

Shaw C, Wu Y, Zimmerman SC, Hayes-Larson E, Belin TR, Power MC, Glymour MM, Mayeda ER. (2022, June). Comparison of imputation strategies for incomplete longitudinal exposure data in lifecourse epidemiology. Poster presented at the Society for Epidemiologic Research Conference, Chicago, IL, USA.

Chapter 1 Introduction

Over 6.5 million individuals —about 1 in 9— in the United States (US) aged 65 years or older are living with Alzheimer's disease (AD) and Alzheimer's disease-related dementias (ADRD), and this number is projected exceed 12.5 million by the year 2050 (Rajan et al. 2021). Nationally-representative cohorts are crucial for monitoring population trends in AD/ADRD incidence, prevalence, and disparities, as well as for understanding determinants of AD/ADRD (Mayeda et al. 2016; Chin, Negash, and Hamilton 2011; Ferretti et al. 2020). Dementia, which is characterized by cognitive impairment severe enough to impact functional ability, is difficult to diagnose due to elusive onset and heterogenous clinical presentation (Alzheimer's Association 2022). Thus, gold-standard dementia adjudication requires consensus diagnosis by an expert panel that triangulates evidence from detailed neuropsychological testing, clinical examination, an informant interview, and medical history (Mayeux et al. 2011; McKhann et al. 2011; Langa et al. 2005). This time- and resource-intensive protocol is infeasible in large population-based surveys, which are of considerable public health interest to develop dementia prevention strategies and treatment and reduce disparities in incidence across subgroups. Algorithmic dementia classification methods have been developed to predict individuals' probability of dementia in large studies (Kasper, Freedman, and Spillman 2013; Gross et al. 2017; Prina et al. 2019), but the lack of available data on neuropsychological testing in population-based surveys has been a persistent limitation of existing dementia classification strategies (Gianattasio et al. 2019).

The Health and Retirement Study (HRS) is a large, ongoing (1992-present) US population-representative study in which a variety of algorithmic dementia classification

methods have been used to predict participants' probability of dementia (Crimmins et al. 2011; Hurd et al. 2013; Langa, Kabeto, and Weir 2009; Wu et al. 2013; Herzog and Wallace 1997). The primary purpose of HRS was to provide a national resource for data on the changing health and economic circumstances associated with aging and thus, the HRS survey gathers extensive information on sociodemographic characteristics, lifestyle and health variables, and general cognitive assessments in an aging population (Sonnega et al. 2014). Though HRS does not have detailed neuropsychological assessments available for all participants, substudies were initiated to collect these important measures. Specifically, the Aging, Demographics, and Memory Study (ADAMS [2001-2009]) (Langa et al. 2005) and the Harmonized Cognitive Assessment Protocol (HCAP [2016-present]) (Weir, Langa, and Ryan 2016) collected neuropsychological test data for a subset of HRS participants. The ADAMS study performed gold-standard dementia adjudication for all participants and was the first study of its kind to do so in a national cohort (Langa et al. 2005; Heeringa et al. 2009). Though HCAP was designed to be a larger, contemporary follow up to the ADAMS study (HCAP n=3496; ADAMS n=856), the scale of the HCAP study precludes gold-standard dementia adjudication for all HCAP study participants (Langa et al. 2020). Thus, ADAMS remains an engine for algorithm development to classify dementia in the larger HRS sample.

An important limitation identified in several algorithmic dementia classification methods is differential misclassification by race/ethnicity. Recognizing that this impacts validity of algorithmic dementia classification methods for use in racial/ethnic disparities research, the most recent dementia classification algorithms developed in ADAMS utilized race-specific score cutoffs— a somewhat controversial method (Gasquoine 2009; Manly 2005; Manly and Echemendia 2007). The algorithms are more sensitive than other existing algorithms and

produce accurate classification across racial/ethnic groups (Gianattasio, Ciarleglio, and Power 2020; Gianattasio et al. 2019). The same data limitation persists in all algorithmic dementia classification methods that rely on ADAMS clinical dementia diagnosis for training underlying models, however— existing models cannot include detailed neuropsychological tests, which are critical in gold-standard dementia diagnoses and available in ADAMS but not available in the full HRS sample. The limited size of the ADAMS study (n=856), however, and the lack of refresher samples to supplement the original cohort recruited 20 years ago contribute to a need for algorithmic dementia classification methods that incorporate information from newer data sources. Because existing algorithmic dementia classification methods can only include predictors available for all participants in the population they aim to classify, they cannot be adapted to include newer and more comprehensive cognitive data such as the data available in the 2016 HCAP study (Langa et al. 2020).

This dissertation is motivated by (1) the need for dementia classification algorithms that can propagate available information from existing data sources into population-representative cohorts that do not contain a gold-standard sub-study, (2) the importance of developing algorithms that can support inferences about health disparities without relying on differential standards across racial-ethnic groups, and (3) the appeal of developing a model flexible enough to incorporate newer and more comprehensive cognitive test data. Using innovative applications of Bayesian statistical methods and latent class mixture modeling, the aim of this work is to build an algorithmic dementia classification framework that incorporates additional predictors important in gold-standard dementia diagnosis (e.g., detailed neuropsychological measures that are available in a substudy of a population-based survey) thereby strengthening algorithmic dementia classification in population-representative samples.

Results from this work will support studies of AD/ADRD incidence, prevalence, and disparities as well as studies evaluating determinants of AD/ADRD in nationally-representative samples that to date have not been used for these purposes. The methodological development in this dissertation uses existing data efficiently, adds insight to available studies, and lays the groundwork for further development of cutting-edge research in algorithmic dementia classification.

This dissertation is organized as follows: **Chapter 2** provides descriptions of relevant datasets, summarizes current dementia adjudication methods in cohort studies and developments in algorithmic dementia classification, and concludes with a brief overview of the core statistical topics on which the methods developed in this dissertation are built; **Chapter 3** describes the novel statistical framework for algorithmic dementia classification and provides a detailed illustrative example of the methods using ADAMS; **Chapter 4** gives an overview of the simulation study framework used in this dissertation and provides details on a simulation study that evaluated statistical properties of the Bayesian latent class mixture model and the role of HRS/HCAP study sample size when information from ADAMS is used to formulate the priors; **Chapter 5** discusses expanded simulation studies that evaluated the use of information from subsets of HCAP to formulate priors that are better calibrated to observed data and presents strategies for combining results from these analyses with results from analyses that used ADAMS to specify priors; **Chapter 6** presents results for algorithmic dementia classification in HCAP using the proposed Bayesian latent class mixture model and compares results to recently published estimates from HCAP investigators (Manly et al. 2022); the dissertation concludes with a discussion of future directions for research in **Chapter 7**.

This dissertation used computational and storage services associated with the Hoffman2 Shared Cluster provided by UCLA Institute for Digital Research and Education's Research Technology Group. All analyses and visualizations were done using R version 4.1.0 (R Core Team 2020), and all code is available on GitHub: *https://github.com/cshawsome/link-transport-integrate*.

Chapter 2 Background & Literature Review

The methods developed for this dissertation were motivated by challenges in algorithmic dementia classification in population-representative studies. This chapter provides background information on the motivating context and statistical methods in the dissertation. This chapter begins with a description of the three cohort studies used in this dissertation followed by a discussion of methods for dementia adjudication implemented in those and similar studies. The second half of the chapter provides an overview of statistical concepts necessary for understanding the methods development described in **Chapter 3**.

2.1: Dataset descriptions

Cohort studies are important for monitoring and understanding disease burden in a population. Several US-based and international cohort studies on aging and dementia exist (Manly et al. 2005; Lee and Dey 2020; Steptoe et al. 2013; Mejia-Arango et al. 2020; Zhao et al. 2014; Gómez-Olivé et al. 2018; Bienias et al. 2003; Lopez et al. 2003; Knopman et al. 2016; Kukull et al. 2002; Langa et al. 2005). These studies have enabled measurement of dementia incidence and prevalence (Mehta and Yeo 2017; Tang et al. 2001; Plassman et al. 2007; Hebert, Scherr, Bienias, Bennett 2003; Kuller et al. 2016), aided in risk factor studies (Kuller et al. 2016; Walter et al. 2016; Fishman 2017; Tschanz et al. 2013; Alpérovitch et al. 2003), and highlighted disease disparities (Chin, Negash, and Hamilton 2011; Cunningham et al. 2017; Babulal et al. 2019; Vable et al. 2018; Buckley et al. 2018; Mayeda et al. 2016). The methods developed in this dissertation use three existing US-based studies: (1) The Health and Retirement Study (HRS); (2) the Aging, Demographics, and Memory Study (ADAMS); and (3) the Harmonized Cognitive Assessment Protocol. A brief description of each study follows. **Figure 2.1** depicts study

timelines, sample sizes, availability of different measures, and the relationships between HRS and ADAMS and HCAP, which are both substudies of HRS.

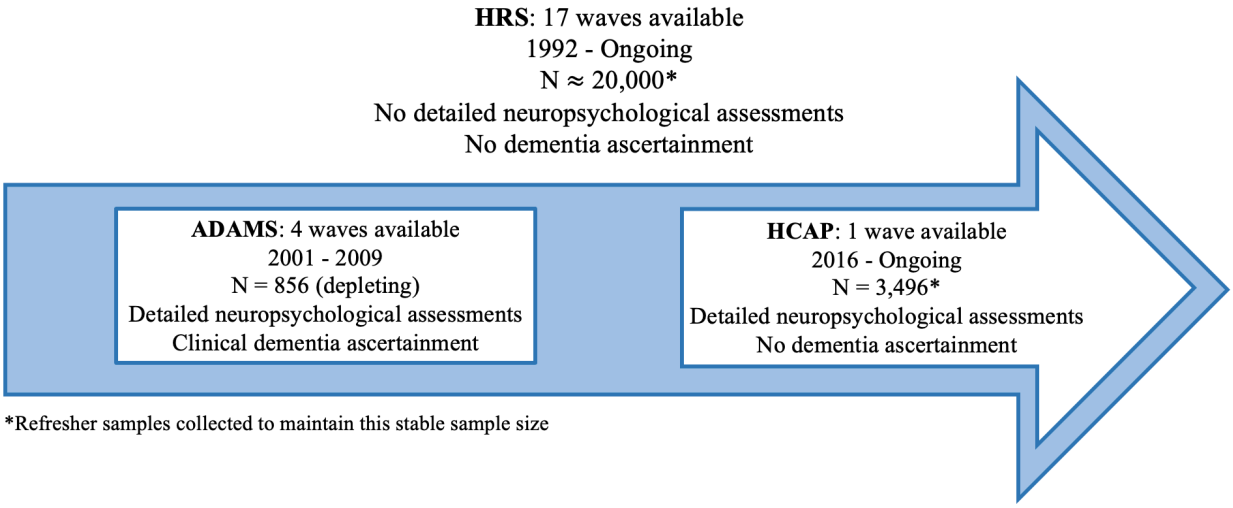


Figure 2.1: Timelines, sample sizes, and relationships among HRS, ADAMS, and HCAP studies.

2.1.1: The Health and Retirement Study

HRS is an ongoing, longitudinal, nationally-representative cohort study of Americans aged 50+ and their spouses (Sonnega et al. 2014). HRS was initiated in 1992 by the University of Michigan with follow-up every two years, alternating between phone and in-person interviews for about half of the sample— 50% of the sample is assigned to in-person interviews in one cycle while the other 50% of the sample completes a telephone interview, then each half is assigned to the other mode of interview at the next HRS follow-up wave. Over 37,000 individuals were initially enrolled; refresher cohorts are added every 6 years to maintain a stable sample size of about 20,000 participants. HRS data is publicly available; the study aims to be a resource for US trends in changing economic and health patterns associated with aging at both the individual and population level. As such, HRS collects data in four broad domains: income and wealth; health, cognition, and the use of healthcare services; work and retirement; and family connections. Most interviews are self-response, but proxy interviews are conducted if necessary to help maintain

coverage of the cognitively impaired. Beginning in 2006, in-person interviews included physical measures (e.g., height, weight, blood pressure, grip strength, walking speed). Thus, physical measures are available for HRS participants every four years. A complete description of the sample and available measures can be found in (Sonnega et al. 2014). Cognitive measures available in HRS are listed in **Appendix Table A.1**. Sample characteristics for the 2016 wave of HRS participants aged 70+ (the sample relevant to this work) are provided in **Table 4.1**.

In addition to the primary HRS survey, the University of Michigan conducts supplemental surveys in subsamples of HRS with more specific aims. For example, ADAMS and HCAP were designed to collect more detailed neuropsychological assessments on a subset of HRS participants. Gold-standard clinical dementia adjudication was conducted for ADAMS participants only.

2.1.2: The Aging, Demographics, and Memory Study

The Aging, Demographics, and Memory Study (ADAMS) was a substudy of HRS participants aged 70+ that performed gold-standard dementia adjudication for all participants (Heeringa et al. 2009; details provided in **Section 2.2.1:**). There are four waves of ADAMS data (A-D; 2001-2009); the goal of ADAMS was to follow a group of at least 850 individuals for dementia onset. To capture a wide range of cognition, ADAMS participants were sampled within 5 strata based on HRS cognitive assessment performance (range: low functioning-high normal) with oversampling of participants in the low-functioning category. The cognitively normal group was further stratified by age (70-79, 80+) and sex. ADAMS wave A comprised n=856 participants, and the sample was depleted as participants were either diagnosed with dementia or died. Neuropsychological assessments available in ADAMS are listed in **Appendix Table A.1**.

Sample characteristics for ADAMS wave A (baseline wave) are provided in **Table 4.1**.

2.1.3: The Harmonized Cognitive Assessment Protocol

HCAP is an ongoing substudy of HRS participants initiated in 2016 (Weir, Langa, and Ryan 2016). One wave of HCAP data is currently available, and the second wave of data collection is underway with delays due to the COVID-19 pandemic. The main goals of HCAP are (1) to provide a new HRS data resource to better assess determinants, prevalence, and the current and future costs and consequences of MCI and dementia in the US and (2) to facilitate harmonization of cognitive measurement with existing ongoing international studies on aging to study determinants, prevalence, and the impact of MCI and dementia worldwide (Langa et al. 2020). HCAP participants were randomly sampled from 2016 HRS participants who completed their interview prior to HCAP initiation. Participants were sampled based on their age eligibility (65+) and 2016 HRS marital status— half of age-eligible, uncoupled HRS participants and half of age-eligible, coupled participants were selected. Of those selected for HCAP, n=3496 participated. HCAP interviews were conducted in pairs when possible: (1) 1-hour target-respondent interview that included a detailed neuropsychological assessment and (2) 20-minute informant interview completed by an individual nominated by the target responder.

HCAP was modeled after the ADAMS study, but the HCAP research team used insights from fielding ADAMS to improve HCAP data collection. One limitation of ADAMS was the high cost of in-home assessments administered to participants, which required 3-4 hours and trained specialized personnel (see **Section 2.2.1:**). The ADAMS research team also performed gold-standard dementia adjudication for all participants, which added to the burden of conducting the study. HCAP, however, used a computer-assisted neuropsychological assessment

that only required 1 hour to complete and did not perform dementia adjudication for participants. Thus, the HCAP team was able to significantly reduce the cost per assessment and include a sample nearly 4 times larger than ADAMS. Models have recently been developed by HCAP investigators, however, to predict probability of impairment for participants (Manly et al. 2022). Sample characteristics for HCAP 2016 participants aged 70+ (the sample relevant to this work) are provided in **Table 4.1.**

2.2: Dementia adjudication methods

2.2.1: Gold-standard dementia adjudication and adaptations

Diagnosing dementia in large studies like HRS is a major challenge. The accepted gold-standard dementia adjudication method involving consensus diagnosis by an expert panel that evaluates results from detailed neuropsychological testing, clinical examination, an informant interview, and medical history, is incredibly time and resource intensive. For example, in ADAMS, the dementia adjudication protocol involved a 3-4 hour structured in-home assessment that required a nurse and specially trained neuropsychology technician. The assessment included (1) a battery of neuropsychological measures, (2) a self-reported depression measure, (3) a standardized neurological examination, (4) measured blood pressure, (5) collection of buccal DNA samples for APOE genotyping (AD risk factor), and (6) a 7-minute videotaped segment for portions of the cognitive status and neurological exams. An informant interview was also conducted; a knowledgeable informant was asked to provide information on (1) chronological history of cognitive symptoms, (2) medical history, (3) current medications, (4) current neuropsychiatric symptoms, (5) measures of severity of cognitive and functional impairment, (6) family history of memory problems, and (7) a caregiving questionnaire detailing the time and strain associated

with providing care. All testing was scored by the original technician, a second technician, and the supervising neuropsychologist. A preliminary diagnosis was assigned by a team including a geropsychiarist, neurologist, and cognitive neuroscientist after reviewing all in-home assessment and informant interview materials. The geropsychiatrist reviewed available medical records and revised preliminary analyses when warranted. Participants were placed into one of the following categories: normal/non-case, mild cognitive impairment, dementia of undetermined etiology, frontal lobe dementia, alcoholic dementia, ALS with dementia, probable AD, possible AD, probable vascular dementia, possible vascular dementia, probable Lewy Body dementia, Parkinson's disease, Huntington's disease, progressive supranuclear palsy, normal pressure hydrocephalus, Pick's disease, severe head trauma (with residual), mild ambiguous, depression, psychiatric disorder, mental retardation, alcohol abuse (past/current), stroke, other neurological conditions, other medical conditions, cognitive impairment no dementia (non-specified). Cases that were ambiguous were assigned to return in the next wave of data collection for re-evaluation (Langa et al. 2005).

The resource-intensive dementia adjudication protocol was a major limitation to the ADAMS study as it precluded recruiting a large sample (Langa et al. 2020). To circumvent the intensity of gold-standard dementia adjudication for all participants in a study, dementia researchers have implemented hybrid/augmented, or adaptive designs as an alternative. Hybrid/augmented designs use clinical dementia adjudication in a subset of participants to predict impairment in the rest of the sample (Bennett et al. 2002; Knopman et al. 2016; Bennett et al. 2012). Adaptive designs use stages of evaluation that increase in intensity as participants are flagged at each stage as either unimpaired or requiring further evaluation. Adaptive designs typically begin with a low-burden screening test in stage one and conclude with clinical dementia

adjudication in final stages (Tang et al. 2001; Lopez et al. 2003; Trittschuh et al. 2011; Demirovic et al. 2003; Fillenbaum et al. 1998). A 2021 study in the Kaiser Healthy Aging and Diverse Life Experiences (KHANDLE) cohort used an innovative hybrid adaptive design to adjudicate impairment status for all 1700 of its study participants. In KHANDLE, participants were either randomly assigned to undergo full clinical dementia evaluation or were evaluated for impairment using an adaptive design (Mungas et al. 2021).

2.2.2: Algorithmic dementia classification

In large, population-representative surveys like HRS, hybrid or adaptive designs are still too time- and resource- intensive to implement for all participants at each wave of data collection. Fully algorithmic dementia classification methods have been developed as an alternative. Much like hybrid/augmented designs, existing algorithmic dementia classification methods fit models (e.g., logistic models, probit models, ordered probit models) in studies with gold-standard adjudicated dementia cases and use estimated effects of covariates to predict probabilities of impairment for participants in larger studies. ADAMS has been an engine for the development of algorithms to predict impairment for participants in HRS (Crimmins et al. 2011; Wu et al. 2013; Hurd et al. 2013; Herzog and Wallace 1997). These algorithms were limited to including covariates that existed in HRS, however, and could not include detailed neuropsychological assessments available in ADAMS, which are critical measures in gold-standard dementia adjudication. Further, the sensitivity and specificity of early models differed by race/ethnicity, which merits careful consideration of whether their use is valid in disparities research (Gianattasio et al. 2019).

Recent dementia classification algorithms developed specifically for racial/ethnic disparities AD/ADRD research achieved non-differential sensitivity and specificity across racial/ethnic subgroups (Gianattasio, Ciarleglio, and Power 2020). Improvements were accomplished by considering main effects and two-way interactions between sociodemographic and health characteristics on an individual's dementia status and by using race-specific score cutoffs for dementia classification. Using racial/ethnic-specific score cutoffs is a topic of debate, however, because their use ignores effects of cultural or educational factors for which race is a proxy and can lead to denial of necessary services to these groups (Gasquoine 2009; Manly 2005; Manly and Echemendia 2007). Further, this method does not address the primary limitation of existing algorithmic dementia classification methods— they cannot be adapted to include comprehensive neuropsychological data from studies like ADAMS or newer and more comprehensive data from studies like HCAP.

The goal of this dissertation is to develop an algorithmic dementia classification framework flexible enough to incorporate neuropsychological data in the prediction model, thereby strengthening algorithmic dementia classification in population-representative studies.

2.3: Statistical topic review

The methods developed in this work and detailed in **Chapter 3** build on concepts in the statistical areas of finite mixture modeling and general location models. Models will be embedded in a Bayesian analysis framework to accommodate prior information (e.g., information from the ADAMS study) in analyses. Additional statistical tools are used throughout this work— bootstrapping is used to overcome challenges related to small sample sizes in **Section 3.4:** analyses and to create a superpopulation for the simulation study in **Chapter 4**,

missing data is addressed using multiple imputation and stratified hotdeck imputation as alternatives to complete-case analyses to preserve sample size in analyses, and standardization is used for dementia prevalence comparisons among racial/ethnic groups in analyses. What follows is a relatively brief but sufficient overview of necessary material from the statistical areas that form the core of methods development in this dissertation: the Bayesian analysis framework, finite mixture models, and general location models. Brief explanations of additional statistical tools used in analyses will be provided in context as they become relevant.

2.3.1: Bayesian analysis framework

There are two primary statistical paradigms: frequentist and Bayesian. Frequentist analyses are traditionally taught in introductory statistics courses and thus tend to be more familiar to applied researchers. Bayesian methods, though conceptually intuitive, can be computationally intense. For this reason, Bayesian methods have gained traction in the last 50 years due to increased data storage capacity and computational power. Frequentist and Bayesian methods diverge in their treatment of data being analyzed and parameters underlying the data-generating process. In the frequentist paradigm, data are considered random while parameters underlying the data generating process are considered fixed. Frequentist inference relies on large sample theory (e.g., the sample mean, $\bar{x}$, approaches the population mean, μ, as the sample size, n, increases).

In contrast, the Bayesian framework views data as fixed and parameters underlying the data generating process as random (i.e., coming from a probability distribution). Thus, Bayesian inferences are conditional on the data being analyzed. The primary reference for Bayesian methods in this dissertation is *Bayesian Data Analysis* by Gelman et al. (2014).

Bayesian analyses rely on expressing the joint distribution of model parameters and observed data as the product of a prior distribution and the sampling distribution (likelihood), then applying Bayes rule to arrive at an expression for the posterior distribution of the parameters θ given the data y:

$$p(\theta|y) = \frac{p(\theta, y)}{p(y)} = \frac{p(y|\theta)p(\theta)}{p(y)}. \quad (2.1)$$

Since (2.1) is conditional on the data y, $p(y)$ is a constant and we write

$$p(\theta|y) \propto p(y|\theta)p(\theta), \quad (2.2)$$

where $p(y|\theta)$ is the likelihood and $p(\theta)$ is the prior. In words, (2.2) reads, "the posterior is proportional to the likelihood times the prior." The goal of Bayesian analyses is to develop a model for the joint distribution $p(\theta, y)$ and perform the computations necessary to summarize the posterior distribution $p(\theta|y)$.

Appealing features of the Bayesian framework include the intuitive nature of the model being a tug-of-war between the prior and the likelihood (i.e., strong priors require more evidence in the data to overcome prior assumptions; analyses using uninformative priors are driven by the data and yield inferences similar to frequentist analyses), natural interpretations of the Bayesian credible interval (the probability of the model parameter lying in the interval), and the flexibility of this framework for analyzing complex statistical problems.

Sampling from the posterior distribution can be computationally challenging in complex models. When posterior distributions are more complex than the standard named distributions (normal, gamma, binomial, etc.), special techniques can be used to explore the sample space. Markov Chain Monte Carlo (MCMC) is an iterative sampling technique based on drawing parameter values θ from an approximate distribution and correcting the draws at the next

iteration to better approximate the target posterior distribution. When the prior and posterior distributions belong to the same distributional family, they are said to be conjugate. If the model can be specified using conjugate prior distributions, sampling from the posterior distributions is straightforward because conjugate distributions yield a closed-form expression for the posterior that can be sampled from directly. An MCMC technique that can leverage conjugate distributions nicely is the Gibbs sampler. Gibbs sampling, also known as alternating conditional sampling, draws components of the parameter vector θ conditional on other values of the parameter vector. Gibbs sampling is the technique used in this dissertation because components of the parameter vector in this work can be specified using conjugate distributions and posterior parameters can be sampled from named distributions. Several conjugate distributions exist; the relevant distributions for this work are listed in **Table 2.1**.

Prior	**Likelihood**	**Posterior**
Dirichlet	Multinomial	Dirichlet-Multinomial
Inverse Wishart	Normal	Normal-Inverse-Wishart

Table 2.1: Conjugate distributions appearing in dissertation methods.

Gabry et al. (2019) emphasize the role of data visualization at every stage of Bayesian modeling and outline the workflow for (1) prior predictive checks, (2) model diagnostics, and (3) posterior predictive checks using a real data example. Gabry et al. (2019) envision each stage of the Bayesian analysis workflow as part of an iterative process of specifying models, fitting models, evaluating model fit, and updating model specifications. The following sections provide general descriptions of each stage of the workflow; each stage as it relates to this project is illustrated in analyses in subsequent chapters.

2.3.1.1: Prior predictive checks

Understanding how assumptions encoded in the prior interact with the likelihood to affect posterior inferences is an important part of Bayesian modeling. Prior predictive checks are visualizations of synthetic datasets generated from prior distributions only (i.e., sampled from the prior predictive distribution). These checks are meant to assess compatibility between the prior and the data and whether assumptions encoded in the prior lead to realizations of the data that capture the full range of possible values? Ideal prior predictive checks would yield prior predictive distributions that are wider than but roughly centered around the observed data, demonstrating compatibility and that the full range of plausible values was captured by the prior. Priors may need to be tuned by hyperparameters (parameters of the prior distribution) to achieve this. Examples of prior predictive checks for this work are displayed in **Figure 3.4** and **Figure 3.5**.

2.3.1.2: Model convergence diagnostics

Once models are tuned so that prior predictive checks are satisfactory, several draws from the posterior predictive distribution should be sampled and model parameter convergence assessed across draws. Trace plots of model parameters are used to assess convergence. All model parameters should be monitored for convergence, but in cases where models contain many parameters, a subset of important parameters may be chosen instead. Ideal trace plots look like "fuzzy caterpillars" that stabilize around some value. If there is any type of pattern in the trace plot (e.g., increasing mean value across iterations), this indicates a lack of model convergence. Additional model tuning by adjusting hyperparameters or additional iterations may be required to

achieve proper convergence. Examples of convergence plots for this work are displayed in **Figure 3.6**.

Bayesian inferences are based on several independent draws from the posterior distribution. Iterative sampling can create challenges for drawing independent data as values may be correlated with those at previous iterations. In general, though, once trace plots show convergence, each subsequent draw is from the posterior distribution. However, this does not guarantee that the posterior distribution is insensitive to starting values in the parameter space. Model stability related to starting values is assessed by monitoring convergence of multiple chains initiated at different locations in the parameter space. Ideal diagnostic plots for multiple chain convergence would look like overlapping "fuzzy caterpillars" which would demonstrate that model convergence is robust to the starting point in the parameter space (Gelman et al. 2014). An example of multiple chain convergence plots is displayed in **Figure 3.7.**

2.3.1.3: Posterior predictive checks

Once proper model convergence is achieved, posterior predictive checks assess whether salient features of the data were captured by the model. "Salient features" are measured by statistics of interest to the researcher. In this work, I followed the example in Gabry et al. (2019) and assessed medians and skew for continuous variables. I also monitored counts for cross-classified categorical variables. Analogous to prior predictive checks, ideal posterior predictive distributions would be centered around the observed value of the statistic of interest. Examples of posterior predictive checks for this work are displayed in **Figure 3.8 - Figure 3.9**.

2.3.1.4: Summarizing the posterior distribution

Measures of centrality and spread are used to summarize posterior predictive distributions and perform Bayesian inference. Means or medians could be used for centers of posterior distributions and credible intervals are used to measure uncertainty. Analogous to frequentist confidence intervals, Bayesian α-level credible intervals are constructed by taking the lower $\frac{\alpha}{2} \times 100\%$ and upper $\left(1 - \frac{\alpha}{2}\right) \times 100\%$ tails of the posterior predictive distribution for a parameter of interest. The interpretation of a Bayesian credible is more intuitive than the interpretation of frequentist confidence intervals. A Bayesian 95% credible interval for a parameter θ, for example, is a set of values for which there is a 95% probability of θ belonging to that set.

2.3.1.5: Should we expect Bayesian and frequentist analyses to always coincide?

The choice between a Bayesian or a frequentist analysis can sometimes be more philosophical than pragmatic. For example, some statistical practitioners may feel hesitant to place prior distributions on model parameters since this can feel like a subjective process. Bayesian analyses are becoming more widely accepted, however, even in carefully monitored spaces like the FDA-regulated pharmaceutical industry (Boulanger and Carlin 2021).

Bayesian and frequentist analyses usually coincide when uninformative priors are used in a Bayesian model. Strictly specifying uninformative priors, however, ignores available existing information about the problem and robs the analyst of the greatest strength of the Bayesian framework: leveraging prior information to increase accuracy and efficiency of analyses. It would be more desirable to conduct a Bayesian analysis using available prior information and show that the Bayesian model is "well-calibrated" by frequentist standards.

Well-calibrated Bayesian models are desirable because analysts running these models are often not experts in Bayesian statistics or in the content-area application, thus, they will not have the skillset necessary to fine tune their model to the specific application context (Rubin 1984). Thus, Bayesian statisticians develop models that would ideally be applicable to a broad set of problems. Bearing this in mind, summaries of Bayesian analyses in the simulation studies in this dissertation include the frequentist metric of "95% interval coverage" to assess model calibration to frequentist standards.

There is no requirement that a 95% credible interval have nominal coverage across simulation runs because a 95% credible interval is simply a summary of the posterior predictive distribution. Rubin (1984) and more recently Gelman et al. (2020) discuss that one should expect at least nominal coverage from a 95% credible interval when the prior is *correct*. Knowledge of whether the prior is correct is only possible in a simulation study and indeed, Rubin and Gelman both discuss assessing the validity of Bayesian modeling procedures by generating synthetic data using a specified prior, analyzing the synthetic data using the same prior, and assessing coverage. Under these circumstances, interval estimates from a well-calibrated Bayesian model would be expected to have nominal coverage. This issue will be further discussed in context as it relates to results presented throughout the dissertation.

2.3.2: Finite mixture models

Finite mixture models provide a flexible, semiparametric modeling strategy in applications where a single distribution would not adequately capture important features of the data. The underlying assumption in finite mixture models is that the data are composed of a finite number of subgroups with meaningfully different characteristics. Finite mixture models are also useful as

a strategy for modeling complex distributions as a mixture of more familiar or tractable distributions. The probability density function of a finite mixture distribution is

$$f(y) = \sum_{i=1}^{g} \lambda_i f_i(y), \tag{2.3}$$

where λ_i are the mixing proportions (non-negative and sum to 1) and $f_i(y)$ are the component densities. Component densities are usually assumed to be from the same family (e.g., normal distributions with different means and variances), but it is possible to have a mixture of different distributions. The inference goal in finite mixture models is the correct mixing proportions λ_i for the component densities.

Mixture models are foundational in many statistical areas including clustering, discrimination, and latent class analyses and have gained traction as a method in a variety of fields outside of statistics. An example in the psychological literature is Belin and Rubin's (1995) use of a finite mixture model for the reaction times of patients with schizophrenia. For a more comprehensive review of recent methodological developments and applications of finite mixture modeling, see Mclachlan, Lee, and Rathnayake (2019).

One challenge in mixture modeling is choosing the number of groups g. This can be included in the inference process or be driven by prior knowledge of the application context. Identifiability is also a concern in some mixture models. A parametric family of distributions is said to be identifiable if distinct values of the parameters result in distinct members of the family of densities. In the case of finite mixture modeling, nonidentifiability of the model leads to an issue called "label switching" where subgroups cannot be distinguished from one another. Titterington (1985) argues, however, that most finite mixtures of continuous densities are identifiable.

2.3.2.1: Bayesian analysis of finite mixture models: priors on group membership

Prior information is incorporated in the finite mixture models used for this work in two ways: (1) the number of mixture distributions is based on clinically meaningful cognitive impairment groups based on ADAMS (4 groups: Unimpaired, MCI, Dementia, Other), see **Section 2.2.1:** and **Table A.2** and (2) mixing proportions are informed by models fit in ADAMS relating important predictors of impairment group membership (sociodemographic characteristics, health and health behavior measures, general cognitive assessments, and detailed neuropsychological assessments) to clinical cognitive impairment diagnosis. Additional details on specifying priors for latent class membership are provided in **Section 3.4.1:**.

2.3.3: General location model

The general location model is a framework for modeling a mix of categorical and continuous variables (Olkin and Tate 1961; Little and Schluchter 1985; Schafer 1997). Using the general location model, an observation's continuous variables are modeled using normal distributions with parameters determined by the observation's contingency cell membership (i.e., cross-classification of categorical variables). Following the notation of Schafer (1997), let $W_1, \dots W_p$ be a set of categorial variables and $Z_1, \dots Z_q$ be a set of continuous variables. Then for a sample Y with n observations, Y can be represented by the $n \times (p+q)$ matrix (W, Z).

2.3.3.1: Unrestricted general location model

The unrestricted general location model uses main effects and all interaction effects of categorical variables W to model continuous variables Z. Suppose W_j takes possible values $1, 2, \dots, d_j$. The categorical data can be summarized by a D-dimensional contingency table with

D = $\prod_{j=1}^{p} d_j$ possible cells. Let $C = \{c_d : d = 1, \ldots, D\}$ represent the vector of cell counts. Another characterization of W comes from defining a d-vector u_i with a 1 in position d if the observation falls into contingency cell d and 0 otherwise. Then, let U be an $n \times D$ matrix with rows $u_i^T, i = 1, \ldots, n$. $U^T U$ = diagonal(C). By the independence assumption on sampled units, all the information in W is also contained in C. Thus, we can model the data Y as

$$C \sim M(n, \pi) \tag{2.4}$$

$$Z_i | u_i \sim N_q(\mu_d, \Sigma). \tag{2.5}$$

For the distribution (2.4), $\pi = \{\pi_d : d = 1, 2, \ldots, D\}$ is a vector of cell probabilities parameterizing the multinomial distribution corresponding to C. Note that the mean of the normal distribution in (2.5) is indexed by d, denoting that the means vary by contingency cell but the covariance structure Σ is assumed to be constant across cells. The model for Z could also be regarded as a multivariate regression $Z = U\mu + \varepsilon$. The number of free parameters in the unrestricted model is thus $(D - 1) + Dq + q(q + 1)/2$.

2.3.3.2: Restricted general location model

The unrestricted general location model is suitable for datasets with many observations relative to the total number of cells D. Even when this is the case, however, challenges can arise if there are too many small contingency cell counts. To address this, the number of free parameters in the model can be reduced by: (1) placing a loglinear restriction on the cell probabilities or (2): defining a linear model for the within-cell means. I used method (2) to restrict the model in analyses.

Instead of considering the model as a multivariate regression $Z = U\mu + \varepsilon$, which includes all main effects and interaction effects of categorical data on continuous values, a design matrix

A can be defined to specify the desired effects to include in the model. Let $\mu = A\beta$. Taking $A = I$ will return the unrestricted model, but a different design matrix will lead to a reduced number of parameters β. For example, analyses for this work used the following design matrix which specifies an intercept and main effects of W_1, W_2, W_3 only:

$$A = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 \end{bmatrix}. \tag{2.6}$$

In this setup, estimability depends on the rank of UA instead of just U and parameters in the model may be estimable even with some random zeros in the contingency cells.

2.3.3.3: Restricted General Location Model in a Bayesian Framework

Prior information is incorporated in the restricted general location model used for this analysis in three ways: (1) distributions of contingency cell counts are based on observed counts in ADAMS and (2) means and (3) variances for continuous variables are based on parameters estimated in ADAMS. Additional details for specifying prior distributions are provided in **Section 3.4.1:**.

Chapter 3 Methods Development: The Bayesian Latent Class Mixture Model

Methods discussed in the statistical topic review presented in **Section 2.3:** were combined and extended to create a novel latent class mixture modeling framework for Bayesian algorithmic dementia classification. I used ADAMS, an HRS substudy that included detailed neuropsychological assessments and gold-standard dementia adjudication for all participants (see **Section 2.1.2:**), to develop and validate the Bayesian latent class mixture model. Details are presented in this chapter which discusses the motivation, formulation, and implementation steps for the Bayesian latent class mixture model.

3.1: Dataset preparation

To simplify matters while formulating the algorithmic dementia classification framework, I restricted analyses to a complete-case subset of ADAMS Wave A data for which all relevant covariate measures (i.e., sociodemographic characteristics, general cognitive assessments, neuropsychological assessments, and health characteristics (see variable selection steps discussed in **Section 3.2:**) were available (n=520). A more general framework that incorporates methods for handling missing data is discussed in **Chapter 4**.

3.1.1: Data cleaning

I collapsed ADAMS marital status categories “single”, “divorced”, and “separated” into a “not married/partnered” category and preserved the ADAMS “married/partnered” and “widowed” categories.

Self-reported measures of health characteristics are only available in the core HRS survey with follow-up waves that do not necessarily align with ADAMS interview waves. ADAMS wave A interview dates ranged from 2001-2004. The relevant HRS interview years are 2000, 2002, and 2004. If ADAMS and HRS interviews were conducted in the same year, ADAMS interviews took place after HRS interviews. I used values for health behaviors and characteristics (history of stroke, hypertension, diabetes, heart disease, cancer; current smoking status; BMI; alcohol consumption) from the HRS wave closest to the ADAMS interview wave (i.e., HRS 2000 for ADAMS 2001, HRS 2002 for ADAMS 2002 and 2003, and HRS 2004 for ADAMS 2004). I characterized alcohol use (no drinking, moderate drinking, heavy/high risk drinking) according to the 2020 Dietary Guidelines for Americans (U.S. Department of Agriculture and U.S. Department of Health and Human Services 2020).

Measures of functional ability, Instrumental Activities of Daily Living (IADLs) and Activities of Daily Living (ADLs), are only available in the HRS core interview. Thus, measures from a representative wave of HRS were used for ADAMS wave A as described above for other self-reported health measures. I used average scores for portions of the assessment that used multiple items to assess the same cognitive or functional domain. These portions of the assessment included IADLs, which asked participants to rate their level of difficulty with using the telephone, taking medication, and handling money; ADLs, which asked participants to rate their level of difficulty with bathing, eating, dressing, walking across a room, and getting in or out of bed; and proxy cognition inventories adapted from the Informant Questionnaire on Cognitive Decline in the Elderly (IQCODE), which asked informants to rate the participants' ability to perform memory-related tasks compared to 2 years ago (Jorm 2004).

Scores on neuropsychological exam and general cognitive assessment tasks are ordinal categorical variables but were analyzed on a continuous scale. The Mini-Mental State Exam (MMSE), a 20-item test of global cognition (Folstein, Folstein, and McHugh 1975), is usually analyzed as a sum score in practice (Arevalo-Rodriguez et al. 2015). I transformed the MMSE sum score from its original scale [0, 30] to a normalized scale [0, 100] using a transformation developed and validated by Philipps et al. (2014). Normalized MMSE possesses better psychometric properties compared to raw MMSE and more closely resembles a normal distribution. A brief discussion about the psychometric properties of raw MMSE scores and the transformation is available in **Appendix Section A.1:**. The transformation function is available in the NormPsy R Package (Proust-Lima and Philipps 2018). I also used sum scores for the Boston naming test, an assessment where participants identify a series of line drawings of common items (Morris et al. 1989). I used the highest score for tasks where multiple trials were conducted (i.e., the CERAD immediate word list recall task (Morris et al. 1989) and backwards counting from 20 and 86 tasks). Prompts and scoring for ADAMS neuropsychological test items is available in **Appendix Table A.1**.

ADAMS clinically adjudicated cognitive impairment classes were collapsed into four categories: (1) Unimpaired (n=211), (2) Mild Cognitive Impairment (MCI, n=65), (3) Dementia (n=158), and (4) Other (cognitive impairment due to conditions other than progression to dementia [e.g., depression, stroke, or other medical/neurological conditions], n = 86). A complete list of ADAMS adjudicated impairment categories and the corresponding collapsed category is provided in **Appendix Table A.2**.

3.1.2: Data splitting

I split the complete-case ADAMS Wave A sample into 70% training (n=364) and 30% hold-out (n=156) sets for internal and external validation of the Bayesian latent class mixture model. Characteristics for the training and hold-out samples for variables chosen for inclusion during the variable selection process (see **Section 3.2:**) are presented in **Table 3.2**. Though the data split was random, there were slight differences between the training and hold-out samples due to small sample sizes. The hold-out sample was composed of a slightly higher proportion of White participants and slightly lower proportion of Black participants compared to the training sample. The hold-out sample had fewer participants with a history of stroke, slightly fewer unimpaired participants, and slightly more participants with dementia.

3.2: Modeling and variable selection of important predictors of impairment

An important initial step in formulating this algorithmic dementia classification framework was understanding important predictors of ADAMS adjudicated cognitive impairment classes. Rather than defining one model for the multi-level categorical outcome (Unimpaired vs. Other vs. MCI vs. Dementia), I used multi-part models to specify separate logistic regression models at each stage of classification. Three models were used to distinguish between impairment classes: (1) Unimpaired vs. Impaired, (2) Other vs. MCI or Dementia, and (3) MCI vs. Dementia. Models (2) and (3) were conditional on individuals being classified as being impaired or having MCI or dementia, respectively, thus models were fit in different subsets of the data. Modeling impairment classes this way naturally accommodates non-linear relationships and different subsets of predictors (Olsen and Schafer 2001). Based on clinical dementia assessments, candidate predictors for models included sociodemographic characteristics, general cognitive

assessments, neuropsychological assessments, and health characteristics. A complete list of candidate predictors for the multi-part models is provided in **Appendix Table A.3**.

Important predictors of ADAMS diagnosed impairment class were chosen by fitting multi-part models in the overall ADAMS sample (n=826). Let $G_i, i = 1, \ldots, 826$, denote the ADAMS adjudicated cognitive impairment class (group) for individual i in the ADAMS training sample,

$$G_i = \begin{cases} 1 & \text{if participant } i \text{ was classified as Unimpaired} \\ 2 & \text{if participant } i \text{ was classified as having Other impairment} \\ 3 & \text{if participant } i \text{ was classified as having MCI} \\ 4 & \text{if participant } i \text{ was classified as having Dementia} \end{cases} \quad (3.1)$$

Letting X denote the vector of candidate predictor variables including an intercept and $\boldsymbol{\gamma_1}, \boldsymbol{\gamma_2}, \boldsymbol{\gamma_3}$ be vectors of regression coefficients, the following logistic regression models were fit in the ADAMS training sample:

$$logit(P(G = 1|X)) = \boldsymbol{\gamma_1} X \quad (3.2)$$

$$logit(P(\mathrm{G} = 2|X, G \neq 1)) = \boldsymbol{\gamma_2} X \quad (3.3)$$

$$logit(P(G = 3\,|X, G \neq 1, G \neq 2) = \boldsymbol{\gamma_3} X. \quad (3.4)$$

Variables were entered sequentially starting with fully observed variables; variables that led to more than 25% missing observations in the model were not considered. As an initial simplifying assumption to identify important predictors, predictors of cognitive status that were significant at the p=0.05 level were retained in the models. Updates to this process are discussed in **Section 4.2.2.3:**. **Table 3.1** lists the specific predictors with non-zero regression coefficients reflecting that $\boldsymbol{\gamma_1}, \boldsymbol{\gamma_2}, \boldsymbol{\gamma_3}$ are distinct from one another. Sample characteristics for these selected predictors in the ADAMS training and testing sample are provided in **Table 3.2**.

Model 1 Unimpaired vs. Impaired	Model 2 Other vs. MCI or Dementia	Model 3 MCI vs. Dementia
Normalized MMSE Score	Normalized MMSE Score	Normalized MMSE Score
Immediate Word Recall	Immediate Word Recall	Immediate Word Recall
Age	Age	
Race/Ethnicity		
Serial 7s		
Word List Recognition (yes)		
Story Recall (Immediate)		
Average Proxy Cognition (Jorm IQCODE)		
	Delayed Word Recall	
		IADLs
		BMI
		Stroke History (yes/no)

Table 3.1: Variables included in multi-part models (**Equations (3.1) - (3.4)**) for predicting cognitive impairment categories in ADAMS.

Ideally, all important predictors of impairment would be available for participants we aim to classify. Crucial measures are often unavailable in the larger sample where we aim to predict impairment but are available in a subset of the study. For example, Total MMSE, which was a consistently important predictor across the multi-part models (**Table 3.1**) is only available in ADAMS, not HRS where we aim to predict dementia. A reasonable strategy would be to impute important predictors for the rest of HRS using methods like multiple imputation (van Buuren 2019; Rubin 1996) or semi-supervised learning (Zhang, Brown, and Cai 2019). While it is possible to include indicators for different classes and even interaction terms by class in these imputation methods, these are ultimately single-class modeling techniques. Several

neuropsychological measures in ADAMS possess distributions that are skewed and differ significantly by impairment class, however, (**Figure 3.1**) and initial modeling attempts using multiple imputation did not recover observed values in the tails of the distributions.

Variable	Overall (Wave A) N = 520	ADAMS Training (Wave A) N = 364	ADAMS Hold Out (Wave A) N = 156
Age, Mean (SD)	82.4 (6.3)	82.6 (6.3)	82.0 (6.5)
Race/Ethnicity, n (%)			
White	382 (73.5%)	263 (72.3%)	119 (76.3%)
Black	89 (17.1%)	67 (18.4%)	22 (14.1%)
Hispanic	49 (9.4%)	34 (9.3%)	15 (9.6%)
BMI, Mean (SD)	25.6 (5.2)	25.2 (4.8)	26.4 (6.1)
History of stroke, n (%)	101 (19.4%)	75 (20.6%)	26 (16.7%)
IADLs, Mean (SD)	0.6 (1.0)	0.5 (1.0)	0.6 (1.0)
Serial 7s, Mean (SD)	2.4 (2.0)	2.3 (2.0)	2.5 (2.0)
Immediate word recall, Mean (SD)	5.7 (2.3)	5.7 (2.3)	5.7 (2.4)
Delayed word recall, Mean (SD)	3.7 (2.7)	3.7 (2.7)	3.7 (2.8)
Total MMSE (normalized), Mean (SD)	54.2 (24.9)	54.3 (25.0)	54.1 (24.7)
Word recall (yes), Mean (SD)	8.2 (2.3)	8.2 (2.3)	8.3 (2.3)
Immediate story recall, Mean (SD)	13.7 (9.0)	13.6 (8.9)	14.0 (9.4)
Average Jorm IQCODE, Mean (SD)	3.3 (0.7)	3.3 (0.7)	3.3 (0.7)
Adjudicated impairment, n (%)			
Unimpaired	211 (40.6%)	151 (41.5%)	60 (38.5%)
MCI	65 (12.5%)	45 (12.4%)	20 (12.8%)
Dementia	158 (30.4%)	108 (29.7%)	50 (32.1%)
Other	86 (16.5%)	60 (16.5%)	26 (16.7%)

Table 3.2: Sample characteristics for complete-case ADAMS overall and stratified by training, and hold-out samples for participants and variables included in the illustrative example.

3.3: The Bayesian latent class mixture model

Transitioning from single-distribution modeling strategies to a mixture of distributions was motivated by the practice in cohort studies with gold-standard dementia adjudication of adjudicating participants into different, clinically meaningful, cognitive impairment groups. One of the challenges of modeling data using mixture distributions is choosing the number of

distributions to use (see **Section 2.3.2:**). The 4-class mixture used in this analysis was motivated by ADAMS clinically adjudicated cognitive impairment categories and the practice of classifying individuals as having no cognitive impairment, MCI, or dementia in cohort studies with gold-standard adjudication (Manly et al. 2005; Wilson et al. 2010; Lopez et al. 2012; Knopman et al. 2016; Trittschuh et al. 2011; Demirovic et al. 2003; Bennett et al. 2012; Plassman et al. 2007). Further, visualizing the distribution of continuous variables in ADAMS stratified by cognitive impairment classes revealed that a mixture of distributions might do a better job of recovering the observed overall distribution of these variables (**Figure 3.1**).

Any dataset could be viewed as a mixture of individuals who have no impairment, MCI, dementia, or other impairment. The inference goal of algorithmic dementia classification is identifying the correct mix of these individuals in a given study. Latent class mixture models simultaneously model data and infer individual impairment class membership (see **Section 2.3.2:**). I embedded the three major steps of the proposed latent class mixture modeling approach for algorithmic dementia classification in a Bayesian framework to incorporate prior information from the ADAMS study into the model (see **Section 2.3.1:** for a brief overview of the Bayesian analysis framework). Broadly, the steps of the Bayesian latent class mixture model are: (1) make a synthetic version of a dataset with detailed neuropsychological assessment data but unknown cognitive impairment classification; (2) the latent class mixture model assigns impairment status (unimpaired, MCI, dementia, other) to observations in the synthetic dataset— since the mixture of impairment classes determines the distributions of synthetic data, we have increased confidence in the predicted impairment classes when the synthetic data closely resemble the real

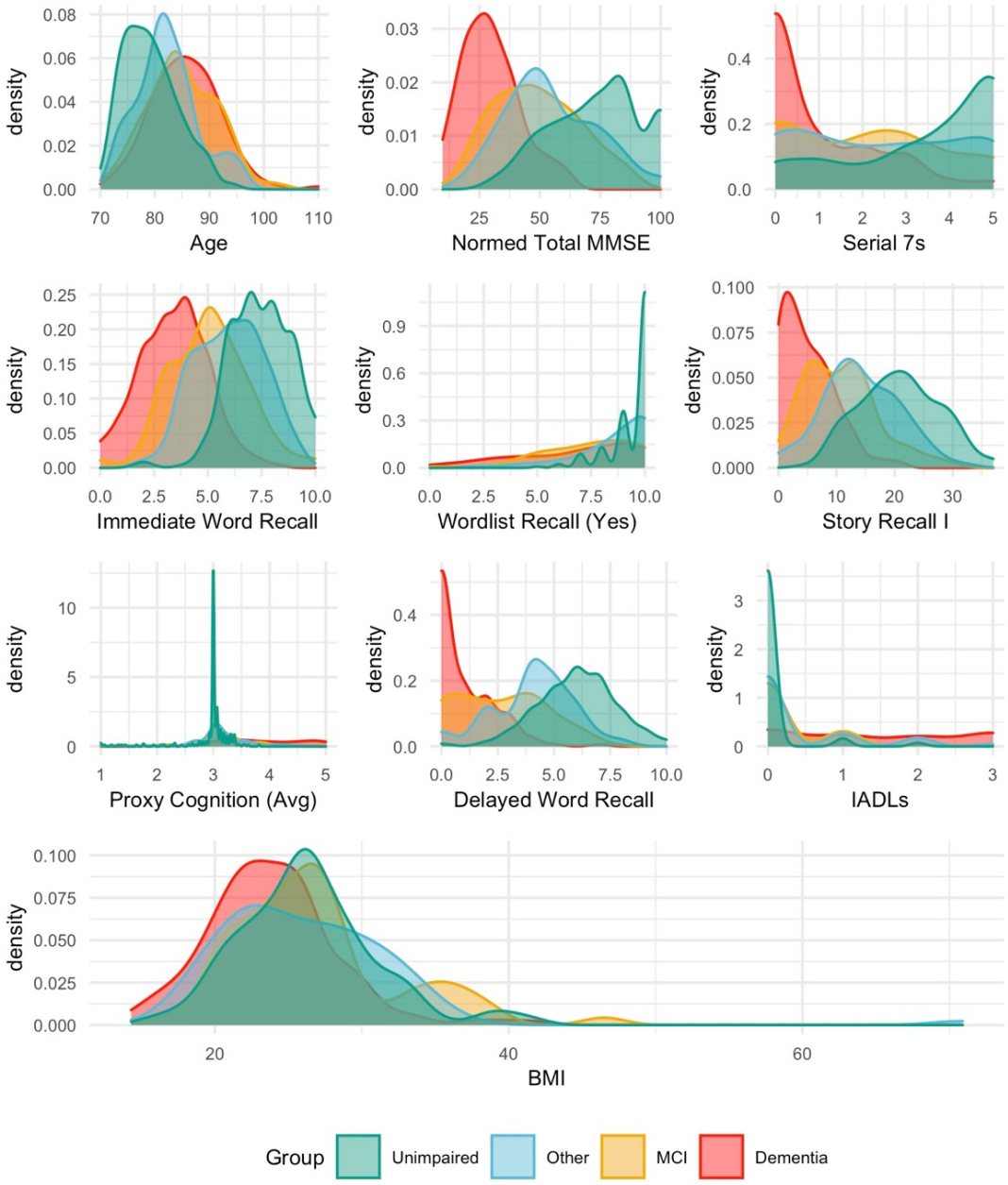


Figure 3.1: Density plots for continuous variables in **Table 3.1** stratified by ADAMS cognitive impairment class.

data (3) generate many synthetic datasets to quantify uncertainty in inference. A schematic of this process is shown in **Figure 3.2**.

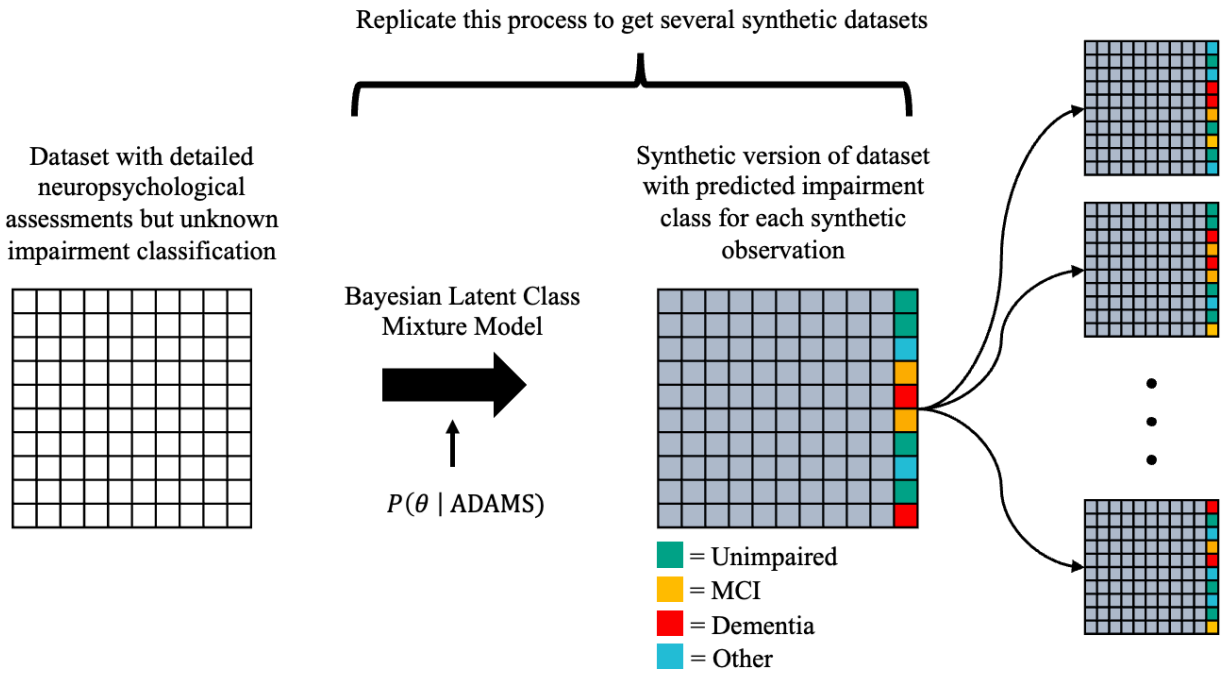


Figure 3.2: Schematic of Bayesian latent class mixture model for generating synthetic datasets. θ = parameters in the model. $P(\theta \mid \text{ADAMS})$ indicates that information from the ADAMS study is used to specify priors in the model.

3.3.1: Extending the general location model

The Bayesian latent class mixture model outlined above is an extension of the general location model which provides a framework for modeling a mix of categorical and continuous variables (see **Section 2.3.3:**). Briefly, the general location model uses normal distributions to model continuous variables with distributional parameters determined by an observation's contingency cell membership (cross-classification of categorical variables). Density plots of continuous variables in **Table 3.1**, stratified by ADAMS cognitive impairment classes are shown in **Figure 3.1**. By inspection, several of the variables could reasonably be modeled as mixtures of normal distributions, which motivated the use of the general location model in this framework. The formulation of the general location model implemented in this dissertation uses contingency-cell specific means but common variances across contingency cells.

Following the notation of Schafer (1997), let W_1 and W_2 be the categorical variables in **Table 3.1**, race/ethnicity (White, Black, Hispanic) and stroke history (ever/never), and let $Z_1, Z_2, \ldots Z_{10}$ be the continuous variables in **Table 3.1**. Let $X = (W, Z)$ be an $n \times 12$ matrix of observed data. Contingency cell membership was determined by cross classification of two categorical variables, race/ethnicity and stroke history, yielding 6 possible cells. Let $C = \{c_d : d = 1, 2, \ldots 6\}$ be the vector of observed counts for each contingency cell and let U be an $n \times 6$ matrix with rows u_i^T, where u_i is a 6-vector with a 1 in position d if observation i falls into cell d and 0s in all other position. All the information about W is contained in C, thus the distribution of X within each cognitive impairment group (latent class) G can be characterized by $f(X|\theta_G) = f(W, Z|\theta_G) = f(C, Z|\theta_G) = f(Z|C, \theta_G) f(C|\theta_G)$, where

$$C|\theta_G \sim M(n_G, \pi_G) \tag{3.5}$$

$$Z_i|u_i \sim N_{10}(\mu_{d_G}, \Sigma_G) \tag{3.6}$$

and θ_G is a vector of cognitive impairment group-specific parameters for the model. For the distribution in (3.5), $\pi_G = \{\pi_{d_G} : d_G = 1, 2, \ldots, 6\}$ is a vector of cell probabilities parameterizing the multinomial distribution corresponding to C in impairment group G. Note that the mean of the normal distribution in (3.6) is indexed by d_G, denoting that the means vary by contingency cell and across cognitive impairment groups but with an assumed constant covariance structure Σ_G across cells within an impairment group. Due to small cell counts, the unrestricted form of the general location model (see **Section 2.3.3.1:**) which would estimate all race/ethnicity by stroke interactions effects on continuous variables Z could not be used. Thus, the restricted formulation was used instead (see **Section 2.3.3.2:**) to include only main effects of race/ethnicity and stroke on continuous variables Z. In this model, the 6×10 matrix μ_G is restricted to the form

$$\mu_G = A\mathrm{B}, \tag{3.7}$$

where $\mu_G^T = [\mu_1 \mid \mu_2 \mid ... \mid \mu_6]$ and A is a 6×4 ANOVA-like design matrix that specifies an intercept and main effects for race/ethnicity and stroke on Z only,

$$A = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 \end{bmatrix}. \tag{3.8}$$

3.3.2: Generating synthetic datasets

I embedded the restricted general location model described above in a Bayesian latent class mixture model to generate synthetic versions of ADAMS Wave A data. I used Gibbs sampling (see **Section 2.3.1.2:**) to sample model parameters from their posterior distributions. A diagram of the data generating model at iteration b of the MCMC chain is provided in

Figure 3.3, including the specific posterior distributions from which variables were sampled. I derived posterior distributions using similar techniques to those outlined in Schafer (1997) and Gelman et al. (2014). Complete derivations are provided in **Appendix C**.

At iteration b of an MCMC chain, let $p(G_i^b)$ be the predicted probability of cognitive impairment group membership for the ith participant based on the multi-part models defined by **(3.1) - (3.4)**. Individuals were classified into distinct cognitive impairment classes by drawing from a Bernoulli distribution with probability of success equal to predicted probabilities from the models. Because the multi-part models were conditionally specified, Bernoulli draws were made sequentially using $p(G_i^b = \text{Unimpaired})$, $p(G_i^b = \text{Other} \mid G_i^b \neq \text{Unimpaired})$, $p(G_i^b = \text{MCI} \mid G_i^b \neq \text{Unimpaired}, G_i^b \neq \text{Other})$, $p(G_i^b = \text{Dementia} \mid G_i^b \neq \text{Unimpaired}, G_i^b \neq \text{Other}, G_i^b \neq \text{MCI})$. Once a participant was assigned to a cognitive impairment group, no other predicted probabilities were

used since membership in a more severely impaired group was conditional on non-membership in the less impaired groups.

Let G_i^b be a participant's cognitive impairment group membership at iteration b. The restricted general location model was used to model data within each subset of participants belonging to the same impairment group G^b. The impairment group-specific Bayesian formulation of the restricted general location model at iteration b was

$$Z_i|u_i \sim N_{10}(\mu_{d_G}, \Sigma_G) \tag{3.9}$$

$$C_G \sim M(n_G, \pi_G) \tag{3.10}$$

$$\mathrm{B}_G|\Sigma_G \sim MN_{4\times 10}(\mathrm{B}_{0_G}, V_{0_G}, \Sigma_\mathrm{G}/\kappa_0) \tag{3.11}$$

$$\Sigma_G \sim W_{\nu_0}^{-1}\left(\Lambda_{0_G}^{-1}\right) \tag{3.12}$$

$$\pi_G \sim D(\alpha_G) \tag{3.13}$$

with hyperparameters κ_0, ν_0, and where u_i indicates the contingency cell membership of an observation and $\mu_G = A\mathrm{B}_G$.

Sampling in this framework was fast and convenient by properties of conjugate distributions (see **Section 2.3.1:**), thus computational time was not a challenge in this model. Small contingency cell counts in the ADAMS training sample, however, led to difficulty specifying priors. See **Section 3.4.1:** for a detailed discussion on techniques for overcoming this

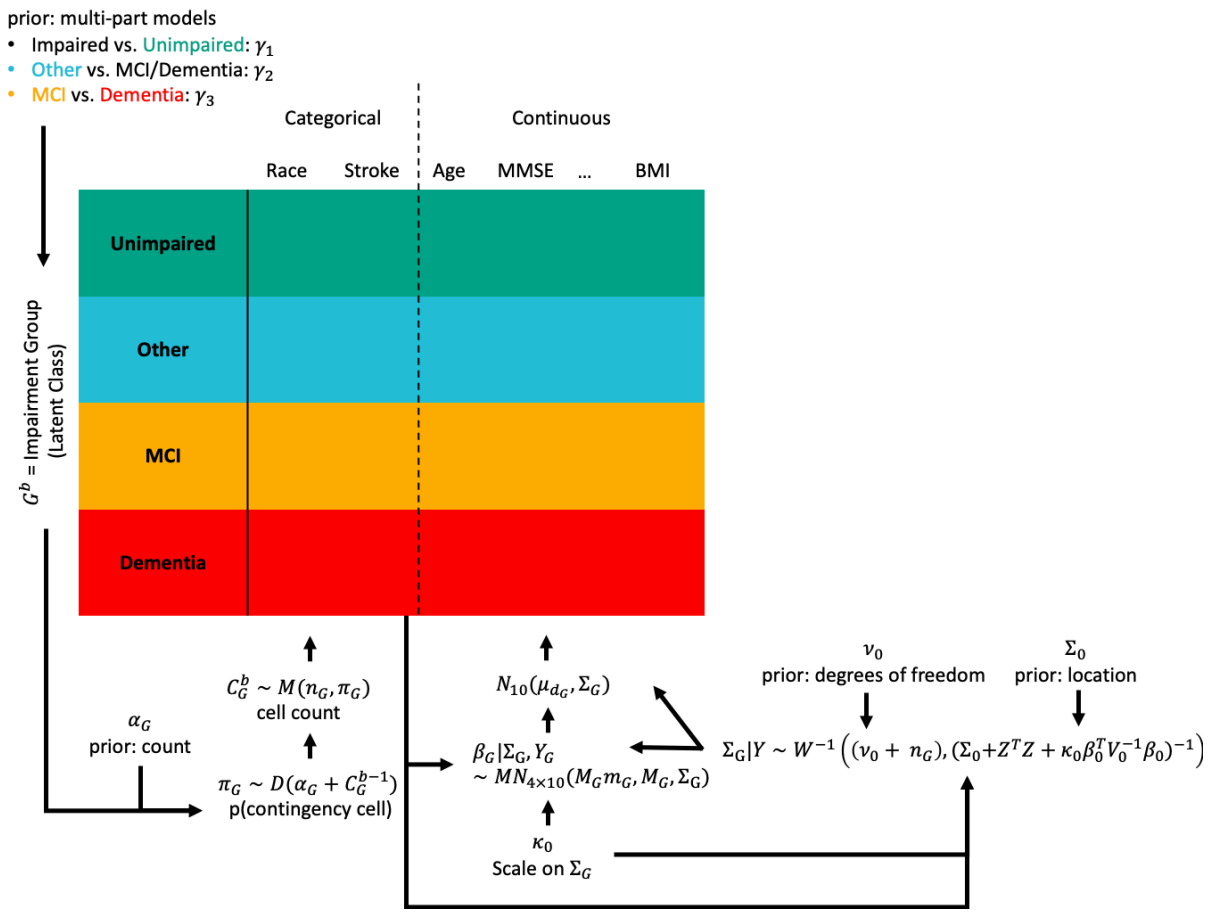


Figure 3.3: Data-generating model for iteration b of an MCMC chain for parameters in the Bayesian latent class mixture model. M = Multinomial distribution; D = Dirichlet distribution; N = Normal distribution; MN = Matrix normal distribution; W^{-1} = Inverse Wishart distribution.

challenge. To simplify monitoring convergence of continuous variable variances, all continuous variables were standardized prior to the creation of synthetic datasets. Thus, prior distributions were specified using standardized versions of continuous variables.

3.4: Illustrative example: algorithmic dementia classification in the Aging, Demographics, and Memory Study

I trained the Bayesian latent class mixture model in the ADAMS training sample (n=364) and used it to create 1000 synthetic versions of the ADAMS training sample (n=364; internal

validation) and the ADAMS hold-out sample (n=156; external validation). Details on specifying prior distributions, model diagnostic plots, and algorithmic dementia classification results are presented below for both training and hold-out samples.

3.4.1: Specifying prior distributions

The ADAMS hold-out sample was reserved purely for external model validation; thus, prior distributions were specified using the ADAMS training sample only. There were 14 contingency cells out of a total 24 in the ADAMS training sample with fewer than 10 participants. These small cell counts created difficulties in estimating parameters for continuous variables conditional on their contingency cell membership. For example, there were no Hispanic participants with stroke history in the Unimpaired, MCI, or Other groups; and there was only one White participant with stroke history and one Black participant with stroke history in the MCI group.

I made an initial attempt with non-informative priors, but those created model convergence issues. Attempts to increase prior cell counts by using observed cell counts in the larger HRS sample improved convergence but led to poor model fit due to differences in sample characteristics between HRS and ADAMS (e.g., ADAMS was oversample for impaired participants, so race/ethnicity by stroke distributions differed between HRS and ADAMS). Specifying prior distributions by bootstrap sampling (resampling with replacement (Efron and Tibshirani 1994)), proved to be a valuable technique to overcome these challenges.

Since the goal of this analysis was to create realistic synthetic versions of ADAMS, bootstrapping seemed like a promising way to reproduce the empirical distribution of the data while avoiding the need to make oversimplifying distributional assumptions. Prior distributions

were obtained through a three-step process: (1) resample ADAMS data with replacement, drawing a sample of equal size to the original sample (non-parametric bootstrap). (2) calculate and store parameter estimates characterizing effects of covariates on cognitive impairment class membership, contingency cell counts, and effects of contingency cell membership on continuous covariates. (3) repeat the process 10,000 times to represent both sampling variability and estimation uncertainty in model parameters. This process was motivated by Bayesian non-parametric methodology (Rubin 1981) and empirical Bayes concepts (Casella 1985).

Priors on latent classes were specified using the multi-part models described by **Equations (3.1)-(3.4)**. The models were fit in each of the 10,000 bootstrapped ADAMS datasets. Vectors of effect estimates ($\gamma_1, \gamma_2, \gamma_3$) for covariates in each model (see **Table 3.1**) were saved from each bootstrap sample, yielding a distribution of 10,000 values for each parameter. A random vector of effects was sampled for each model at each iteration b of the MCMC chain for synthetic data generation.

Priors on categorical variables were specified by cross-classifying observations in each ADAMS bootstrap sample into race/ethnicity x stroke contingency cells. Cognitive impairment-group specific race/ethnicity x stroke contingency cell counts (α_G) were saved from each ADAMS bootstrap sample, yielding a distribution of 10,000 values for each vector of contingency cell counts for each cognitive impairment group. A random vector of counts was sampled as the prior count at each iteration b of the MCMC chain for synthetic data generation.

Priors on continuous variables were specified using group- and contingency cell-specific estimates of means (μ_{d_G}) and group-specific estimates of covariances matrices (Σ_G) from each ADAMS bootstrap sample, yielding a distribution of 10,000 values for each parameter. A

random mean-covariance matrix pair was sampled and used as the prior at each iteration *b* of the MCMC chain for synthetic data generation.

3.4.2: Results

3.4.2.1: Prior Predictive Checks

Prior predictive checks were performed for distributions of contingency cell counts and continuous variables stratified by ADAMS cognitive impairment classes in both the training and hold-out samples. For both samples, 1000 synthetic ADAMS datasets were generated by drawing from prior distributions only. Prior predictive checks for contingency cell counts and normalized MMSE are presented in **Figure 3.4** and **Figure 3.5**, respectively, for the ADAMS training sample. Prior predictive checks of other continuous variables included in analyses (**Table 3.1**) were similarly satisfactory and code for creating those figures is provided in the associated GitHub repository.

Prior predictive distributions for the 1000 synthetic contingency cell counts were centered around true count as desired and expected since priors were based on bootstrap sampling of the ADAMS training sample. By nature of resampling the data, any observed 0 count cell remained 0 in prior predictive distributions (Hispanic + Stroke group in Unimpaired, MCI, and Other impairment groups (**Figure 3.4(a), Figure 3.4(b), Figure 3.4(d)**).

Prior predictive distributions of normalized MMSE were slightly wider (more variable) than observed distributions, ensuring that the full range of values was captured by features encoded in the priors. **Figure 3.5** shows prior predictive distributions of normalized MMSE overlayed on observed normalized MMSE in ADAMS training data for one synthetic dataset;

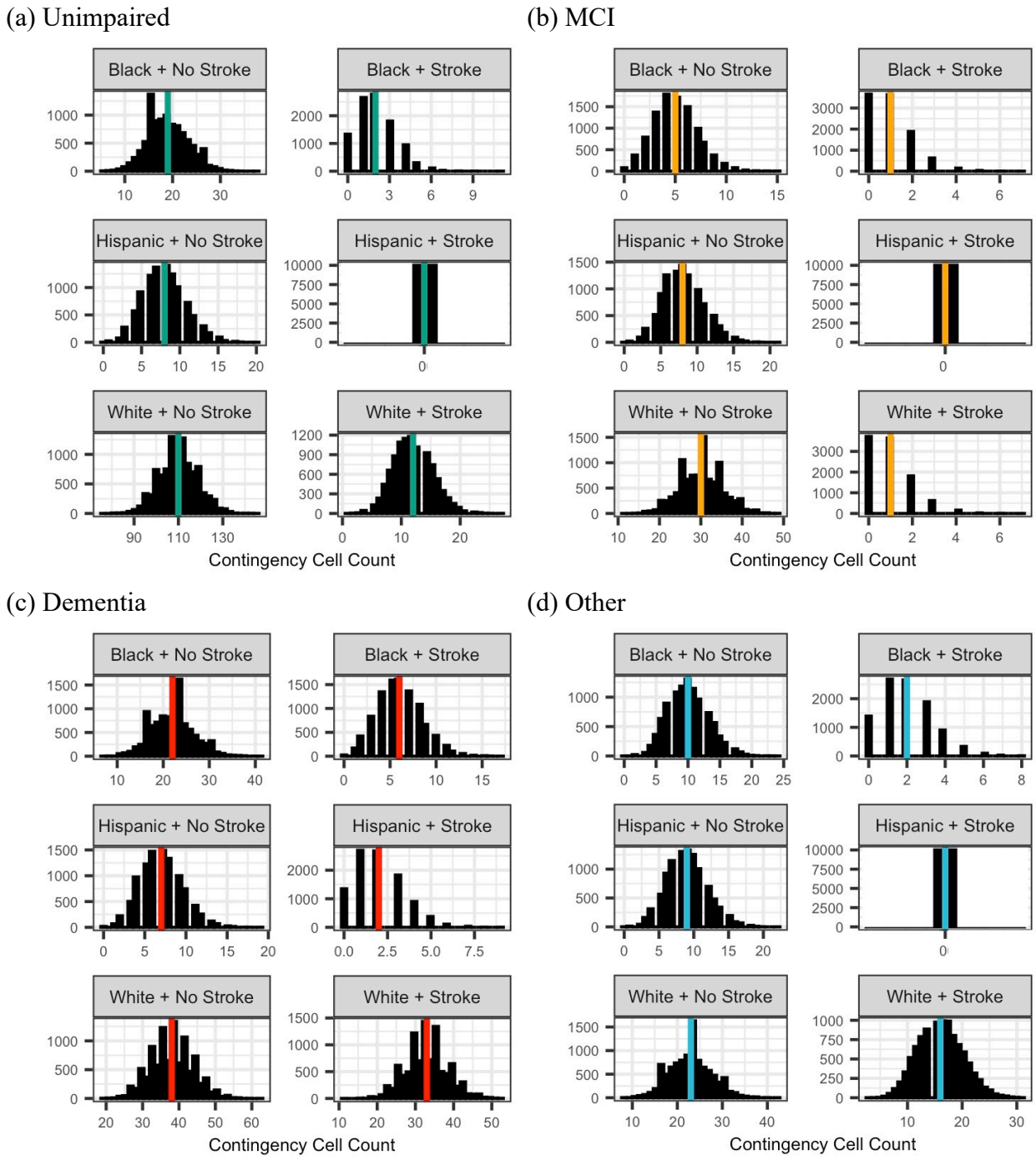


Figure 3.4: Prior predictive distributions of contingency cell counts based on 1000 synthetic ADAMS training datasets by impairment group: (a) Unimpaired, (b) MCI, (c) Dementia, (d) Other. Colored vertical lines in each panel represent observed cognitive impairment group-specific contingency cell counts in the ADAMS training sample.

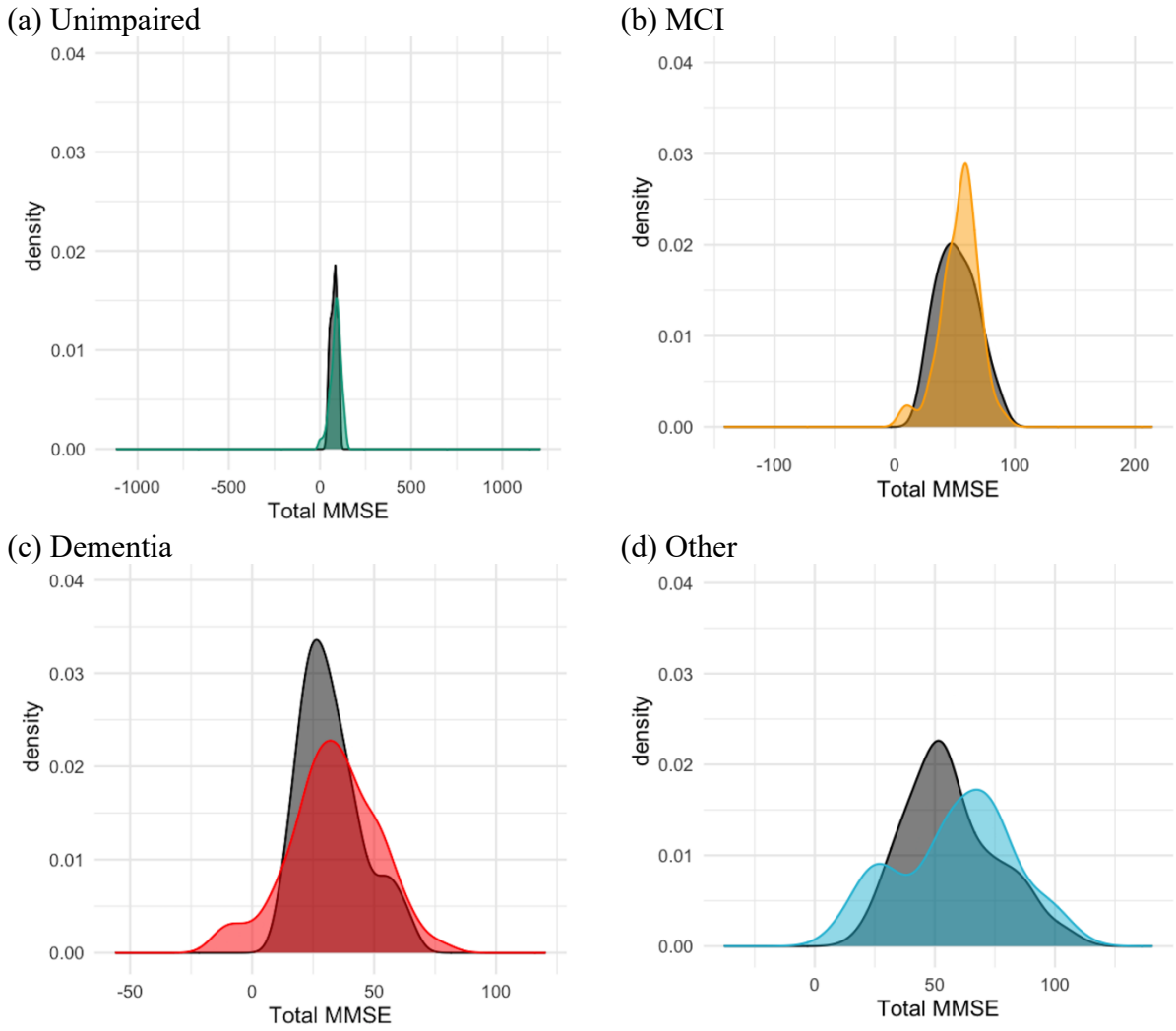


Figure 3.5: Prior predictive distributions of normalized MMSE (colored densities) based on 1000 synthetic ADAMS training datasets by impairment group: (a) Unimpaired, (b) MCI, (c) Dementia, (d) Other. Gray densities are observed impairment group-specific distributions of MMSE in the ADAMS training sample.

an animated gif cycling through all 1000 synthetic datasets was used to determine whether there was enough variability across the 1000 synthetic datasets. Prior predictive checks for the ADAMS hold-out sample were similar. Code to produce .gif files is available on the associated GitHub repository.

3.4.2.2: Assessing model convergence

MCMC chains were produced for each parameter in this analysis, but the primary chains monitored for convergence across runs were impairment class proportions and impairment group-specific variances for continuous variables. All chains converged in both training and holdout samples. MCMC chains for the ADAMS training sample are shown in **Figure 3.6**, and code for the MCMC chains for the ADAMS hold-out sample are provided in the associated GitHub repository.

To assess model stability, MCMC chains were initiated in different parts of the parameter space of proportions of cognitive impairment class membership. Five chains were monitored: (1) a "warm start" chain with cognitive impairment class proportions close to observed ADAMS proportions (40% Unimpaired, 10% MCI, 30% Dementia, 20% Other), (2) a "random" chain with equal proportions for all impairment classes (25% Unimpaired, 25% MCI, 25% Dementia, 25% Other), (3) "mostly dementia" chain where proportion of dementia dominated other impairment classes (10% Unimpaired, 30% MCI, 40% Dementia, 20% Other), (4) "mostly MCI" chain where proportion of MCI dominated other impairment classes (10% Unimpaired, 40% MCI, 20% Dementia, 30% Other), and (5) "mostly impaired" chain where only 5% of participants were initiated in the unimpaired class (5% Unimpaired, 25% MCI, 55% Dementia, 15% Other). Analyses of multiple chains of impairment class proportions showed good convergence and mixing, demonstrating that the model was stable regardless of the starting point (**Figure 3.7**).

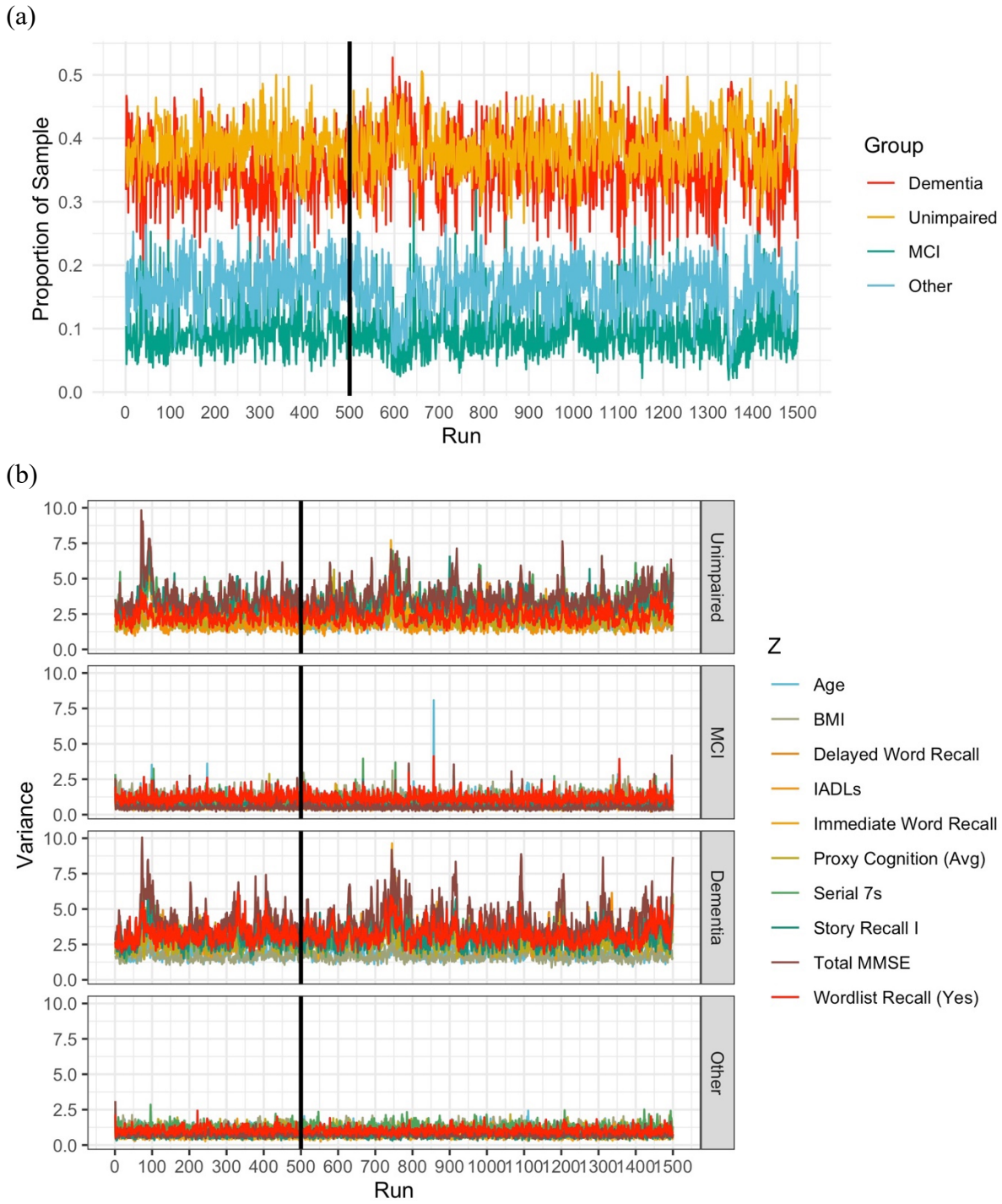


Figure 3.6: MCMC chains of (a) proportions of cognitive impairment class membership and (b) impairment group-specific variances of continuous variables based on 1000 synthetic ADAMS training datasets. Black vertical lines mark the end of the burn-in period (500 runs).

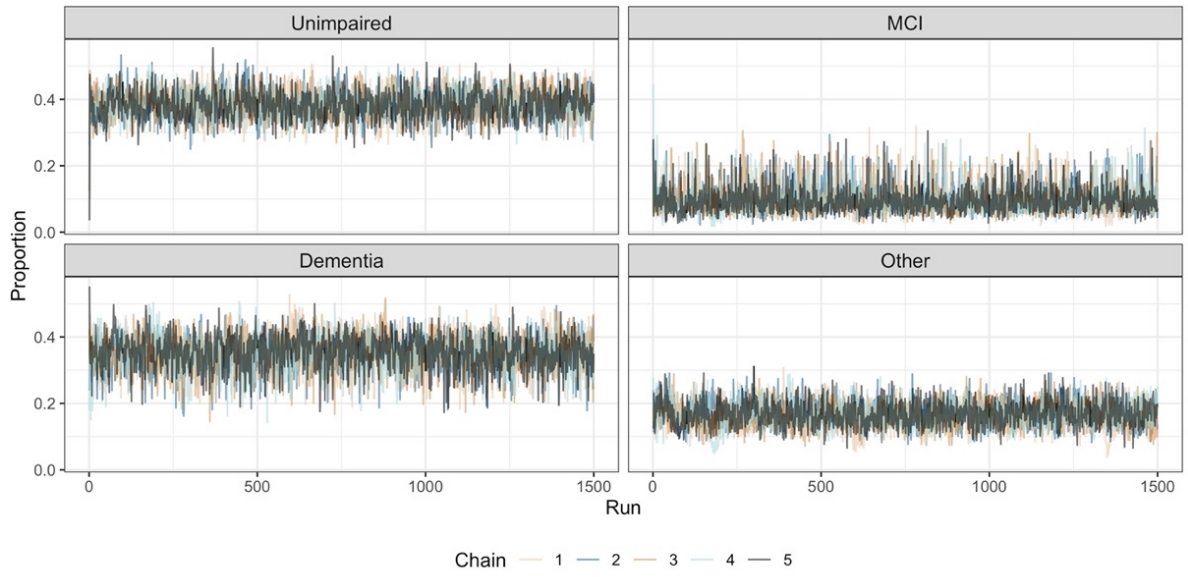


Figure 3.7: Five MCMC chains of proportions of cognitive impairment class membership. Each chain was initiated at different points in the parameter space. Chain 1 (warm start): 40% Unimpaired, 10% MCI, 30% Dementia, 20% Other; Chain 2 (random chain): 25% Unimpaired, 25% MCI, 25% Dementia, 25% Other; Chain 3 (mostly dementia): 10% Unimpaired, 30% MCI, 40% Dementia, 20% Other; Chain 4 (mostly MCI): 10% Unimpaired, 40% MCI, 20% Dementia, 30% Other; Chain 5 (mostly impaired): 5% Unimpaired, 25% MCI, 55% Dementia, 15% Other.

3.4.2.3: Posterior predictive checks

Posterior distributions of contingency cell counts and median and skew for continuous variables were assessed for both ADAMS training and hold-out samples. All posterior statistics were stratified by predicted cognitive impairment class. Posterior predictive distributions for cell counts and median and skew of normalized MMSE in the ADAMS training sample are presented in **Figure 3.8 - Figure 3.10**. Posterior predictive distributions for the remaining continuous variables were similar and code for producing them is provided in the associated GitHub repository.

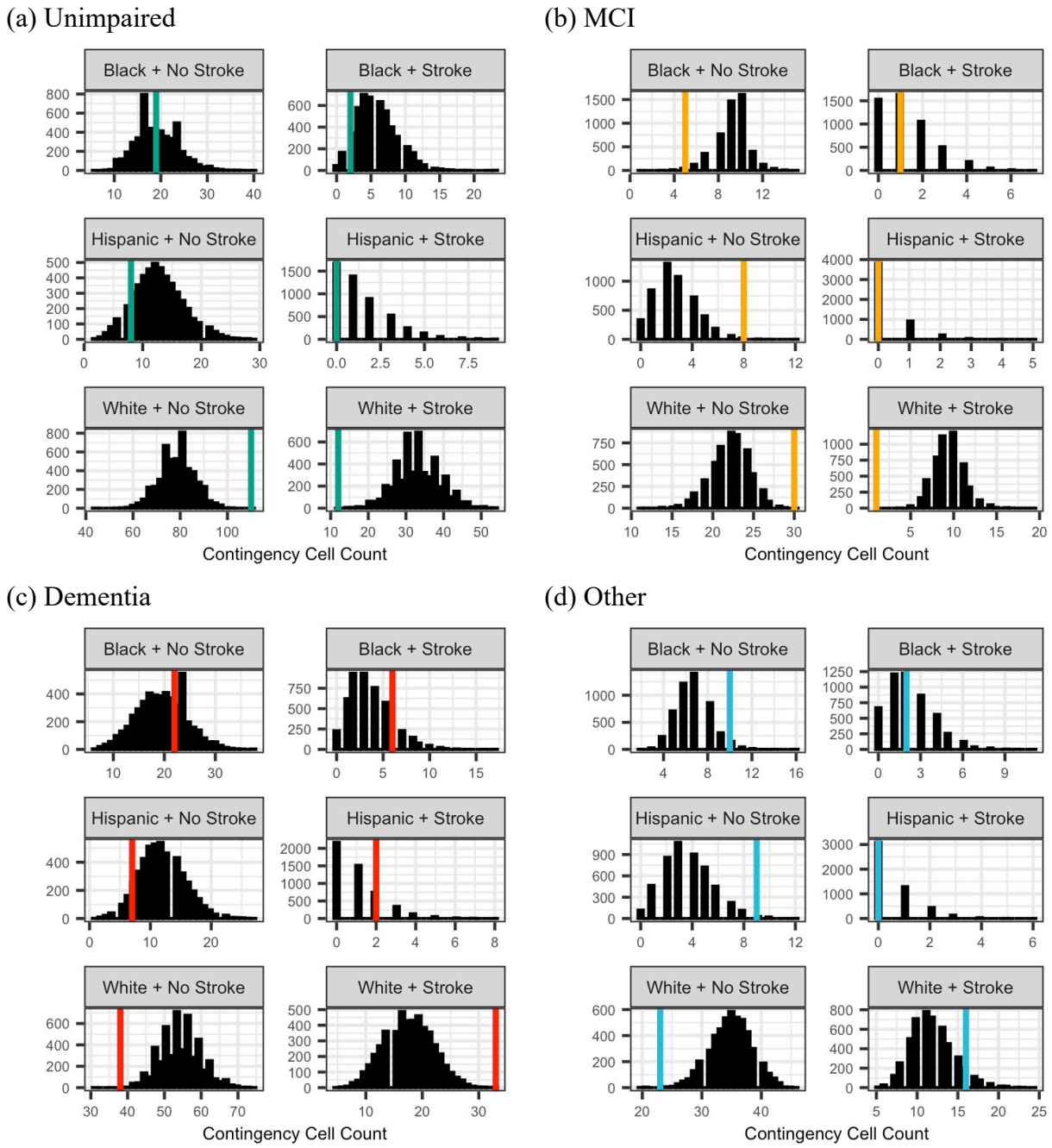


Figure 3.8: Posterior predictive distributions of contingency cell counts based on 1000 synthetic ADAMS training datasets by cognitive impairment group: (a) Unimpaired, (b) MCI, (c) Dementia, (d) Other. Colored vertical lines in each panel represent observed impairment group-specific contingency cell count in the ADAMS training sample.

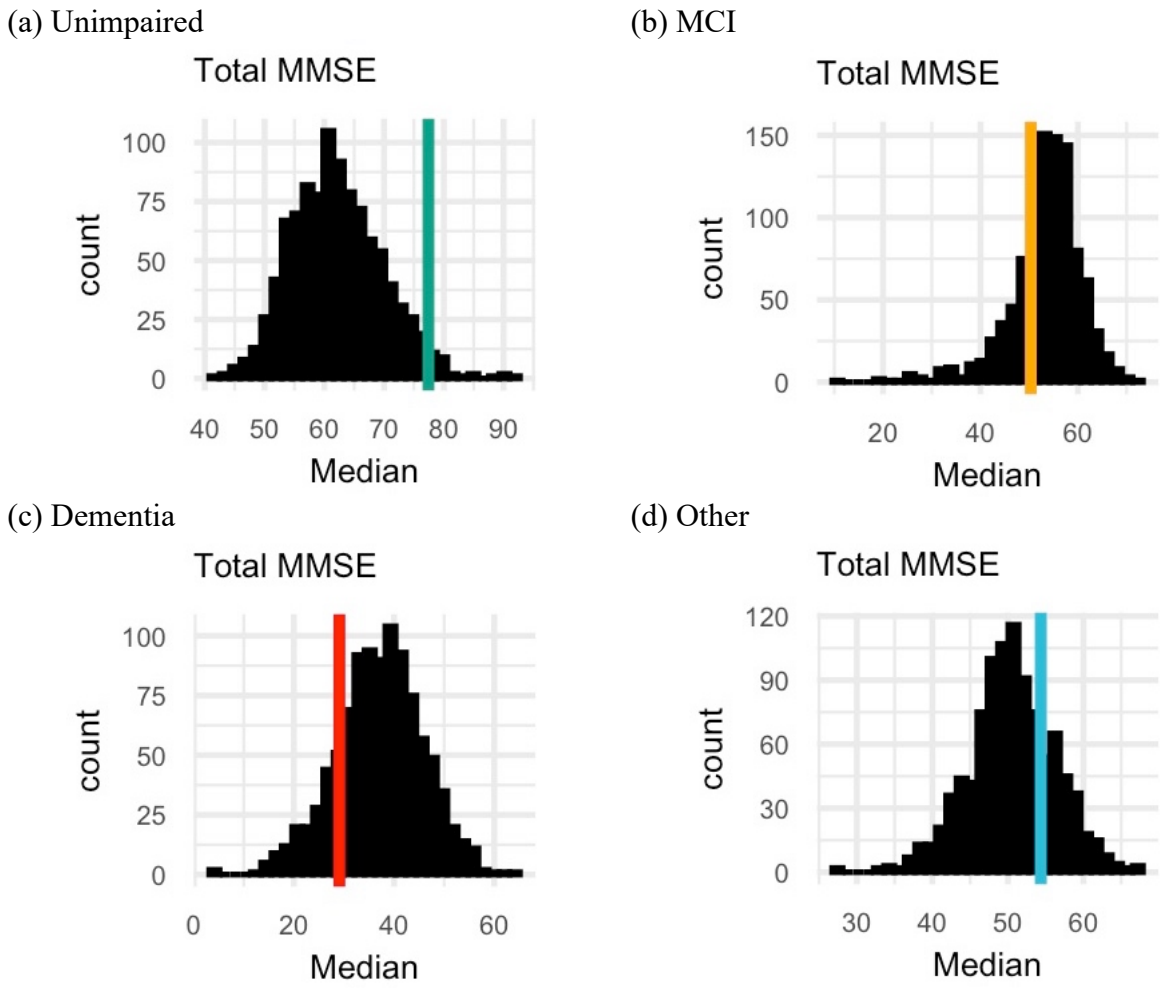


Figure 3.9: Posterior predictive distributions of median MMSE based on 1000 synthetic ADAMS training datasets by impairment group: (a) Unimpaired, (b) MCI, (c) Dementia, (d) Other. Colored vertical lines denote observed impairment group-specific medians of MMSE in the ADAMS training sample.

Posterior cell counts for Black and Hispanic participants were satisfactory across impairment groups while those for White participants were less so since posterior distributions for White participants only included observed counts in the tails of the distributions.

Posterior distributions of continuous variable medians were roughly centered around observed medians in the ADAMS data. Posterior distributions of skewness of synthetic continuous variables were centered around 0, demonstrating that synthetic variables were more symmetric than observed continuous variables which for the most part had non-zero skewness.

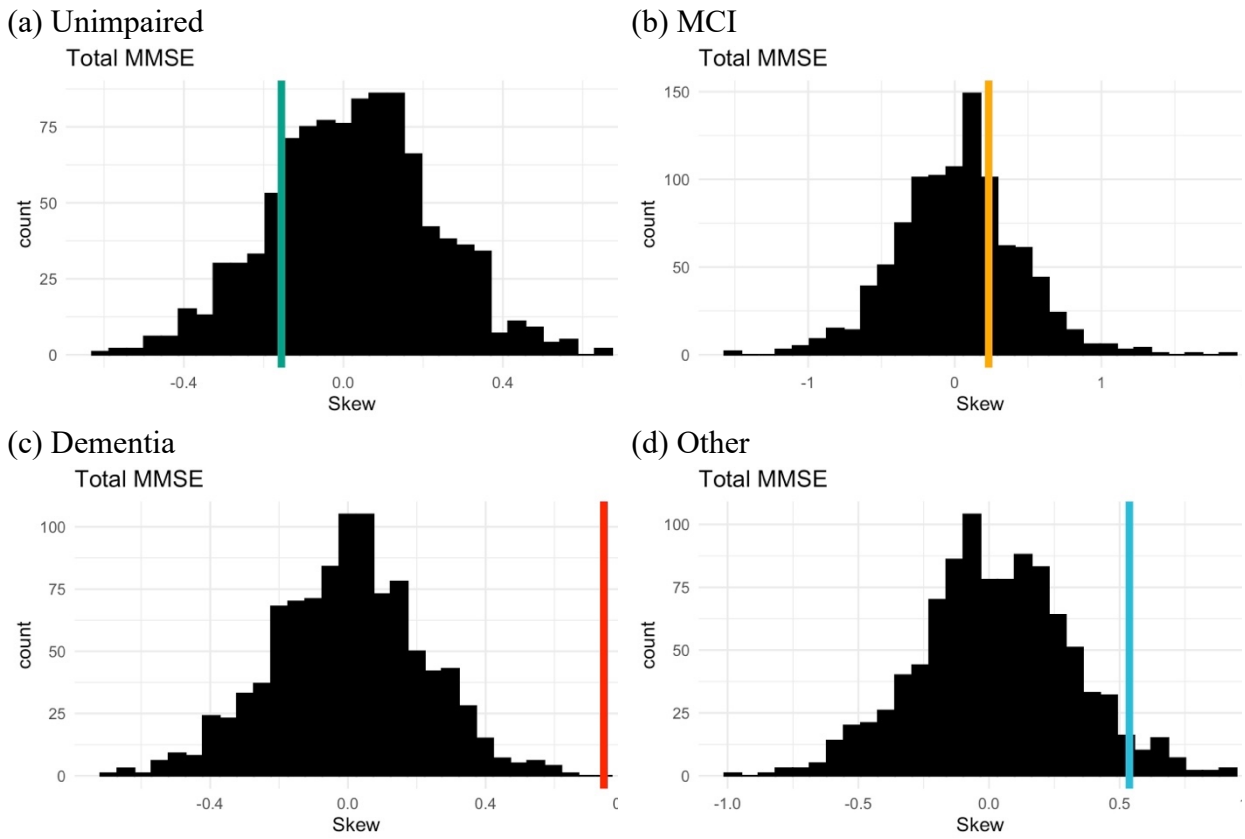


Figure 3.10: Posterior predictive distributions of MMSE skew based on 1000 synthetic ADAMS training datasets by impairment group: (a) Unimpaired, (b) MCI, (c) Dementia, (d) Other. Colored vertical lines denote observed impairment group-specific normalized MMSE skew in the ADAMS training sample.

This is expected since synthetic variables are mixtures of normal distributions which can, but are not guaranteed, to produce skewed distributions. Posterior predictive distributions for the ADAMS hold-out sample were similar, and code for constructing them is provided in the associated GitHub repository.

3.4.2.4: Algorithmic dementia classification

Figure 3.11 shows 95% credible intervals of participant counts in each cognitive impairment class across 1000 synthetic ADAMS training and hold-out datasets. In the ADAMS training data, every credible interval captured the observed ADAMS cognitive impairment class count; the largest discrepancy in mean count was in the dementia group where the model overestimated the

count by just 25 people on average. In the ADAMS hold-out set, again, every credible interval captured the observed ADAMS impairment class count, and in the hold-out sample, the largest discrepancy was in the MCI group where the model underestimated the count by just 5 people on average.

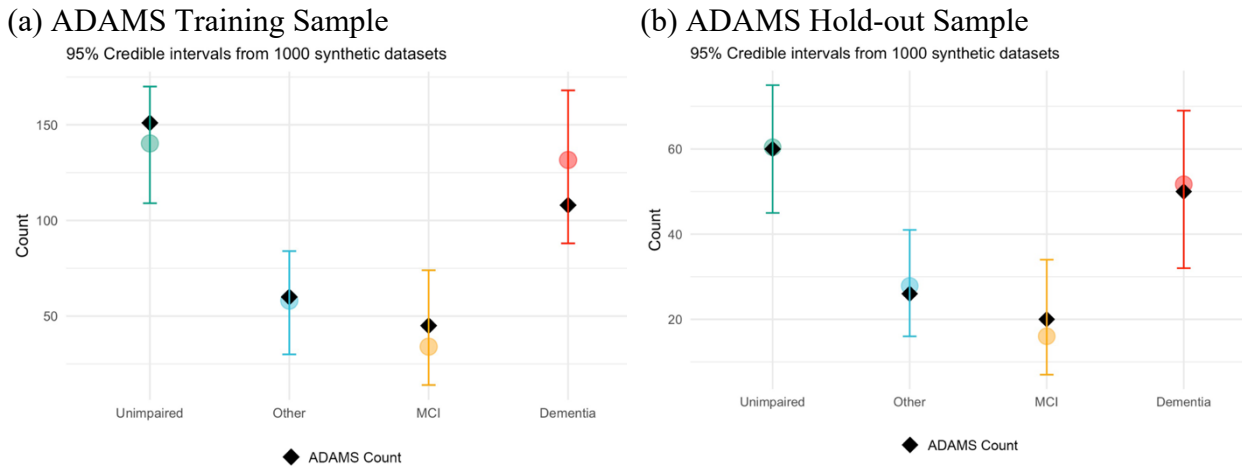


Figure 3.11: Algorithmic dementia classification in the (a) ADAMS training sample and (b) ADAMS hold-out sample. Colored intervals are 95% interval estimates of participant counts within each cognitive impairment group across 1000 synthetic datasets. Black diamonds are group-specific observed counts in each dataset.

3.5: Discussion

Using the Bayesian latent class mixture model, I was able to create synthetic versions of the ADAMS data that preserved important features of the real ADAMS study and resulted in accurate predicted cognitive impairment classes in both the training and hold-out samples. I observed some lack of fit for categorical variables, specifically distributions of White participants in stroke history x cognitive impairment class were not well replicated. This is somewhat counterintuitive since White participants are well-represent in ADAMS, so we would expect estimates to be more accurate for them. The better fit of predicted categorical variable distributions for groups of smaller sample sizes, however, is a feature of multinomial

distributions, which are used to draw categorical data in this model. Groups of smaller sizes make up a smaller proportion of the sample and will thus have less variability in their estimates.

A notable desirable property of this model is that cell counts that were 0 in the prior (i.e., Hispanic participants with stroke history in all cognitive impairment groups other than dementia) did not remain zero in the posterior. Since the 0 cell counts in the prior were random, not structural, realistic replicates of the dataset would be expected to have small, non-zero cell counts.

Overall, the results presented in the illustrative example demonstrate promising performance for the Bayesian latent class mixture modeling framework for algorithmic dementia classification. The lack of fit for some categorical variables and the lack of skew in synthetic continuous variables did not impact the prediction accuracy for cognitive impairment classes in the illustrative example but does call for more careful consideration of potential impacts of using these synthetic datasets in subsequent analyses. I conducted several simulation studies to better understand the statistical properties of this model in datasets of different sizes and using prior data obtained through different sampling frames, the performance of this algorithmic dementia classification procedure for participants of different race/ethnicities, and the use of synthetic datasets generated from the Bayesian latent class mixture model in downstream analyses. These studies are described in full detail in **Chapter 4** and **Chapter 5**.

Chapter 4 Simulation Study: Bayesian Latent Class Mixture Model using Prior Information from the Aging, Demographics, and Memory Study

Chapter 3 presented details for a Bayesian latent class mixture modeling framework for algorithmic dementia classification that incorporates information commonly used in algorithmic dementia classification algorithms (i.e., sociodemographic characteristics, health characteristics, general cognitive assessments) and detailed neuropsychological assessments which are crucial in clinical dementia assessments; but due to a lack of availability in large studies, are currently not used in dementia classification algorithms. Understanding relationships between detailed neuropsychological measures and participants' cognitive impairment status was an important first step to incorporating these key measures in dementia classification algorithms. By successfully creating synthetic versions of the ADAMS study using a Bayesian latent class mixture model (see **Section 3.4:**), I demonstrated how translating key measures like detailed neuropsychological assessments to population-based surveys can strengthen algorithmic dementia classification in these samples.

The ultimate goal of this project is to use the Bayesian latent class mixture modeling framework to algorithmically classify cognitive impairment status in the 2016 HCAP study which does not currently have a clinically adjudicated subset of participants (see **Section 2.1.3:**) and generalize those results to the 2016 HRS study, the population-representative study from which HCAP was sampled (see **Section 2.1.1:**). Before applying the Bayesian latent class mixture model to 2016 HCAP, I conducted a simulation study to assess the statistical properties

of the proposed algorithmic dementia classification framework and the impact of using the generated synthetic datasets in downstream analyses. I used a synthetic superpopulation in the simulation study that was based on 2016 HRS and 2016 HCAP data with cognitive impairment classification based on information from the ADAMS study. I was interested in model performance at two levels: (1) "Can the model accurately classify cognitive impairment for participants in an HCAP-type study with available detailed neuropsychological assessments?" and (2) "Do algorithmic classifications in the HCAP-type study yield valid inferences after they are generalized to a population-representative HRS-type study?" I anchored the simulation study in a question dementia researchers may be interested in answering using HRS: estimating standardized race-specific dementia prevalences and racial/ethnic differences in prevalent dementia in the 2016 HRS study.

The illustrative example in **Chapter 3** using ADAMS data revealed some issues with model fit and included simplification steps that are adjusted in this chapter. This chapter is organized as follows: I start by outlining broad simulation study steps that provide a high-level overview of the study; then, I move into details about ADAMS, HCAP, and HRS dataset preparation including how I addressed missing data in each study and performed variable selection for important predictors in the Bayesian latent class mixture model; next, I provide details on creating the superpopulation for the simulation study. The last half of the chapter shows model fit diagnostics for the Bayesian latent class mixture model and simulation study results.

4.1: Simulation study outline

Simulation studies are a valuable tool for assessing properties of an analytic method under different controlled conditions (e.g., different sample sizes or different data distributions). In simulation studies, we use a dataset with known truth against which we compare estimated values from our model. By varying conditions in these studies, we can quantify performance of the analytic method as it relates to properties of the data. There are several methods for creating a dataset with known truth for simulation studies. Some simulation studies use synthetic datasets with researcher-specified data-generating processes and thus researcher-specified truth (Shaw et al. 2021; Huque et al. 2018; Drechsler 2015; Grund, Lüdtke, and Robitzsch 2017; Lüdtke, Robitzsch, and Grund 2017; Hayes-Larson et al. 2022; Hron, Templ, and Filzmoser 2010). Parameters in these studies are usually based on or calibrated to real-world data so that synthetic datasets are more realistic. Other simulation studies have opted to use datasets derived from extant studies and consider estimates in those derived datasets as the truth (Dahal et al. 2019; L. Tang et al. 2005; Cao et al. 2022; Shaw et al. 2022). Synthetic datasets constructed using researcher-specified parameters represent simplified versions of real-world phenomenon; it is up to the discretion of the researcher whether the simplified data provides a useful assessment of their model. Studies that opt for datasets derived from extant studies typically do so to preserve complicated joint distributions that would be difficult to model in completely simulated data. This is especially useful in studies of analytic methods that require many variables.

Since I am modeling many variables in Bayesian latent class mixture model whose joint distributions would be difficult to replicate but are an important facet of the model assessment, I chose to preserve these relationships by bootstrapping an extant dataset to create a

superpopulation of n=1,000,000 observations. Thus, estimated values of interest in the superpopulation were considered the truth in this simulation.

This simulation study evaluates the role of sample size for population-representative studies (HRS) and subsamples with detailed neuropsychological measures (HCAP) when using information from ADAMS to specify priors in the Bayesian latent class mixture model. An outline of the simulation study is presented in **Figure 4.1**.

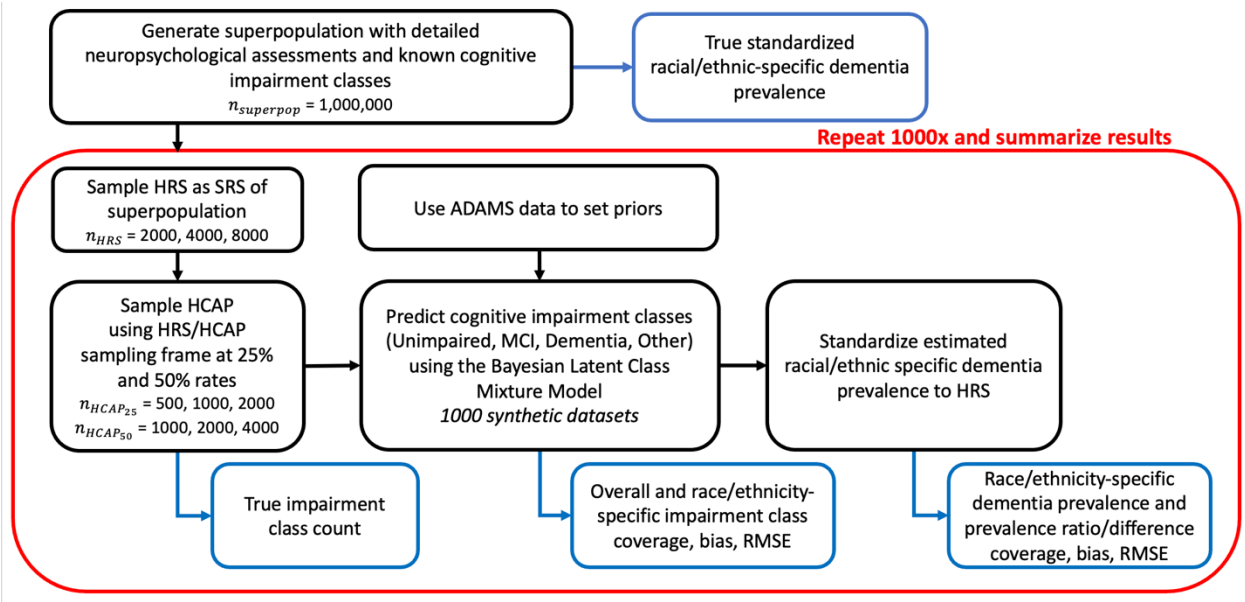


Figure 4.1: Simulation study flow diagram. Black boxes indicate analysis steps and blue boxes indicate calculation steps. The red box denotes the sequence of steps that are repeated 1000 times in the simulation study.

First, I generated a superpopulation of $n_{superpop}$=1,000,000 observations with detailed neuropsychological assessments and known impairment class. Specific methods for creating the superpopulation are discussed in **Section 4.3:**. Prevalence of cognitive impairment overall and by race/ethnicity and racial/ethnic differences in prevalent dementia in the superpopulation were considered the truth in the simulation study.

I evaluated simulation scenarios with HRS and HCAP sample sizes that enveloped sizes of the HRS 2016 70+ sample (n=6,313) and the HCAP 2016 70+ sample (n=2,235), which are the relevant samples for this analysis (see below for a discussion of sample selection). For each simulation run, I sampled HRS studies as a simple random sample (SRS) from the superpopulation using three different sample sizes n_{HRS} = 2000, 4000, 8000. Then, I sampled HCAP studies using the sampling frame implemented in the real HRS/HCAP study—stratified random sampling by married/partnered status (Langa et al. 2020). I sampled HCAP studies at 25% and 50% rates from each HRS study resulting in 6 simulation scenarios with the following sample sizes: $n_{HCAP_{25}}$ = 500, 1000, 2000 and $n_{HCAP_{50}}$ = 1000, 2000, 4000.

For each HCAP sample in each simulation run, I algorithmically classified cognitive impairment status and quantified estimation uncertainty by using the Bayesian latent class mixture model to create 1000 synthetic HCAP datasets and taking the mean and 95% credible interval of posterior predicted distributions of cognitive impairment class counts. Then, I age and sex-standardized race/ethnic specific estimates of dementia prevalence to HRS studies from which the HCAP study was sampled. To assess model performance related to cognitive impairment class prediction, I assessed interval estimate coverage, bias, and RMSE of predicted cognitive impairment class proportions overall and by race/ethnicity across 1000 simulation runs. To assess model performance related to inferences at the population-representative level, I assessed interval coverage, bias, and RMSE of standardized race/ethnic specific estimates of dementia prevalence in HRS; dementia prevalence ratios and differences for Black vs. White and Hispanic vs. White participants in HRS; and interval coverage, bias, and RMSE of dementia prevalence ratios and differences for Black vs. White and Hispanic vs. White participants in HRS across 1000 simulation runs.

4.2: Dataset preparation

I used extant data to create a realistic superpopulation and to specify prior distributions in the Bayesian latent class mixture model. Specifically, I used data from participants aged 70+ in the HRS 2016 wave, the ADAMS 2000 (wave A, baseline) wave, and participants aged 70+ in the HCAP 2016 (baseline) wave. Brief descriptions of each dataset are provided in **Sections 2.1.1:-2.1.3:**. Dataset-specific preparation steps for the simulation study are discussed below.

4.2.1: Health and Retirement Study

I collapsed HRS employment categories to align with ADAMS study employment categories. "Working" and "retired" categories were preserved while "unemployed and looking for work", "temporarily laid off", "disabled", "homemaker", "other", and "on leave" were collapsed into the "not working" category. I characterized alcohol use (no drinking, moderate drinking, heavy/high risk drinking) according to the 2020 Dietary Guidelines for Americans (U.S. Department of Agriculture and U.S. Department of Health and Human Services 2020).

General cognitive assessment scores on the backwards count from 20 and CERAD immediate word recall tasks were cleaned in the same way as described for the ADAMS study in the illustrative example in Chapter 3 (see **Section 3.1.1:**).

4.2.1.1: Sample selection

The HRS sample selection flow diagram is shown in **Figure 4.2**. I restricted the sample to participants aged 70+ and who were not missing race/ethnicity and who self-identified as White, Black, or Hispanic (n=13,745) to match sociodemographic characteristics of the ADAMS study. I dropped n=4 participants missing years of education, n=24 participants missing employment

status, n=45 participants missing alcohol consumption, and n=207 participants missing any information on self-reported chronic conditions (history of stroke, hypertension, diabetes, heart disease, cancer). I also dropped n=573 participants missing any HRS cognitive assessment. The final 2016 HRS analytic sample comprised n=6,313 participants. Relevant sample characteristics are displayed in **Table 4.1**.

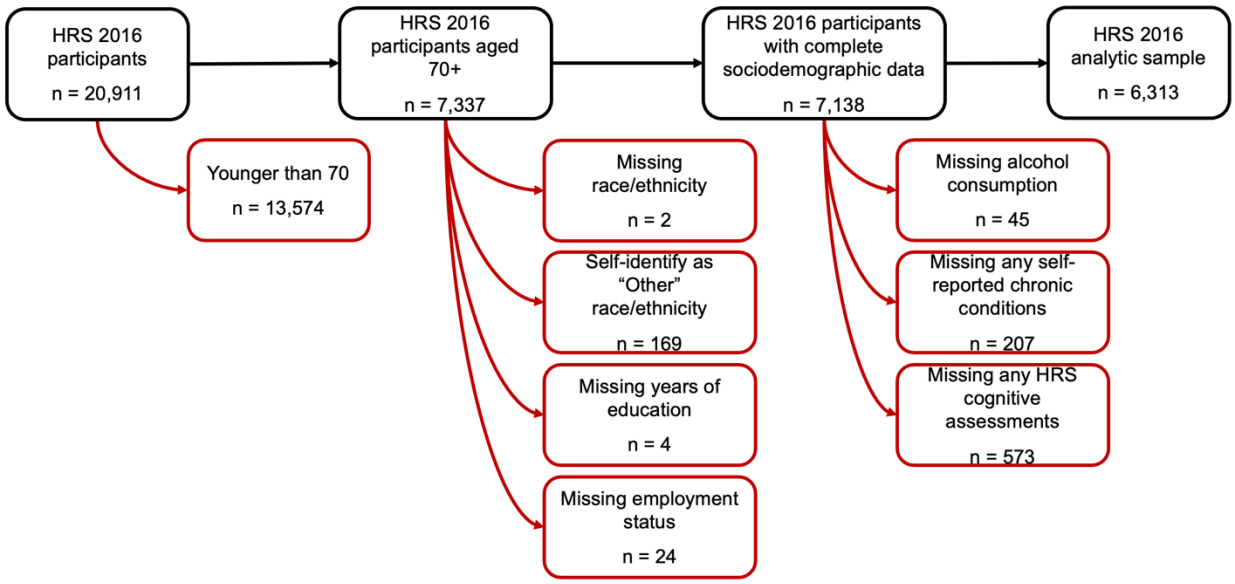


Figure 4.2: Sample selection flow diagram for the 2016 wave of HRS. Red boxes show exclusion criteria.

4.2.2: Aging, Demographics, and Memory Study

ADAMS data cleaning steps and descriptions of cognitive impairment categories were presented previously (see **Section 3.1:**).

4.2.2.1: Sample selection

The ADAMS sample selection flow diagram for the simulation study is shown in **Figure 4.3**. Data from ADAMS is used to specify priors in this simulation study. Unlike the illustrative example in Chapter 3, I did not require complete data on all measures for all ADAMS

participants and instead used multiple imputation (MI) to address missing data in the sample (details below). Participants missing all HRS general cognitive assessments and all ADAMS neuropsychological assessments were dropped from the sample (n=30). These participants were likely severely cognitively impaired, implying the data were not missing at random, which is a violation of MI assumptions and could negatively impact the quality of imputed data (see **Appendix Section B.1:**). Thus, I required ADAMS participants to complete at least one cognitive assessment so that some information on participants' cognitive function would be available for MI models. Relevant sample characteristics are displayed in **Table 4.1**.

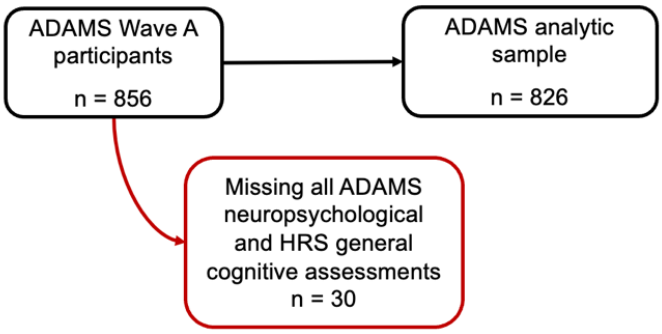


Figure 4.3: Sample selection flow diagram for the 2000 wave (wave A) of ADAMS. The red box shows exclusion criteria.

4.2.2.2: Imputing missing data using multiple imputation

There was minimal missing data in ADAMS on sociodemographic and health variables (range: 0.1% - 0.2%) and a fair amount of missing data on cognitive measures (range: 2.2% - 21.8%) (**Table 4.1**). The ADAMS sample is small, so it was important to preserve sample size by imputing missing data using multiple imputation (MI) for participants with at least one cognitive measure available. I implemented MI using fully conditional specification with predictive mean matching (PMM) through a custom implementation of the PMM option in the miceFast R

package (Nasinski 2021). See **Appendix Section B.1:** for a more detailed discussion of MI, fully conditional specification, and PMM.

I used sociodemographic variables as predictors in imputation models for all variables. In imputation models for variables collected in HRS only (marital status, health characteristics, general cognition measures), I included other HRS-only variables as predictors. In imputation models for variables collected in ADAMS that were also collected in HRS (general cognitive assessments and self-rated memory), predictors included corresponding measures collected in HRS and other measures collected in ADAMS including clinically adjudicated cognitive impairment group. In imputation models for ADAMS-only measures (detailed neuropsychological assessments), I only included other measures collected in ADAMS and clinically adjudicated impairment group as predictors.

Since the maximum proportion of missing data for a variable in ADAMS was 21.8%, I followed recommendations in Bodner (2008) and White, Royston, and Wood (2011) to set the number of imputations to exceed the amount of missing data and used 25 imputations. I monitored traceplots of imputed values and their standard deviations for each imputed variable and chose 15 iterations per imputation.

An updated table of sample characteristics with ADAMS imputed values is displayed in **Appendix Table E.1**. Comparing **Appendix Table E.1** to **Table 4.1**, we see those who were missing data in ADAMS were likely in slightly poorer health and would have performed slightly worse on cognitive assessments compared to ADAMS participants without missing data. Missing participants were imputed as more likely to be retired, less likely to be married/partnered, more likely to have heart disease or hypertension and more likely to consume alcohol. Average imputed scores for general cognitive assessments and detailed neuropsychological assessments

measured on a continuous scale (i.e., items other than correct yes/no items) were slightly lower than observed scores.

4.2.2.3: Variable selection

As discussed in Chapter 3, an important first step in building the Bayesian latent class mixture model was determining which variables were important for predicting cognitive impairment class membership. After imputing ADAMS data using MI, the pool of potential variables was much larger than in the illustrative example in Chapter 3. I had previously hand-selected variables to include using forwards selection and a p-value cutoff, but I needed a more time-efficient, systematic, and objective way of selecting variables for inclusion from this larger pool of variables.

Least Absolute Shrinkage and Selection Operator (LASSO) regression is commonly used for variable selection in problems with many covariates and is available in several software packages. LASSO shrinks coefficients towards zero using a penalty term that can be user specified or algorithmically chosen via cross-validation. See Tibshirani (1996) for an introduction to LASSO regression for variable selection.

I implemented LASSO regression using the R package glmnet (Friedman, Hastie, and Tibshirani 2010). I included all sociodemographic variables, health characteristics, HRS general cognitive assessments, and ADAMS detailed neuropsychological assessments in the pool of variables (38 potential variables). The importance of variables and their magnitude of association likely differ by cognitive impairment class. Instead of fitting one multinomial outcome model, I fit a different model for each impairment class (Unimpaired, MCI, Dementia, Other) to accommodate different subsets of predictors for each model. Let $G_i, i = 1, \ldots, 826$, denote the

ADAMS adjudicated cognitive impairment class (group) for individual i in the ADAMS training sample,

$$G_i = \begin{cases} 1 & \text{if participant } i \text{ was classified as Unimpaired} \\ 2 & \text{if participant } i \text{ was classified as having Other impairment} \\ 3 & \text{if participant } i \text{ was classified as having MCI} \\ 4 & \text{if participant } i \text{ was classified as having Dementia} \end{cases} \quad (4.1)$$

Letting X denote the vector of candidate predictor variables including an intercept and $\boldsymbol{\gamma_1}, \boldsymbol{\gamma_2}, \boldsymbol{\gamma_3}, \boldsymbol{\gamma_4}$ be vectors of regression coefficients, the following logistic regression models were fit in the ADAMS sample using LASSO:

$$logit(P(G = 1|X)) = \boldsymbol{\gamma_1} X \quad (4.2)$$

$$logit(P(\mathrm{G} = 2|X)) = \boldsymbol{\gamma_2} X \quad (4.3)$$

$$logit(P(G = 3\,|X)) = \boldsymbol{\gamma_3} X \quad (4.4)$$

$$logit(P(G = 4\,|X)) = \boldsymbol{\gamma_4} X \quad (4.5)$$

I multiply-imputed ADAMS prior to the variable selection step, thus I had 25 versions of the ADAMS sample. I followed recommendations in Du et al., (2022) and stacked imputed datasets, weighting observations by the inverse of the number of imputed datasets when specifying the LASSO model. I used default settings in the glmnet function and used the cross-validation option to select the best shrinkage parameter. To ensure stability of the LASSO results, I ran the cross-validation function 1000 times and used the median lambda from those runs as my shrinkage parameter. There was minimal shrinkage from the LASSO model, so I rounded model coefficients to two decimal places and dropped any variables that were rounded to zero as a result. There was still minimal shrinkage after this additional rounding step; the most shrinkage was in the model for the Other cognitive impairment class which had four predictors removed. **Appendix Table E.2** shows variables selected for each impairment class model.

Sample characteristics presented in **Table 4.1** include variables selected for at least one of the models.

4.2.3: Harmonized Cognitive Assessment Protocol

HCAP employment categories and alcohol use were cleaned and derived in the same way as described in **Section 4.2.1:** for the HRS sample. Scores on neuropsychological and general cognitive assessments were cleaned in the same way described for the ADAMS sample in the illustrative example in Chapter 3 (see **Section 3.1.1:**).

4.2.3.1: Sample selection

The HCAP sample selection flow diagram is shown in **Figure 4.4**. I restricted the sample to participants aged 70+ and who self-identified as White, Black, or Hispanic to match sociodemographic characteristics of the ADAMS sample; this eliminated 945 participants. I also dropped participants missing information on key sociodemographic and health variables: n=1 participant missing years of education, n=1 participant missing employment status, n=15 participants missing alcohol consumption, and n=73 participants missing any information on self-reported chronic conditions (history of stroke, hypertension, diabetes, heart disease, cancer).

Because HRS total cognition was an important variable used to impute missing neuropsychological measures in HCAP (details below), I removed participants missing HRS total cognition scores from the sample (n=163) resulting in a final 2016 HCAP analytic sample with n=2,298 participants. Sample characteristics are displayed in **Table 4.1**.

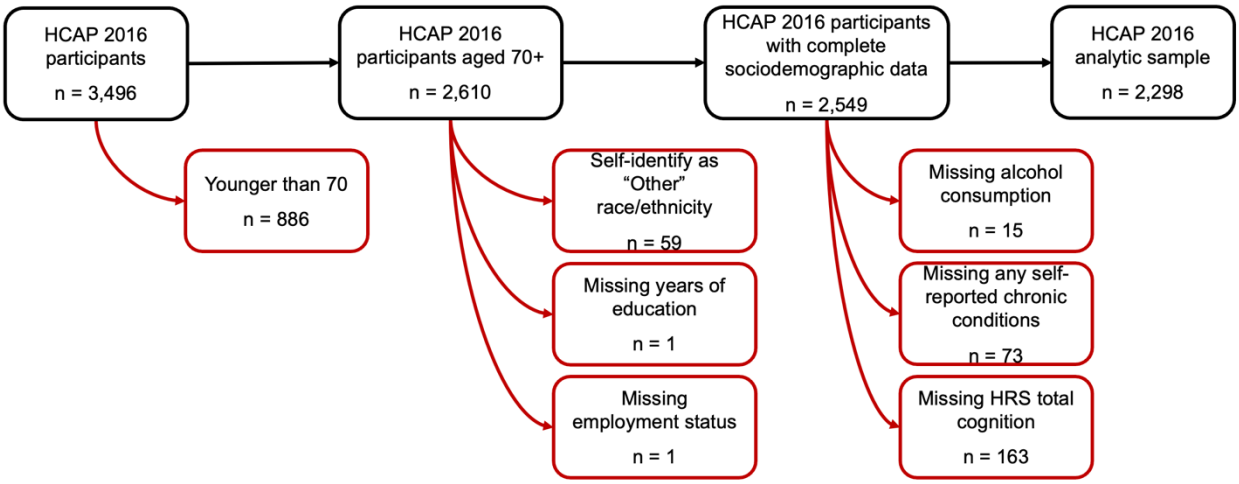


Figure 4.4: Sample selection flow diagram for the 2016 wave of HCAP. Red boxes show exclusion criteria.

4.2.3.2: Imputing missing neuropsychological data using hotdeck imputation

All neuropsychological measures are missing for HCAP respondents who completed their interviews via proxy (n=129 for HCAP 70+) and subsets of the neuropsychological exam are missing for participants who were likely very impaired and could not complete all the items. Participants with proxy-only interviews are likely more cognitively impaired than those who completed the interview as self-responders, thus dropping proxy-only participants would lead to a complete-case sample over-selected for less impaired participants.

Though MI is generally preferred over single imputation, I used a single-imputation method for the HCAP neuropsychological measures to simplify simulation studies (see **Chapter 6** for an illustration of how to incorporate multiple imputation to the Bayesian latent class mixture model). I used single hotdeck imputation with donor pools defined by stratified matching. Briefly, hotdeck imputation uses a pool of donors with observed values in the dataset from which to randomly draw values for variables that need to be imputed. Thus, hotdecking

always produces imputations within the range of observed values. See **Appendix Section B.2:** for a more comprehensive discussion of hotdeck imputation.

The quality of imputed datasets depends on the quality of the donor pool. There are several strategies for creating a pool of donors. I created pools tailored to the variable being imputed by categorizing continuous variables and matching observations in HCAP using combinations of the most relevant sociodemographic characteristics, HRS general cognition items, and neuropsychological assessments. Since I was imputing all cognitive assessments for proxy participants, it was important to have a measure of total cognition on which to anchor those matches. Thus, I dropped participants missing the HRS total cognition measure.

I considered the cognitive domain a neuropsychological test was designed to measure when selecting variables to match. Some neuropsychological tests are designed to capture performance in a specific cognitive domain while others provide a measure of general cognition (Harvey 2019; Reger et al. 2004). When possible, participants were matched on a measure of total cognition and other cognitive assessments designed to measure the same or relevant domain. Since I imputed missing neuropsychological measures using one hotdeck draw, imputation order was important. Variables that I wanted to include in matching for other neuropsychological measures were imputed first. For example, I imputed MMSE scores before other neuropsychological assessments because they were important for subsequent imputations; I matched HCAP participants on race/ethnicity, sex/gender, age, educational attainment, and HRS total cognition score. For downstream word recall (yes) imputation (memory domain), I used hotdeck-imputed MMSE (a measure of total cognition), and hotdeck-imputed immediate word recall and delayed word recall scores (memory domain).

A challenge with hotdeck imputation is balancing quality matches with donor pool size. It is undesirable to require matches so precise that donor pools are small because identical values would be imputed for multiple participants. It is also possible to refine matches to the point that there are no donors available (i.e., the observation missing was the only observation in the dataset with that combination of characteristics). The desire to create donor pools of at least 15 observations led to coarsening continuous variable categories and reducing the number of variables used for matching for some neuropsychological measures. Continuous variable categorization is described in **Appendix Table E.3**, and the complete list of hotdeck-imputed variables and variables used for matching can be found in **Appendix Table E.4**.

An updated table of HCAP sample characteristics incorporating hotdeck-imputed values is displayed in **Appendix Table E.1**. Comparing **Appendix Table E.1** to **Table 4.1**, we see that contrary to my original assumption, HCAP participants missing neuropsychological test measures were imputed as performing similarly to those with observed data as distributions are nearly identical between observed HCAP and imputed HCAP data.

4.3: Creating the superpopulation

I bootstrapped HRS 2016 70+ to create a superpopulation of $n_{superpop}$ =1,000,000 participants (Efron and Tibshirani 1994). Since HRS is a population-representative study, it made the most sense to base sociodemographic and health characteristics of the superpopulation on this sample. By bootstrapping HRS, I was able to preserve complex joint distributions in the data that would have been difficult to model. When I bootstrapped HRS, I included all variables from the variable selection procedure (see **Section 4.2.2.3:**) that were available in HRS 2016. I created neuropsychological measures in the superpopulation by hotdeck imputing all measures using

observed neuropsychological measures from HCAP 2016 as donors and the same matching strategy described above for hotdeck imputing the 2016 HCAP sample. Continuous variables were categorized the same was as for the HCAP 2016 sample imputation except for HRS total cognition which needed to be coarsened for the superpopulation (**Table E.3**). The complete list of superpopulation hotdeck-imputed variables and variables used for matching is presented in **Appendix Table E.5**.

Observations in the superpopulation were assigned to the Unimpaired, MCI, Dementia, or Other cognitive impairment group by predicting probabilities of class membership using coefficients from the variable selection procedure with models **(4.1)-(4.5)** and assigning the observation to Unimpaired if that was the highest predicted probability for an observation. If an observation was not assigned to the Unimpaired group, it was assigned to MCI, Dementia, or Other using a draw from a categorical distribution (generalization of a Bernoulli distribution) with probabilities of success equal to rescaled predicted probabilities from the cognitive impairment class models.

It was important to create a realistic superpopulation to have the best chance of this simulation study informing real data applications. I conducted quality checks of the superpopulation to assess whether established associations between sociodemographic characteristics, health characteristics, and dementia were also replicated in the superpopulation. Proportions of observations in each cognitive impairment classes in the superpopulation were reasonable but proportions in the Dementia group were slightly higher and proportions in the MCI group slightly lower compared to estimates from population-based studies: 37.3% Unimpaired, 16.4% MCI, 26.0% Dementia, and 20.4% Other (Rajan et al. 2021; Jennifer J Manly et al. 2022). There was slightly more dementia among women in the superpopulation

(women: 26.9%, men: 24.5%). There was more dementia among racial/ethnic minorities after standardizing by sex and age: White: 24.1%, Black: 33.1%, Hispanic: 30.2%. In line with literature, the prevalence of dementia increased with each additional year of age (PR=1.03), history of stroke history vs. no history of stroke (PR=1.72), history of diabetes vs. no history of diabetes (PR=1.03) and decreased with each additional year of educational attainment (PR=0.95) (Alzheimer's Association 2022). Sample characteristics for relevant waves of each dataset and for the superpopulation are provided in **Table 4.1**.

Since the superpopulation was generated by bootstrapping the HRS sample, distributions of sociodemographic characteristics, health characteristics, and general cognitive assessments are identical between HRS and the superpopulation. The HCAP study was designed as a stratified SRS from HRS based on marital status (Langa et al. 2020). Since marital status is unlikely highly correlated with other sample characteristics, we would expect distributions of sociodemographic characteristics and cognitive performance in HCAP to be nearly population-representative. Consistent with this, we observe very little difference between HCAP and HRS sample characteristics.

The ADAMS study is used as a prior in this analysis for inferences in the superpopulation. Large differences between ADAMS and the superpopulation may indicate incompatibility between the prior and observed data and negatively impact the quality of inferences. ADAMS is 3 years older than the superpopulation on average, is slightly more balanced on sex/gender due to the ADAMS design (Langa et al. 2005), has slightly more Black participants, fewer married/partnered participants, less educational attainment on average, and more retired participants compared to the superpopulation. ADAMS participants have more stroke history compared to the superpopulation and slightly more difficulty with ADLs and

IADLs but are healthier than the superpopulation when comparing other health characteristics. ADAMS participants were oversampled for impairment, thus, ADAMS had lower scores on average for HRS total cognition (5 points on average) and normalized MMSE (about 20 points on average) compared to the superpopulation.

Variable	ADAMS Baseline (2002) N = 826	HCAP 70+ Baseline (2016) N = 2,298	HRS 70+ (2016) N = 6,313	Superpopulation N = 1,000,000
Age, Mean (SD)	81.3 (7.0)	78.8 (5.9)	78.8 (6.0)	78.8 (6.0)
Female, n (%)	476 (57.6%)	1,379 (60.0%)	3,765 (59.6%)	596,718 (59.7%)
Race/Ethnicity, n (%)				
White	590 (71.4%)	1,767 (76.9%)	4,837 (76.6%)	765,663 (76.6%)
Black	153 (18.5%)	325 (14.1%)	895 (14.2%)	142,182 (14.2%)
Hispanic	83 (10.0%)	206 (9.0%)	581 (9.2%)	92,155 (9.2%)
Years of Education, Mean (SD)	10.0 (4.4)	12.6 (3.1)	12.6 (3.2)	12.6 (3.2)
Employment status, n (%)				
Working	58 (7.0%)	190 (8.3%)	548 (8.7%)	86,592 (8.7%)
Not working	61 (7.4%)	305 (13.3%)	809 (12.8%)	128,287 (12.8%)
Retired	706 (85.5%)	1,803 (78.5%)	4,956 (78.5%)	785,121 (78.5%)
Missing	1 (0.1%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Married/Partnered, n (%)	327 (39.6%)	1,236 (53.8%)	3,356 (53.2%)	531,414 (53.1%)
Missing	2 (0.2%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
BMI, Mean (SD)	25.9 (5.4)	27.8 (5.6)	27.7 (5.6)	27.7 (5.6)
Missing, n (%)	9.0 (1.1%)	0.0 (0.0%)	0.0 (0.0%)	0.0 (0.0%)
History of stroke, n (%)	151 (18.3%)	307 (13.4%)	796 (12.6%)	125,963 (12.6%)
Missing	1 (0.1%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
History of diabetes, n (%)	158 (19.1%)	681 (29.6%)	1,858 (29.4%)	294,720 (29.5%)
Missing	1 (0.1%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
History of heart disease, n (%)	266 (32.2%)	849 (36.9%)	2,306 (36.5%)	365,608 (36.6%)
Missing	1 (0.1%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
History of hypertension, n (%)	445 (53.9%)	1,676 (72.9%)	4,579 (72.5%)	726,009 (72.6%)
Missing	2 (0.2%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Current smoker, n (%)	60 (7.3%)	143 (6.2%)	407 (6.4%)	64,479 (6.4%)
Alcohol consumption, n (%)				
No drinking	681 (82.4%)	1,543 (67.1%)	4,247 (67.3%)	673,197 (67.3%)
Moderate drinking	107 (13.0%)	567 (24.7%)	1,579 (25.0%)	249,798 (25.0%)
Heavy drinking	36 (4.4%)	188 (8.2%)	487 (7.7%)	77,005 (7.7%)
Missing	2 (0.2%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
ADLs, Mean (SD)	0.9 (1.4)	0.4 (1.0)	0.4 (1.0)	0.4 (1.0)

IADLs, Mean (SD)	0.6 (1.0)	0.2 (0.5)	0.2 (0.5)	0.2 (0.5)
Immediate word recall, Mean (SD)	5.6 (2.4)	7.0 (1.8)	4.8 (1.7)	4.8 (1.7)
Missing, n (%)	35.0 (4.2%)	19.0 (0.8%)	0.0 (0.0%)	0.0 (0.0%)
Delayed word recall, Mean (SD)	3.7 (2.7)	4.9 (2.6)	3.8 (2.0)	3.8 (2.0)
Missing, n (%)	43.0 (5.2%)	28.0 (1.2%)	0.0 (0.0%)	0.0 (0.0%)
Serial 7s, Mean (SD)	2.2 (2.0)	3.3 (1.7)	3.3 (1.7)	3.3 (1.7)
Missing, n (%)	18.0 (2.2%)	0.0 (0.0%)	0.0 (0.0%)	0.0 (0.0%)
Item naming (cactus): correct, n (%)	548 (66.3%)	2,105 (91.6%)	5,781 (91.6%)	915,717 (91.6%)
Missing	180 (21.8%)	18 (0.8%)	0 (0.0%)	0 (0.0%)
Item naming (scissor): correct, n (%)	751 (90.9%)	2,254 (98.1%)	6,213 (98.4%)	984,307 (98.4%)
Missing	43 (5.2%)	18 (0.8%)	0 (0.0%)	0 (0.0%)
President naming: correct, n (%)	598 (72.4%)	2,165 (94.2%)	6,047 (95.8%)	957,739 (95.8%)
Missing	160 (19.4%)	19 (0.8%)	0 (0.0%)	0 (0.0%)
Backwards count (20): correct, n (%)	591 (71.5%)	2,099 (91.3%)	5,779 (91.5%)	915,086 (91.5%)
Missing	55 (6.7%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
HRS total cognition, Mean (SD)	16.4 (6.4)	20.8 (5.3)	20.8 (5.3)	20.8 (5.3)
Missing, n (%)	169.0 (20.5%)	0.0 (0.0%)	0.0 (0.0%)	0.0 (0.0%)
Subjective cognitive status, n (%)				
Same as 2 years ago	475 (57.5%)	1,584 (68.9%)	4,423 (70.1%)	701,102 (70.1%)
Better than 2 years ago	59 (7.1%)	46 (2.0%)	129 (2.0%)	20,412 (2.0%)
Worse than 2 years ago	249 (30.1%)	668 (29.1%)	1,761 (27.9%)	278,486 (27.8%)
Missing	43 (5.2%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Total MMSE (normalized), Mean (SD)	51.3 (25.1)	70.6 (19.1)		70.4 (19.2)
Missing, n (%)	12.0 (1.5%)	18.0 (0.8%)		0.0 (0.0%)
Animal naming, Mean (SD)	11.7 (5.6)	15.5 (6.3)		15.5 (6.2)
Missing, n (%)	36.0 (4.4%)	19.0 (0.8%)		0.0 (0.0%)
Word recall (yes), Mean (SD)	8.2 (2.3)	8.9 (1.8)		8.6 (1.8)
Missing, n (%)	52.0 (6.3%)	26.0 (1.1%)		0.0 (0.0%)
Word recall (no), Mean (SD)	9.1 (2.1)	9.6 (1.2)		9.5 (1.2)
Missing, n (%)	52.0 (6.3%)	26.0 (1.1%)		0.0 (0.0%)
Immediate story recall, Mean (SD)	13.9 (9.1)	16.5 (6.4)		14.2 (6.3)
Missing, n (%)	78.0 (9.4%)	40.0 (1.7%)		0.0 (0.0%)
Delayed story recall, Mean (SD)	9.8 (8.8)	11.8 (7.4)		10.0 (6.7)
Missing, n (%)	98.0 (11.9%)	73.0 (3.2%)		0.0 (0.0%)
Immediate constructional praxis, Mean (SD)	8.6 (2.1)	8.1 (2.3)		8.1 (2.3)
Missing, n (%)	89.0 (10.8%)	39.0 (1.7%)		0.0 (0.0%)
Delayed constructional praxis, Mean (SD)	2.9 (1.1)	5.6 (3.2)		5.5 (3.2)
Missing, n (%)	101.0 (12.2%)	42.0 (1.8%)		0.0 (0.0%)
Trails A, Mean (SD)	83.1 (67.4)	57.5 (34.2)		57.6 (33.3)
Missing, n (%)	138.0 (16.7%)	105.0 (4.6%)		0.0 (0.0%)
Impairment group, n (%)				
Unimpaired	307 (37.2%)			373,112 (37.3%)

MCI	98 (11.9%)		163,587 (16.4%)
Dementia	273 (33.1%)		259,032 (25.9%)
Other	148 (17.9%)		204,269 (20.4%)

Table 4.1: Sample characteristics for relevant waves of ADAMS, HCAP, and HRS and the superpopulation generated for the simulation study. Cells that are grayed out indicated measures that are not available for the dataset.

4.4: Specifying prior distributions

The general strategy for specifying prior distributions in this simulation study is identical to the strategies described in Chapter 3 (see **Section 3.4.1:**) However, since the ADAMS data was multiply-imputed for this simulation study, the distributions of parameters were not obtained by bootstrapping ADAMS but were instead obtained by multiply-imputing the ADAMS study 10,000 times. Further, parameters stored for latent cognitive impairment class prediction were based on models **(4.1)-(4.5)** instead of the multi-part models specified in Chapter 3. The updated overview of parameter storage steps for the priors in this simulation study is (1) impute ADAMS data using FCS with PMM and models described in **Section 4.2.2.2:** (2) store parameter estimates characterizing effects of covariates on cognitive impairment class membership, contingency cell counts, and effects of contingency cell membership on continuous covariates. (3) repeat the process 10,000 times to represent both sampling variability and estimation uncertainty in model parameters. Details for specifying prior distributions for each component of the model are described in **Section 3.4.1:**.

4.5: Pre-simulation study tuning of the Bayesian latent class mixture model

As discussed in **Section 2.3.1:** a proper Bayesian workflow involves model checking and parameter tuning via prior predictive checks, convergence diagnostic plots, and posterior

predictive checks. These steps would be infeasible to perform for each of the 1000 iterations of each simulation study scenario. Instead, I performed all Bayesian workflow steps for one iteration of each simulation scenario and set tuning parameters in the simulation study based on results from those runs. For example, to tune parameters in the simulation scenario with $n_{HRS} = 8000$ and 50% HCAP sampling proportion, I sampled one HRS study of size 8000 from the superpopulation and sampled one HCAP study of size 4000 from the HRS study. Then, I performed prior predictive checks, model diagnostic checks, and posterior predictive checks of synthetic HCAP studies of size 4000 generated from the Bayesian latent class mixture model. Tuning parameters that resulted in satisfactory model checks were used in the simulation study. Selected results for Bayesian workflow steps are presented below.

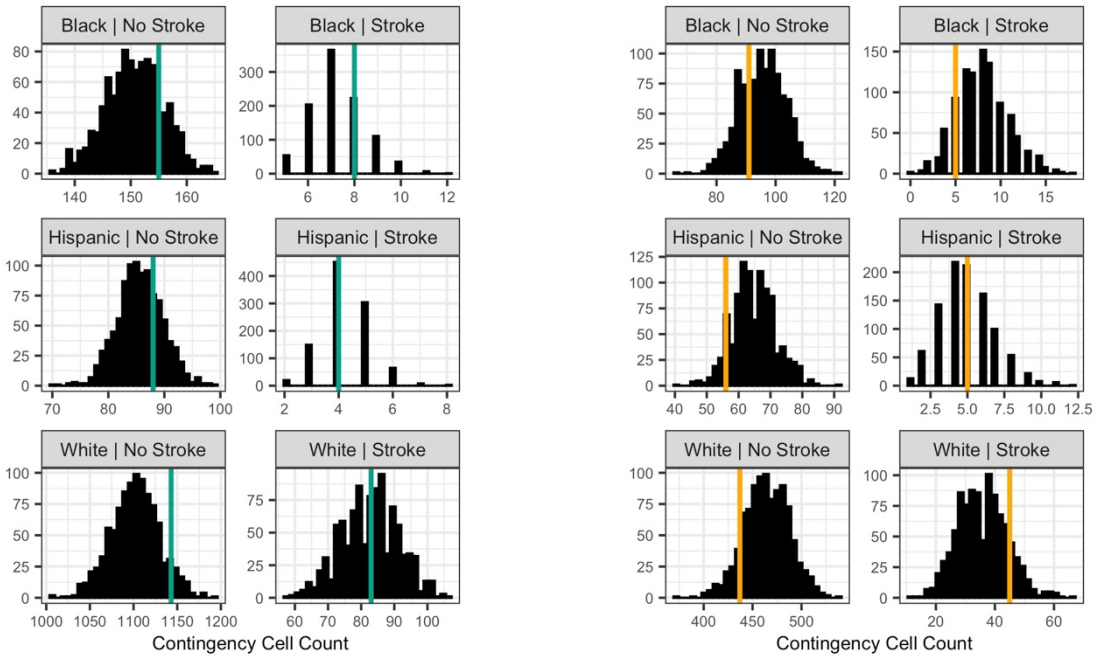
4.5.1: Prior predictive checks

I performed prior predictive checks for distributions of contingency cell counts and continuous variables stratified by cognitive impairment classes. For each simulation scenario, I generated 1000 synthetic HCAP datasets by drawing from prior distributions only. Prior predictive distributions of contingency cell counts and normalized MMSE for the scenario with $n_{HRS} = 8000$ and 50% HCAP sampling proportion are presented in **Figure 4.5** and **Figure 4.6**, respectively.

Prior predictive distributions for the 1000 synthetic contingency cell counts were centered around true counts, as desired. Prior predictive distributions of normalized MMSE

(a) Unimpaired

(b) MCI



(c) Dementia

(d) Other

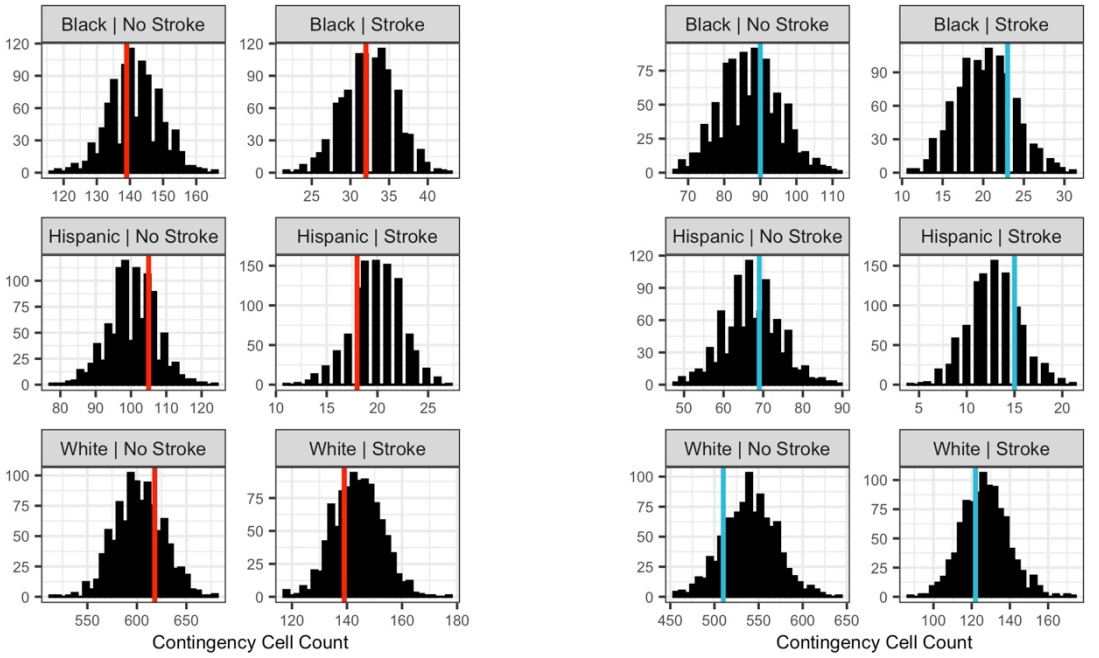


Figure 4.5: Prior predictive distributions of contingency cell counts for the simulation scenario with $n_{HRS} = 8000$ and 50% HCAP sampling proportion stratified by cognitive impairment group: (a) Unimpaired, (b) MCI, (c) Dementia, (d) Other. Distributions are based on 1000 synthetic HCAP datasets of size $n_{HCAP_{50}} = 4000$. Colored vertical lines in each panel represent true cognitive impairment group-specific contingency cell counts.

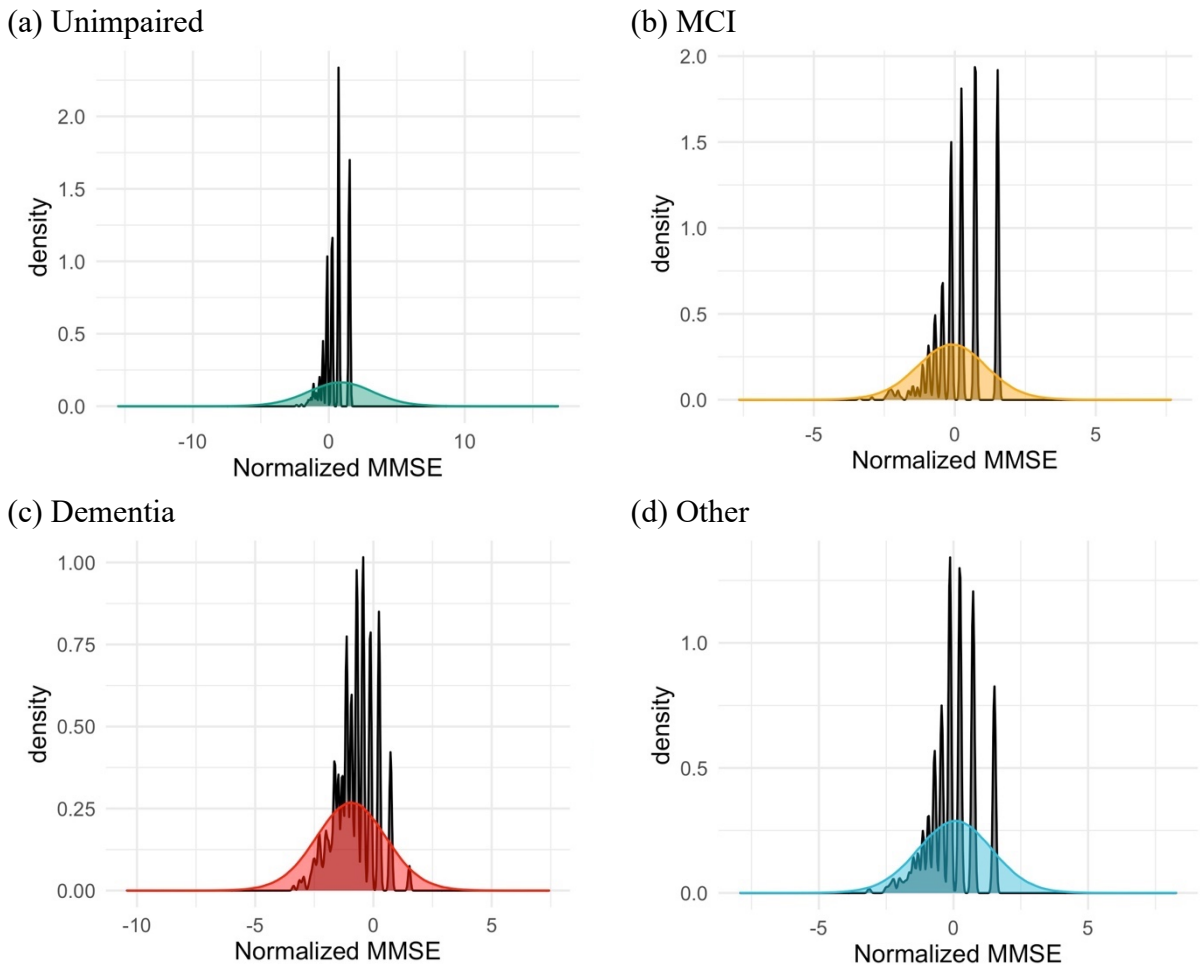
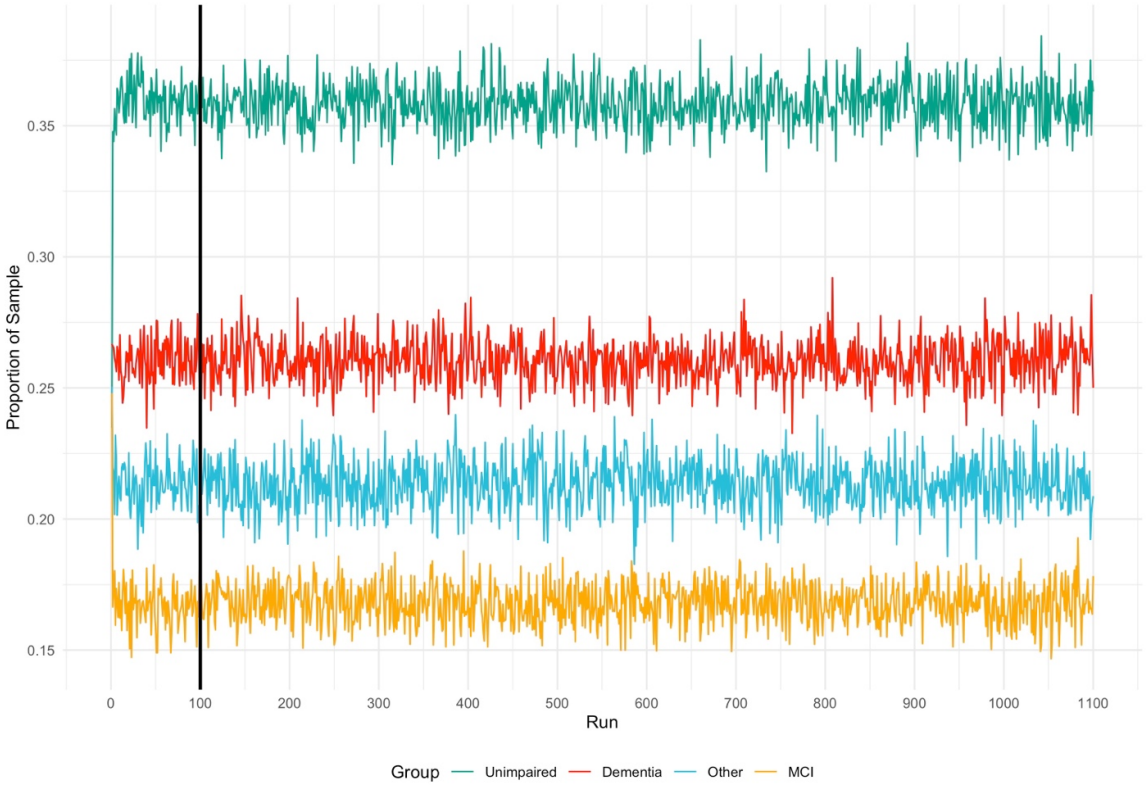


Figure 4.6: Prior predictive distributions of normalized MMSE (colored densities) for the simulation scenario with $n_{HRS} = 8000$ and 50% HCAP sampling proportion stratified by impairment group: (a) Unimpaired, (b) MCI, (c) Dementia, (d) Other. Distributions are based on 1000 synthetic HCAP datasets of size $n_{HCAP_{50}} = 4000$. Gray densities are true impairment group-specific distributions of MMSE.

were slightly wider (more variable) than true distributions, ensuring that the full range of values was captured by features encoded in the priors.

Figure 4.6 shows prior predictive distributions for Normalized MMSE overlayed on true distributions for one synthetic dataset; an animated gif cycling through all 1000 synthetic datasets was used to determine whether there was enough variability across the 1000 synthetic datasets. Prior predictive distributions for the remaining continuous variables and all other simulation scenarios were similarly more variable than true distributions. Code for producing

(a)



(b)

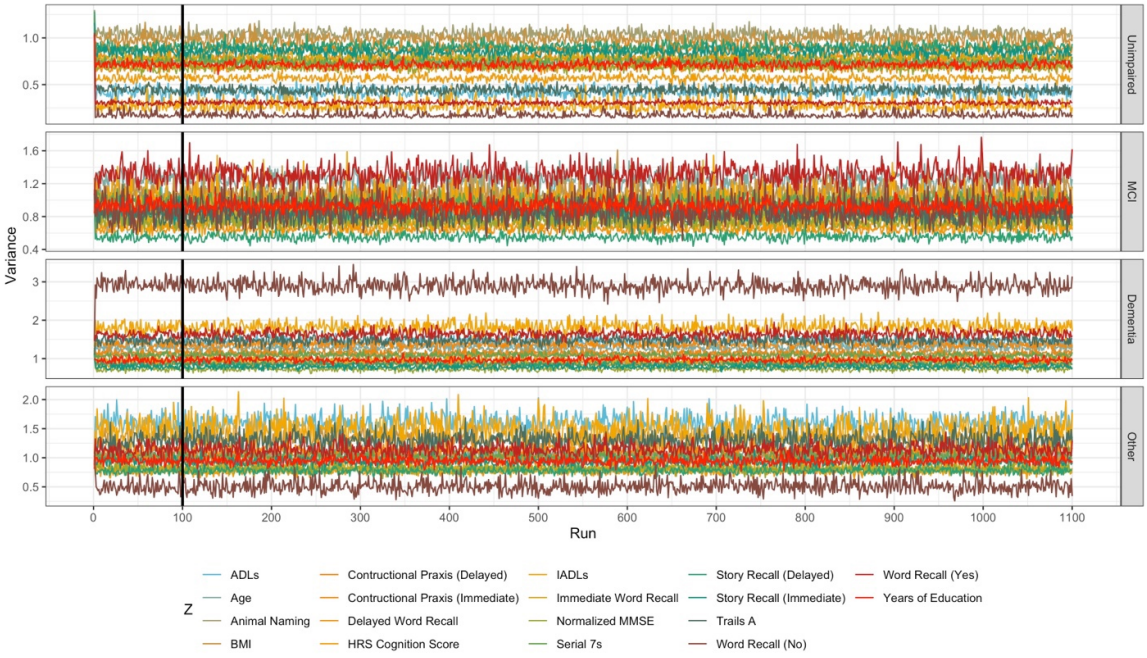


Figure 4.7: MCMC chains of proportions of (a) impairment class membership and (b) impairment group-specific variances of continuous variables for the simulation scenario with $n_{HRS} = 8000$ and 50% HCAP sampling proportion based on 1000 synthetic HCAP datasets of size $n_{HCAP_{50}} = 4000$. Black vertical lines mark the end of the burn-in period (100 runs).

prior predictive checks including .gif files for all simulation scenarios can be found in the associated GitHub repository.

4.5.2: Assessing model convergence

I produced MCMC chains for each parameter in this analysis and for each simulation scenario, but I primarily monitored cognitive impairment class proportion chains and impairment group-specific variances for continuous variables. All chains converged in all simulation scenarios. MCMC chains for the simulation scenario with $n_{HRS} = 8000$ and 50% HCAP sampling proportion are presented in

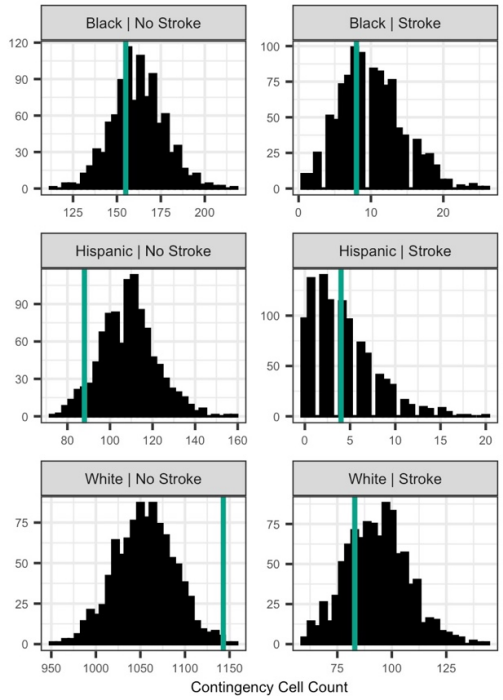
Figure 4.7. Code for producing MCMC chains for the other simulation scenarios can be found in the associated GitHub repository. Stability of the Bayesian latent class mixture model was established in the illustrative example in **Section 3.4.2.2:** by initiating chains in different parts of the parameter space and arriving at similar results for cognitive impairment class membership. Thus, in this simulation study, all chains were initiated in the "random" state with equal proportions of group membership for all cognitive impairment classes (25% Unimpaired, 25% MCI, 25% Dementia, 25% Other).

4.5.3: Posterior predictive checks

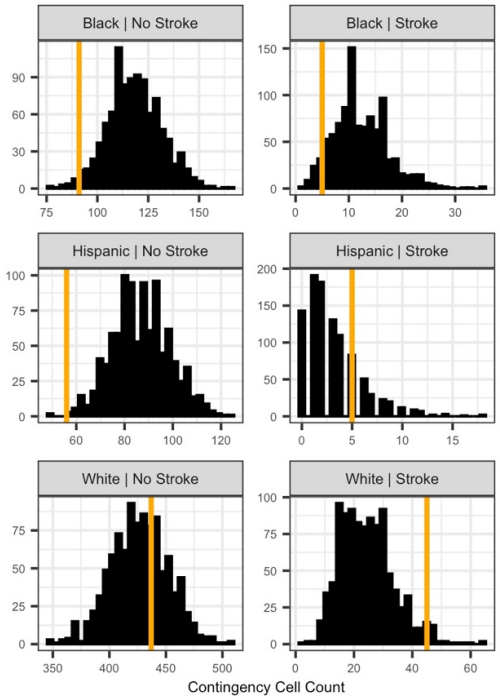
I assessed posterior distributions of contingency cell counts and median and skew for continuous variables for all simulation scenarios. All posterior statistics were stratified by predicted cognitive impairment class. Posterior predictive distributions for cell counts and median and skew of normalized MMSE in the simulation scenario with $n_{HRS} = 8000$ and 50% HCAP

sampling proportion are presented in **Figure 4.8-Figure 4.10.** Posterior cell counts for Unimpaired and Other cognitive impairment groups were mostly satisfactory while those in MCI

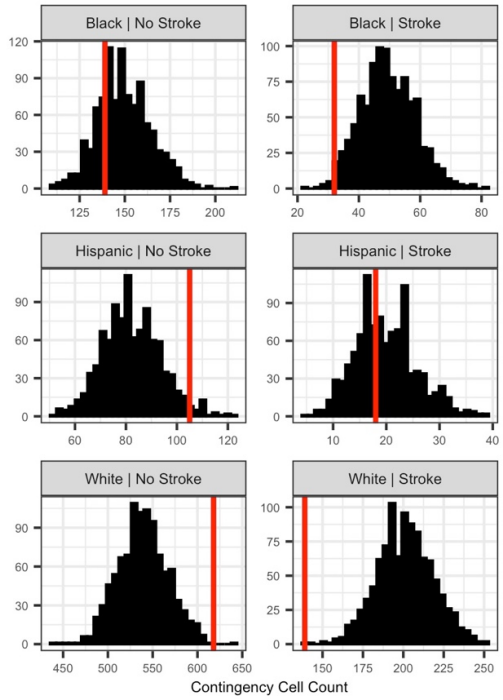
(a) Unimpaired



(b) MCI



(c) Dementia



(d) Other

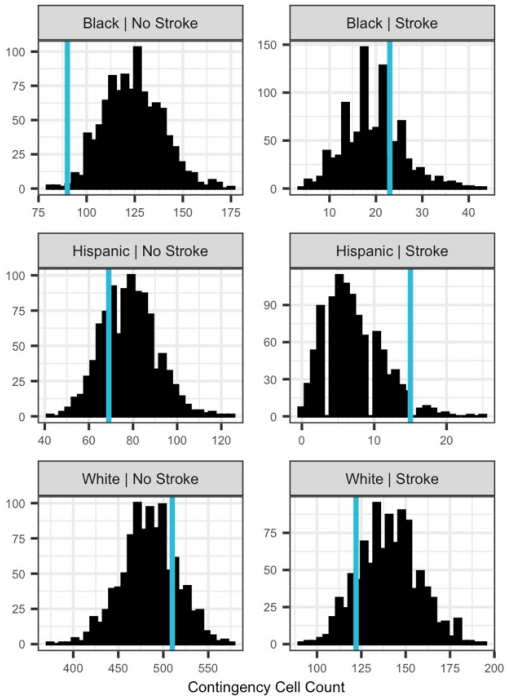


Figure 4.8: Posterior predictive distributions of contingency cell counts for the simulation scenario with $n_{HRS} = 8000$ and 50% HCAP sampling proportion stratified by impairment group: (a) Unimpaired, (b) MCI, (c) Dementia, (d) Other. Distributions are based on 1000 synthetic HCAP datasets of size $n_{HCAP_{50}} = 4000$. Colored vertical lines in each panel represent true impairment group-specific contingency cell counts.

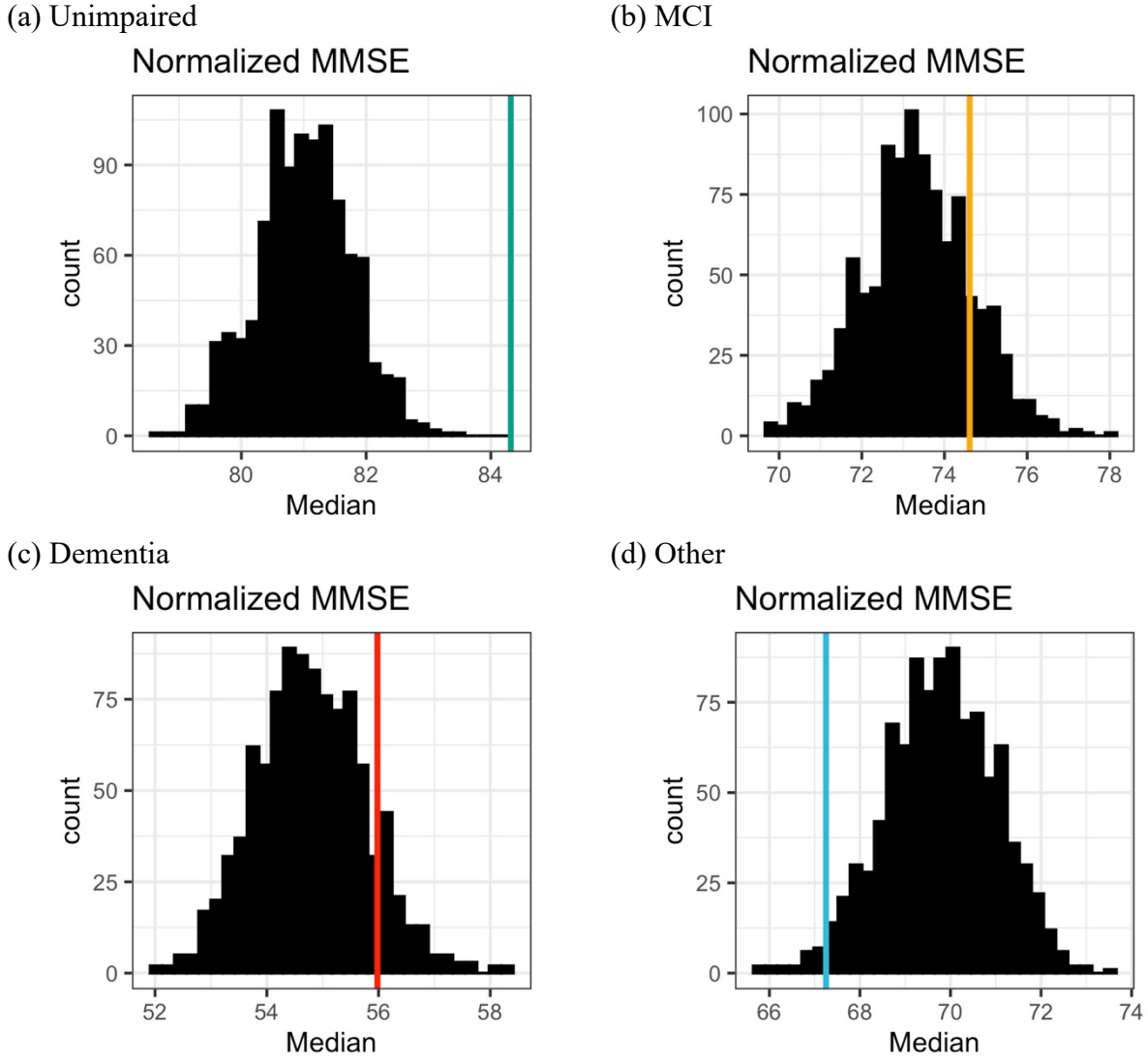


Figure 4.9: Posterior predictive distributions of median normalized MMSE for the simulation scenario with $n_{HRS} = 8000$ and 50% HCAP sampling proportion stratified by cognitive impairment group: (a) Unimpaired, (b) MCI, (c) Dementia, (d) Other. Distributions are based on 1000 synthetic HCAP datasets of size $n_{HCAP_{50}} = 4000$. Colored vertical lines in each panel represent true cognitive impairment group-specific medians of normalized MMSE.

and Dementia groups were less so. Posterior distributions in the MCI group included too many Black and Hispanic participants without stroke history and too few White participants with stroke history.

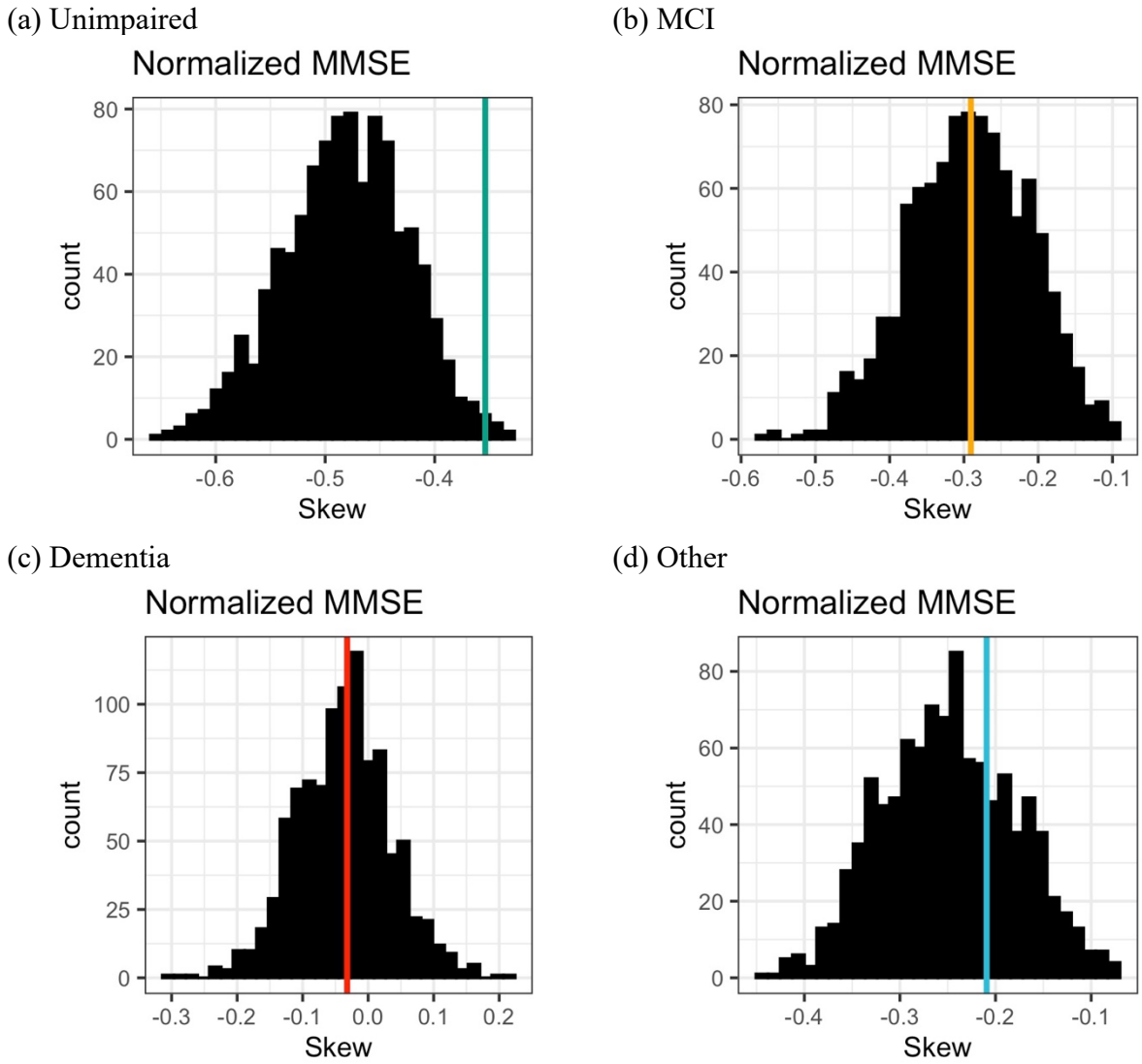


Figure 4.10: Posterior predictive distributions of MMSE skew for the simulation scenario with $n_{HRS} = 8000$ and 50% HCAP sampling proportion stratified by impairment group: (a) Unimpaired, (b) MCI, (c) Dementia, (d) Other. Distributions are based on 1000 synthetic HCAP datasets of size $n_{HCAP_{50}} = 4000$. Colored vertical lines in each panel represent true impairment group-specific skew of normalized MMSE.

In the Dementia group, there were too few White participants without stroke history and too many with stroke history (**Figure 4.8)**. Posterior distributions of median normalized MMSE were

roughly centered around observed medians for MCI and Dementia groups but were only off by about 2 points (on a scale of 0-100) for Unimpaired and Other groups (**Figure 4.9**). Posterior distributions of normalized MMSE skew were roughly centered at true measures of skew, demonstrating that synthetic variables were able to reproduce skewness in these measures (**Figure 4.10**). Posterior predictive distributions of medians and skewness for the remaining continuous variables for this simulation scenario are displayed in **Appendix Figure D.1** and **Appendix Figure D.2**, respectively. Posterior predictive distributions for other simulation scenarios were similar and code for producing figures is available on the associated GitHub Repository.

4.5.4: Algorithmic dementia classification

Figure 4.11 shows 95% credible intervals of participant (a) counts and (b) proportions in each cognitive impairment class across 1000 synthetic HCAP samples for the simulation scenario with $n_{HRS} = 8000$ and 50% HCAP sampling proportion.

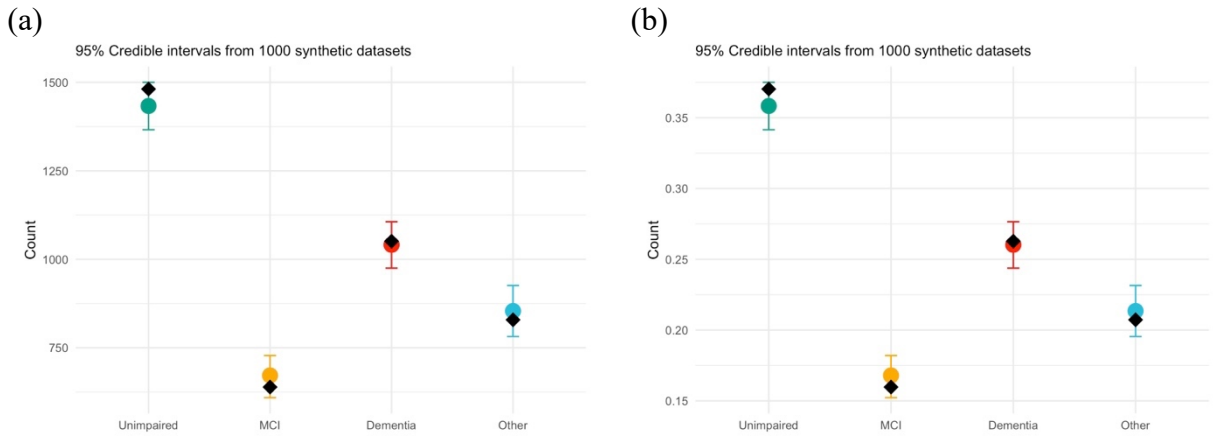


Figure 4.11: 95% interval estimates of participant counts (a) and proportions (b) within each impairment group across 1000 synthetic HCAP datasets for the simulation scenario with $n_{HRS} = 8000$ and 50% HCAP sampling proportion. Black diamonds are group-specific true counts/proportions.

Every credible interval captured the true cognitive impairment class count/proportion, and point estimates were nearly identical to true counts/proportions. The largest discrepancy in mean count was in the Unimpaired group where the model underestimated the count by about 60 people on average (about a 1.5% error).

Results for model tuning were satisfactory, so I proceeded with the simulation study to assess performance of the Bayesian latent class mixture model across repeated runs in the various simulation scenarios.

4.6: Simulation study results

4.6.1: Algorithmic dementia classification

Mean and 95% interval estimates for cognitive impairment class proportions summarized across 1000 simulation runs are presented in **Figure 4.12** by HRS sample size and HCAP sampling proportion. Mean cognitive impairment class proportions were obtained by averaging means of posterior predictive distributions of participant counts in each cognitive impairment class across simulation runs and dividing by HCAP sample size. Upper and lower limits of cognitive impairment class proportions were obtained by averaging 97.5% and 2.5% percentiles of posterior predictive distributions of participant counts, respectively, across simulation runs and dividing by HCAP sample size. Point estimates appear identical across HRS sample sizes and HCAP sampling proportions and are close to superpopulation proportions. The average interval estimates included the truth in all scenarios. As expected, sample size only impacts estimation precision in this simulation due to the nature of SRS sampling.

Figure 4.13 shows that 95% interval estimates of cognitive impairment class proportions achieve at least nominal coverage across all simulation scenarios except in the Unimpaired class which achieved about 94% coverage in scenarios with smaller sample sizes. Bias and RMSE provide additional context for this result. Bias and percent bias in cognitive impairment class proportions across 1000 simulation runs are presented in **Figure 4.14(a)** and **Figure 4.14(b)**.

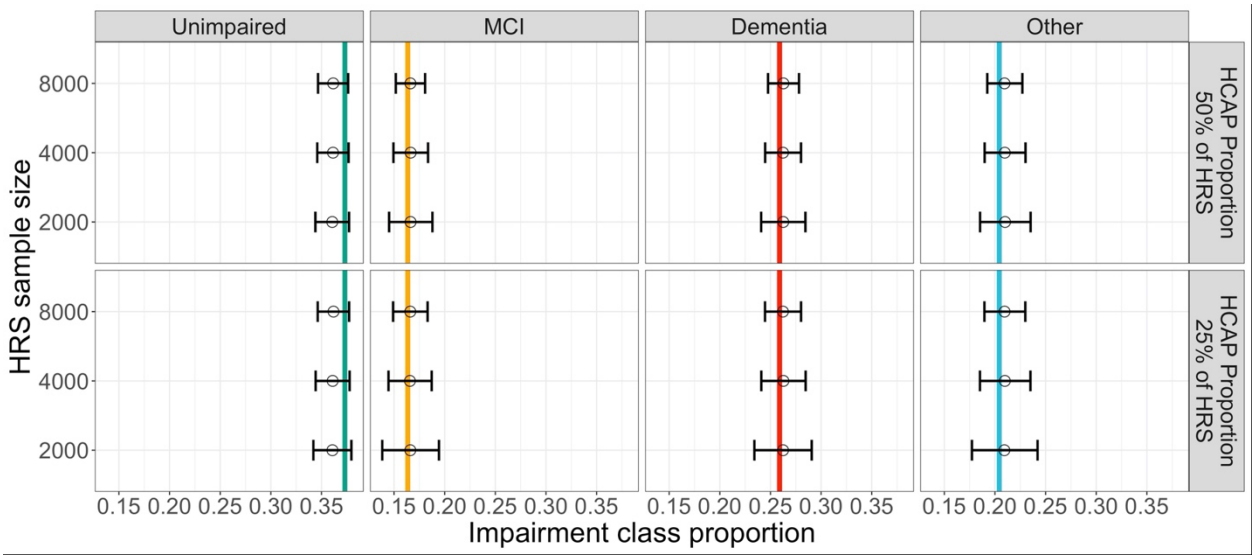


Figure 4.12: Mean and 95% interval estimates for cognitive impairment class proportions by HRS sample size and HCAP sampling proportions averaged across 1000 simulation runs. Colored vertical lines denote true impairment class proportions in the superpopulation.

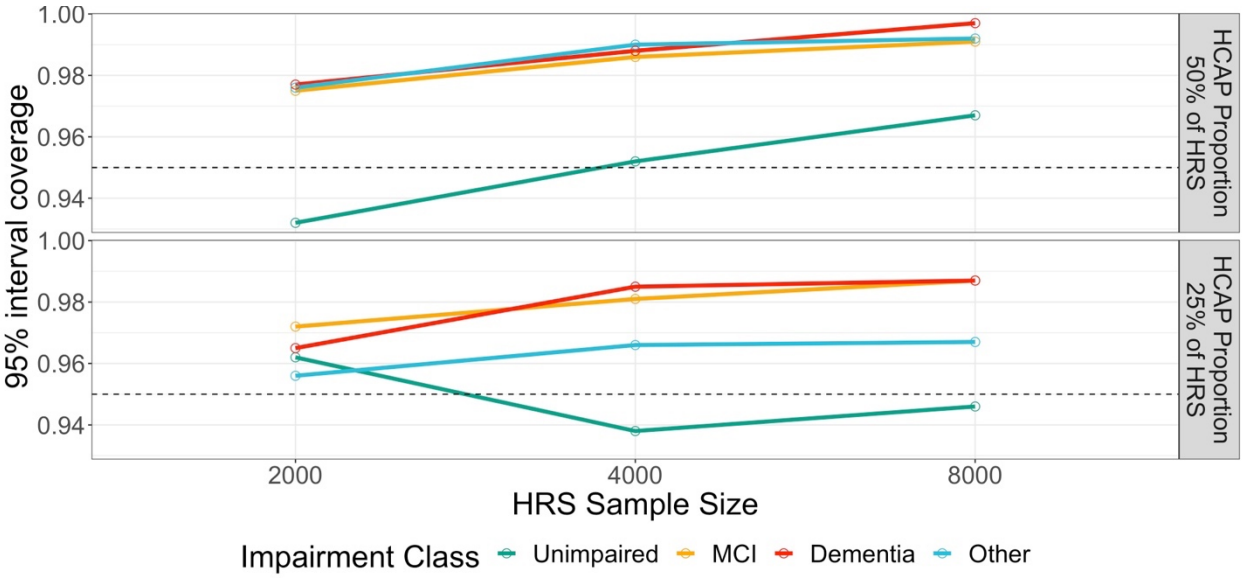
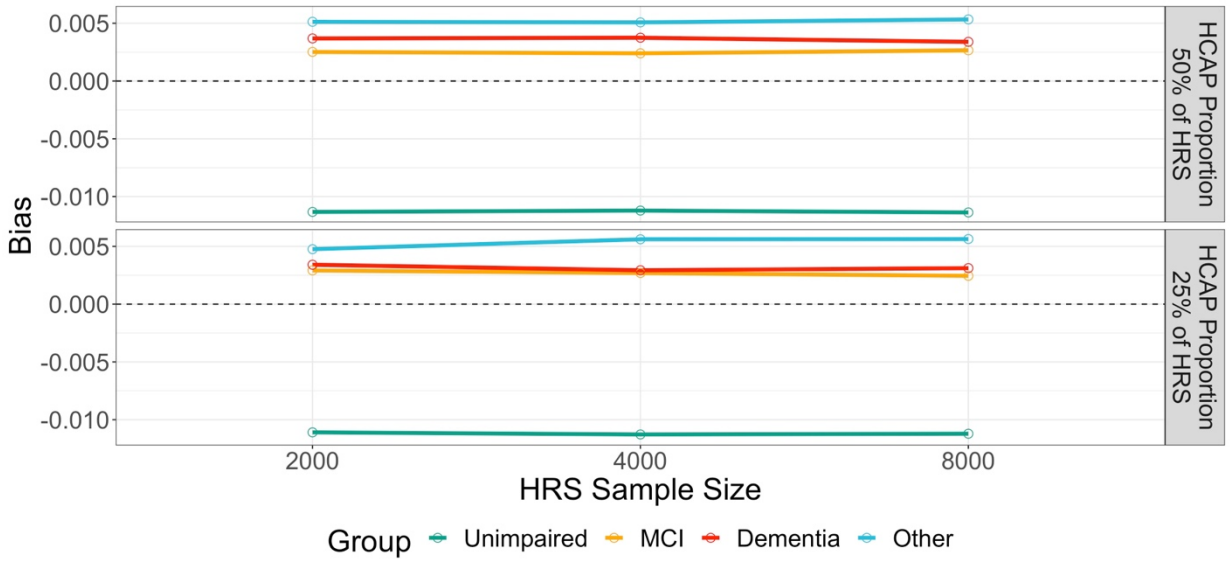


Figure 4.13: 95% interval coverage of true impairment class proportion by cognitive impairment class, HRS sample size, and HCAP sampling proportion across 1000 simulation runs. Horizontal dashed line denotes nominal coverage of 95%.

(a)



(b)

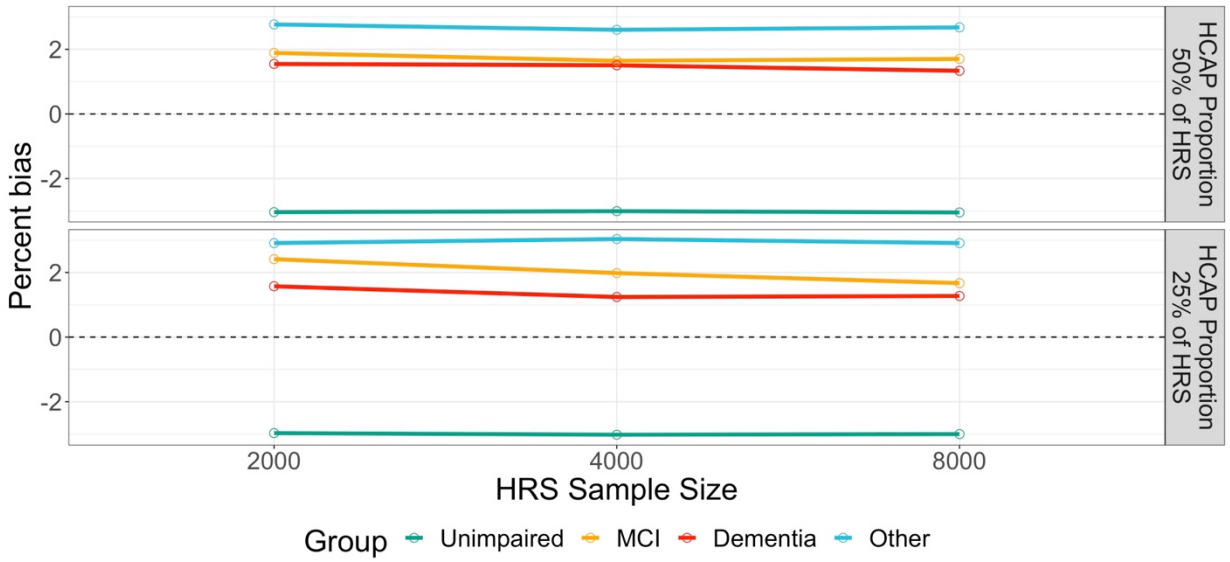


Figure 4.14: Bias (a) and percent bias (b) in cognitive impairment class proportions by HRS sample size and HCAP sampling proportion across 1000 simulation runs. Dashed horizontal line denotes no bias.

Bias was calculated by averaging mean posterior impairment class proportions across runs and subtracting the cognitive impairment class proportion in the superpopulation (e.g., for the Unimpaired group, $\text{bias}_{Unimpaired} = \bar{\hat{\phi}}_{Unimpaired} - \phi_{Unimpaired}$, where $\hat{\phi}$ is the mean of the posterior predictive distribution of proportion of Unimpaired participants). Percent bias is bias divided by the true cognitive impairment class proportion in the superpopulation. We see that bias and percent bias were minimal and stable across HRS samples sizes and HCAP sampling proportions. The model consistently underestimated the proportion of Unimpaired participants and overestimated participants in all other cognitive impairment classes.

Root mean square error (RMSE = $\sqrt{bias^2 + variance}$) gives us insight to bias and variance of estimators simultaneously. Lower values of RMSE are desirable as they indicate less bias/variance. **Figure 4.15** shows RMSE for cognitive impairment class proportions across simulation runs. RMSE for the Unimpaired group is stable across HRS sample size and HCAP sampling proportion but decreases for all other cognitive impairment groups as sample size

increases. Since we observed that bias is stable across simulation scenarios, this indicates increased precision of estimates for proportion of participants in MCI, Dementia, and Other impairment groups but not in the Unimpaired group.

Looking at race-stratified bias plots, we see that the model on average underestimates proportions of White participants and overestimates proportions of Black and Hispanic

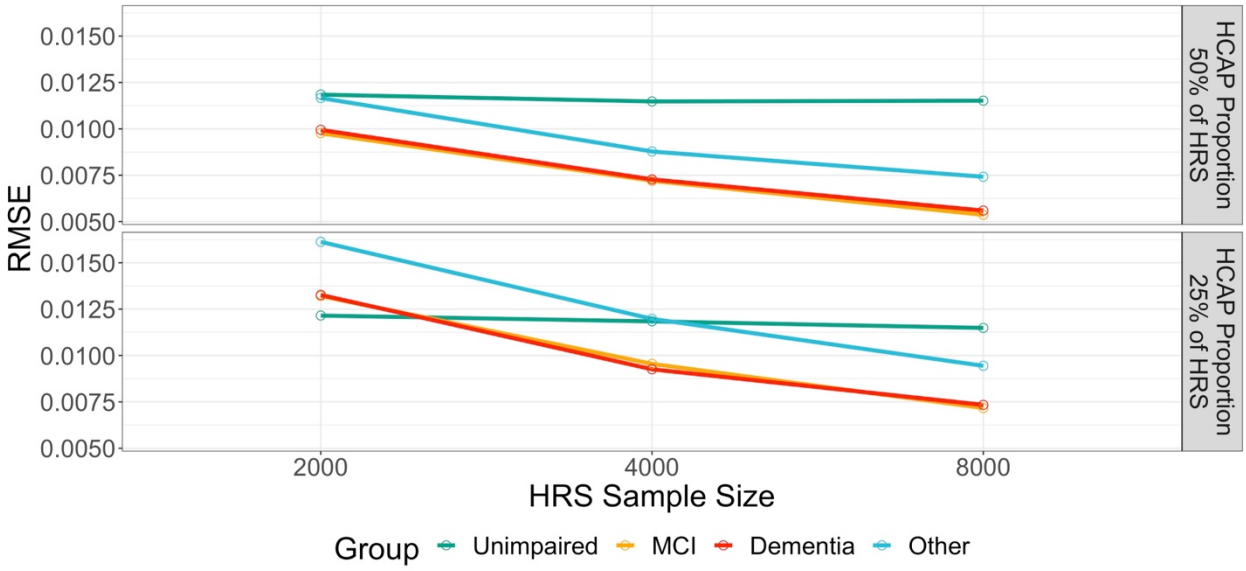
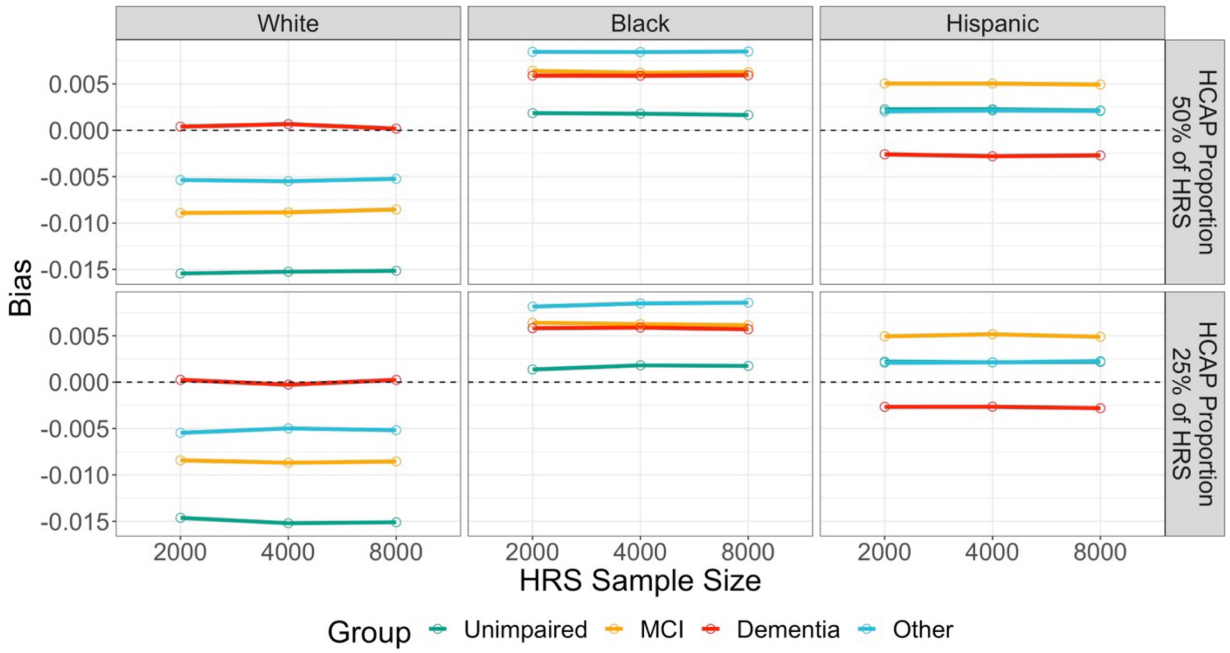


Figure 4.15: Root mean square error (RMSE) for estimated cognitive impairment class proportions by HRS sample size and HCAP sampling proportion across 1000 simulation runs.

(a)



(b)

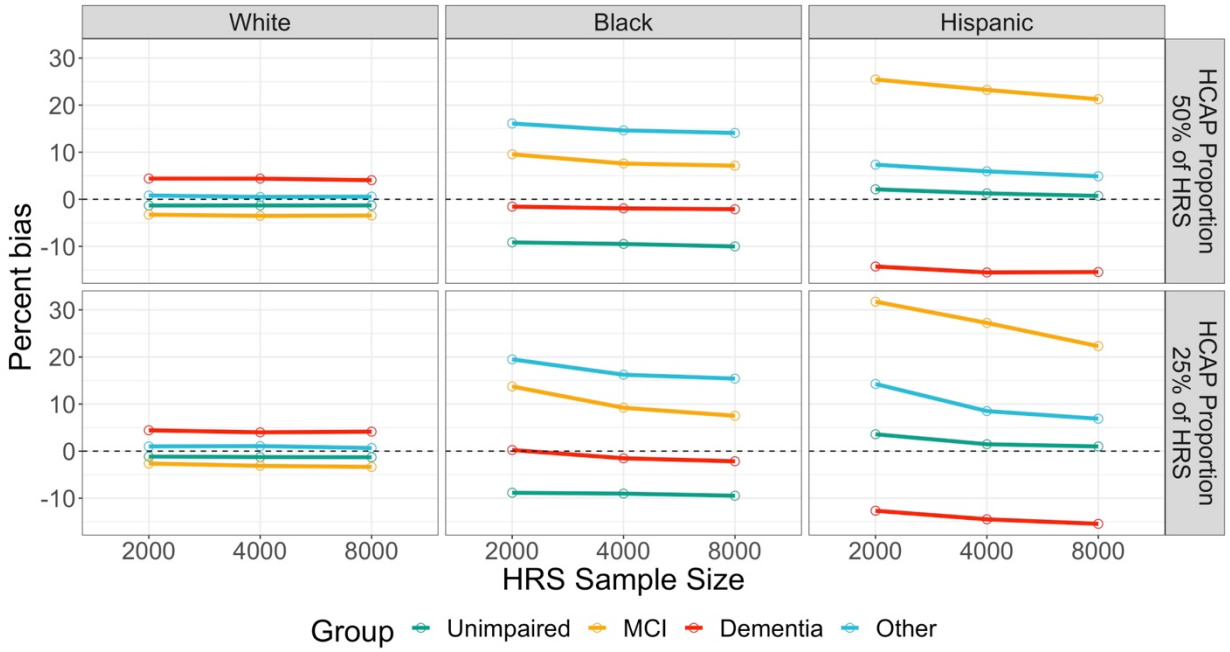


Figure 4.16: Race/ethnicity-specific bias (a) and percent bias (b) in cognitive impairment class proportions by HRS sample size and HCAP sampling proportion across 1000 simulation runs. Dashed horizontal line denotes no bias.

participants in cognitive impairment groups. There was no bias in estimates of proportions of White participants in the Dementia group, though. On an absolute scale, there appears to be

minimal bias in estimates of cognitive impairment class proportions for all race/ethnicities with the maximum bias (-1.5%) in the White Unimpaired group (**Figure 4.16a**). On a relative scale, however, we see the disproportionate amount of bias in Black and Hispanic participants (upwards of 30%) due to smaller sample sizes in those groups (**Figure 4.16b**). Bias decreases slightly across HRS sample sizes and HCAP sampling proportions on a relative scale in all cognitive impairment groups, but the model differentially misclassifies participants by race since it tended to overestimate cognitive impairment group membership in Black and Hispanic participants and underestimate membership in White participants.

Race-stratified RMSE plots are decreasing across HRS sample sizes and HCAP sampling proportions for all cognitive impairment groups for all race/ethnicities (**Figure 4.17**).

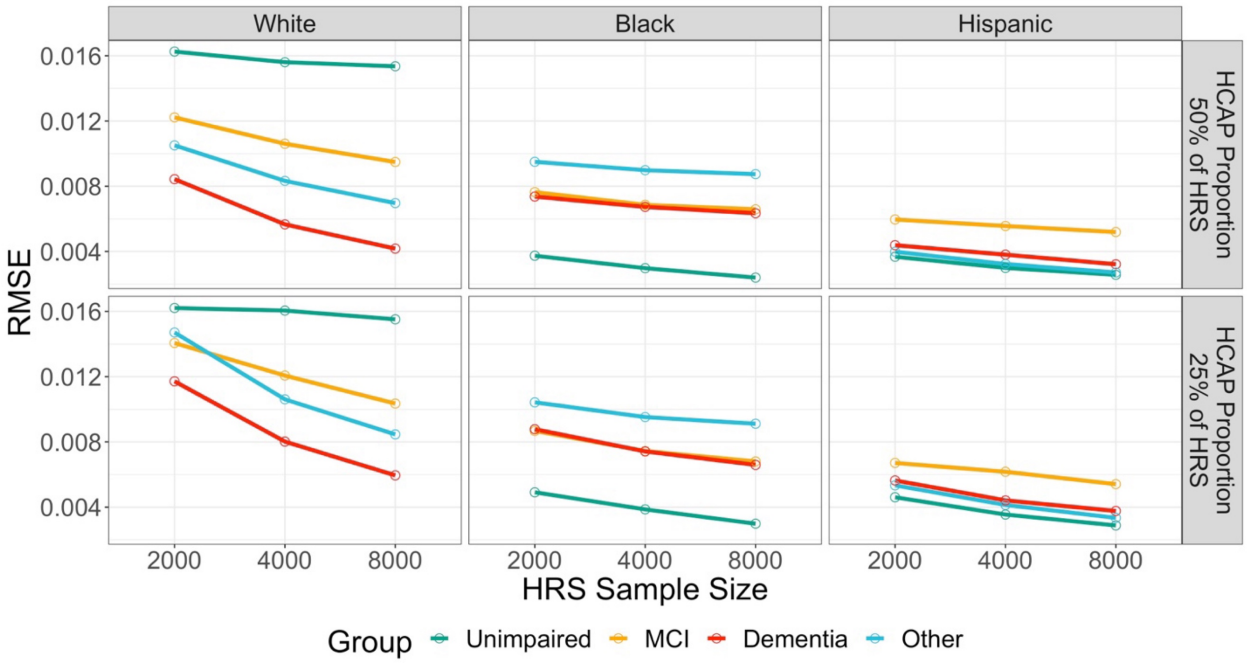


Figure 4.17: Race/ethnicity-specific RMSE for estimated cognitive impairment class proportions by HRS sample size and HCAP sampling proportion across 1000 simulation runs.

The lowest RMSEs are found in the Hispanic group which is likely due to the property of variance of multinomial distributions in small samples discussed at the end of Chapter 3 (see **Section 3.5:**).

4.6.2: Inference in a population-representative study

Algorithmic dementia classification results presented above showed good coverage and minimal bias overall across cognitive impairment groups but demonstrated that the model differentially misclassified participants by race/ethnicity. This simulation study was anchored in an analytic question about racial/ethnic differences in prevalent dementia in the 2016 HRS study. This next set of results assesses the quality of inferences at the population-representative level resulting from algorithmic dementia classification using the Bayesian latent class mixture model.

Figure 4.18 shows mean and 95% interval estimates for age- and sex-standardized race/ethnicity-specific dementia prevalence by HRS sample size and HCAP sampling proportion across 1000 simulation runs. After standardizing race/ethnicity-specific prevalence estimates in all simulation runs, point estimates and upper and lower confidence limits were calculated using the same steps described above for cognitive impairment class proportions presented in **Figure 4.12**. In this case, estimation accuracy and precision improve with increasing sample size even though all sampling is SRS. It seems reasonable that sampling a larger proportion of the population we aim to generalize to improves the accuracy of the generalization. The increased accuracy is more pronounced in Black and Hispanic participants as they comprise a smaller proportion of the sample. Point estimates are closest to the truth in the largest sample sizes, but all interval estimates capture true prevalence of dementia on average. Point estimates for

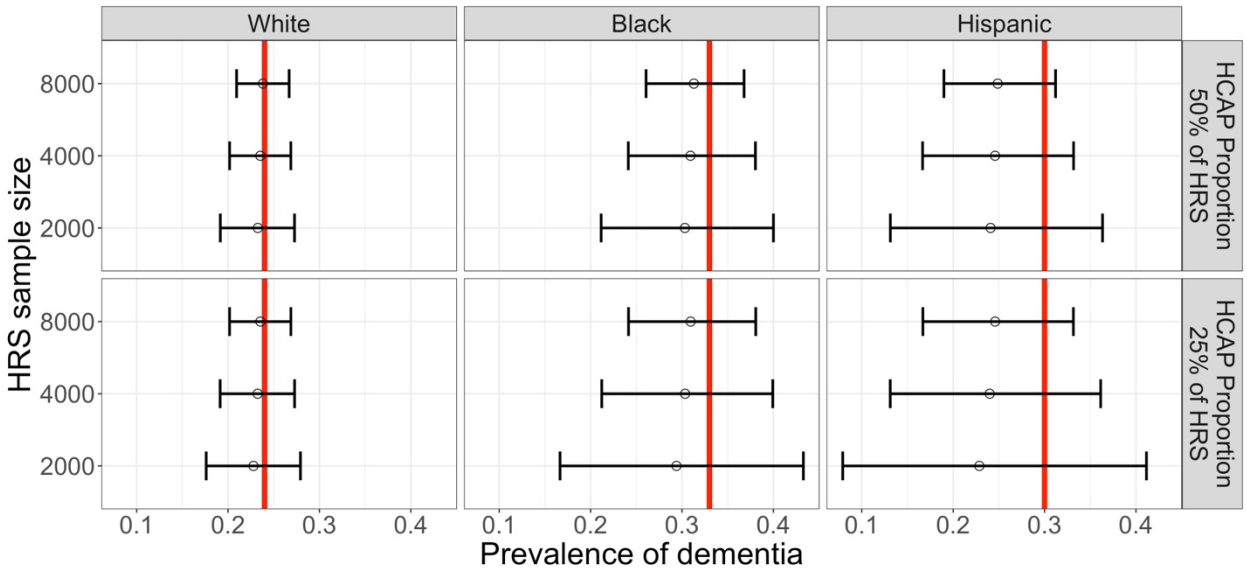


Figure 4.18: Mean and 95% interval estimate for race/ethnicity-specific age- and sex-standardized dementia prevalence by HRS sample size and HCAP sampling proportion across 1000 simulation runs. Prevalence was standardized to synthetic HRS samples at each simulation run. Vertical lines denote true race/ethnicity-specific dementia prevalence in the superpopulation.

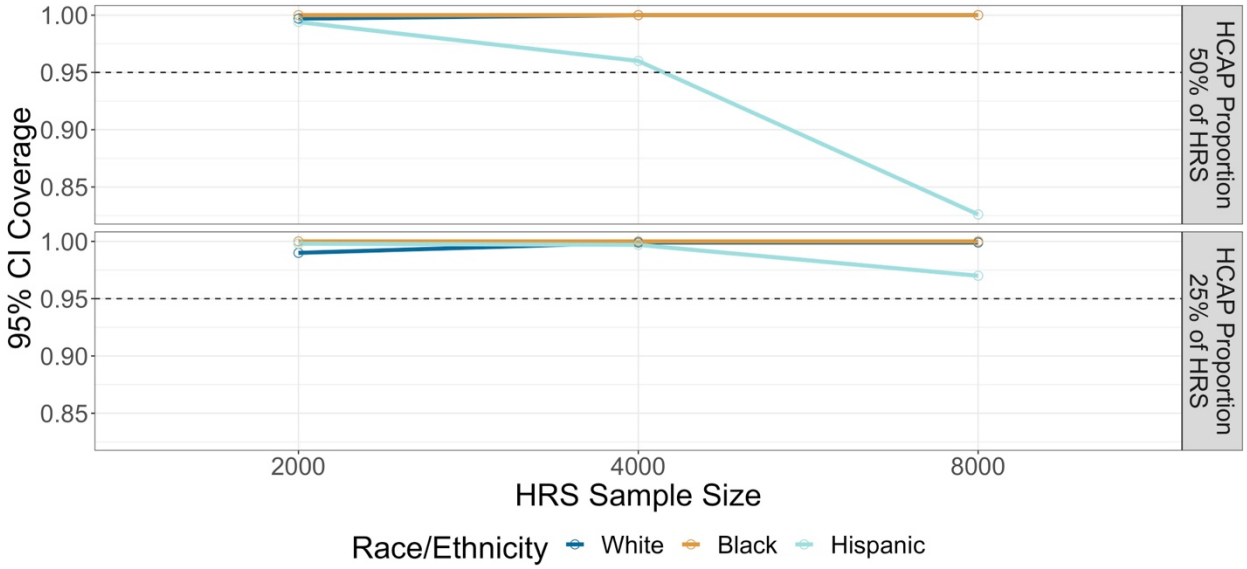


Figure 4.19: 95% interval coverage of true prevalence of dementia by race/ethnicity, HRS sample size, and HCAP sampling proportion across 1000 simulation runs. Dashed horizontal line denotes nominal coverage of 95%.

Hispanic participants are still notably less compared to the truth even in the largest samples, however. On average, the prevalence of dementia among Hispanics was estimated as 25.0% when the truth in the superpopulation was 30.2%. In fact, **Figure 4.19** shows above nominal coverage for all race/ethnicity-specific dementia prevalence interval estimates across HRS sample sizes and HCAP sampling proportions expect for Hispanic participants in the simulation scenario with the largest sample size $n_{HRS} = 8000$ with 50% HCAP sampling proportion.

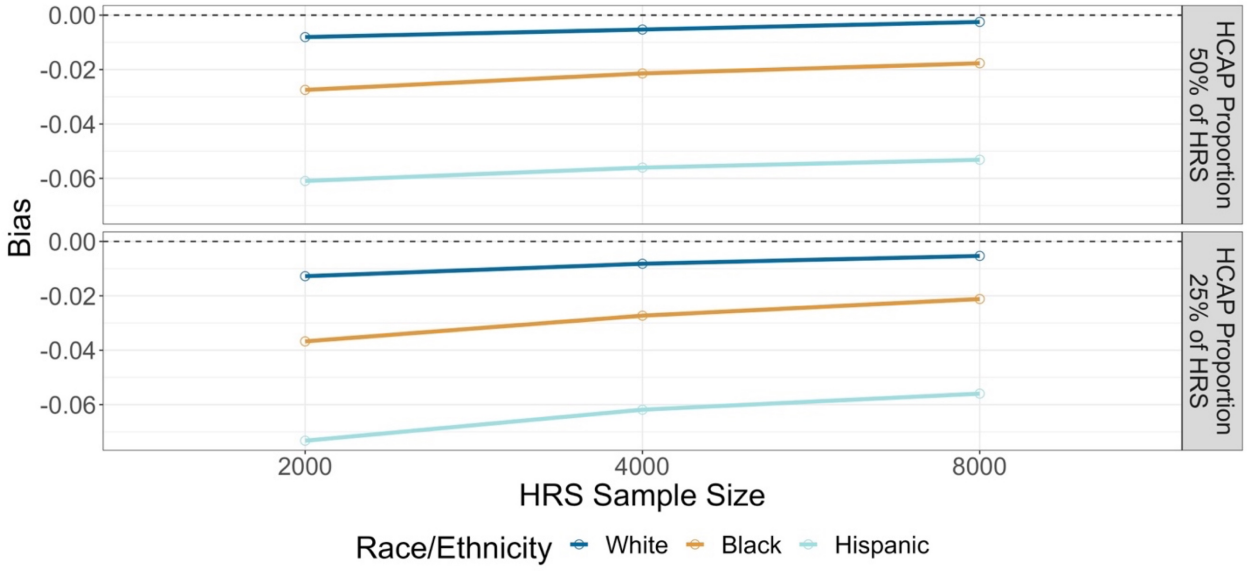


Figure 4.20: Bias in race-specific age and sex-standardized dementia prevalence by HRS sample size and HCAP sampling proportion across 1000 simulation runs. Dashed horizontal line denotes no bias.

Bias and RMSE for race/ethnicity-specific dementia prevalence estimates decreased with increasing sample size (**Figure 4.20** and **Figure 4.21**). It is likely that coverage gets worse as sample size increases for Hispanic participants because variance decreased at a faster rate than bias in this subsample.

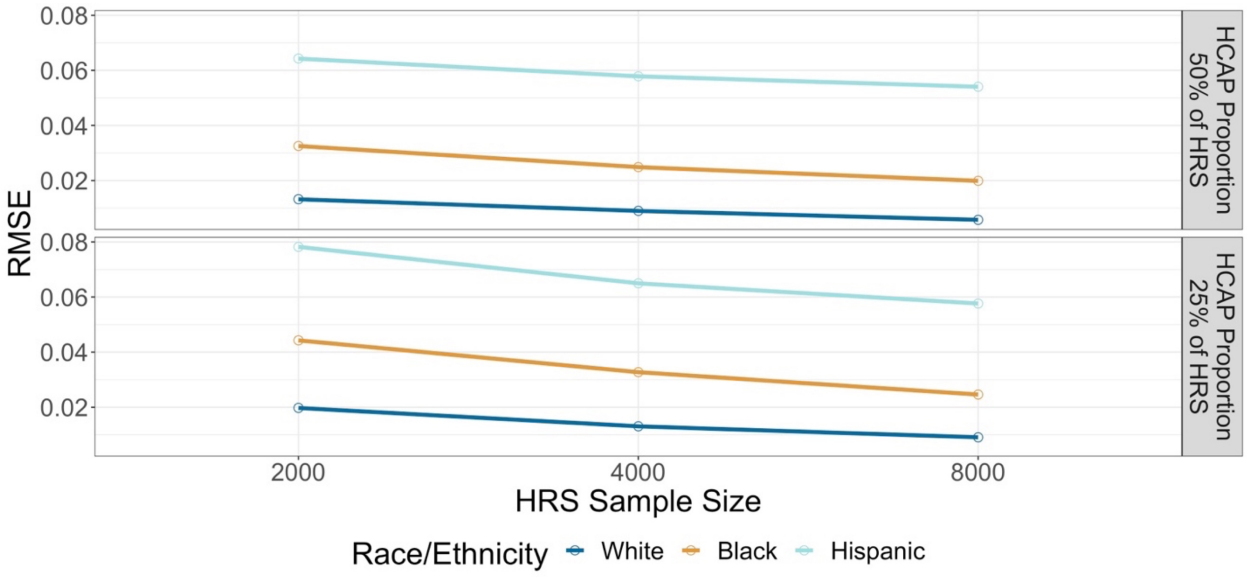


Figure 4.21: RMSE for race-specific age and sex-standardized dementia prevalence by HRS sample size and HCAP sampling proportion across 1000 simulation runs. Dashed horizontal line denotes no bias.

Figure 4.22 shows mean and 95% interval estimates of prevalence ratios (PR= rate in specified population/rate in White population) and differences (PD= rate in specified population - rate in White population) for Black vs. White and Hispanic vs. White participants. In the case of PRs, sample size affected estimation precision only. Point estimates were close to the superpopulation truth across HRS sample sizes and HCAP sampling proportions for PR for Black vs. White participants. However, the only average estimate precise enough to exclude the null was PR for Black vs. White participants in the scenario with $n_{HRS} = 8000$ with 50% HCAP sampling proportion. Increased prevalence for Hispanic vs. White participants were notably underestimated even in the scenario with the largest sample sizes ($n_{HRS} = 8000$ with 50% HCAP sampling proportion). Hispanic participants were estimated to have a 5% increased prevalence of dementia when the truth in the superpopulation was 25% (**Figure 4.22a**). There were marginal improvements with increasing sample size on the difference scale (**Figure 4.22b**).

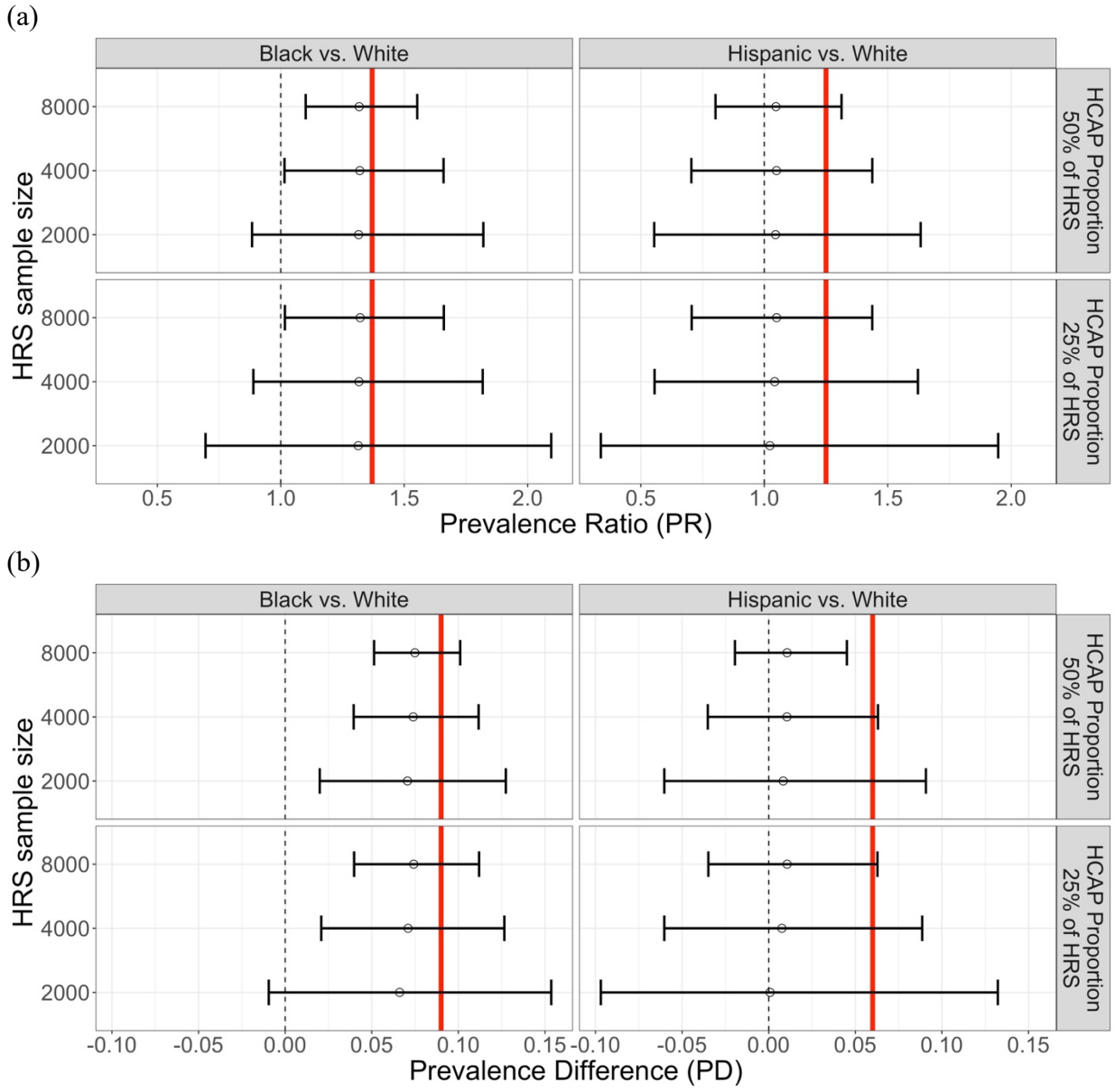


Figure 4.22: Mean and 95% interval estimate for (a) prevalence ratio (PR) and (b) difference (PD) in dementia for Black vs. White and Hispanic vs. White participants by HRS sample size and HCAP sampling proportion across 1000 simulation runs. Red vertical lines denote true PR/PD of dementia in the superpopulation. Black dashed vertical line denotes no racial ethnic differences (PR=1; PD=0).

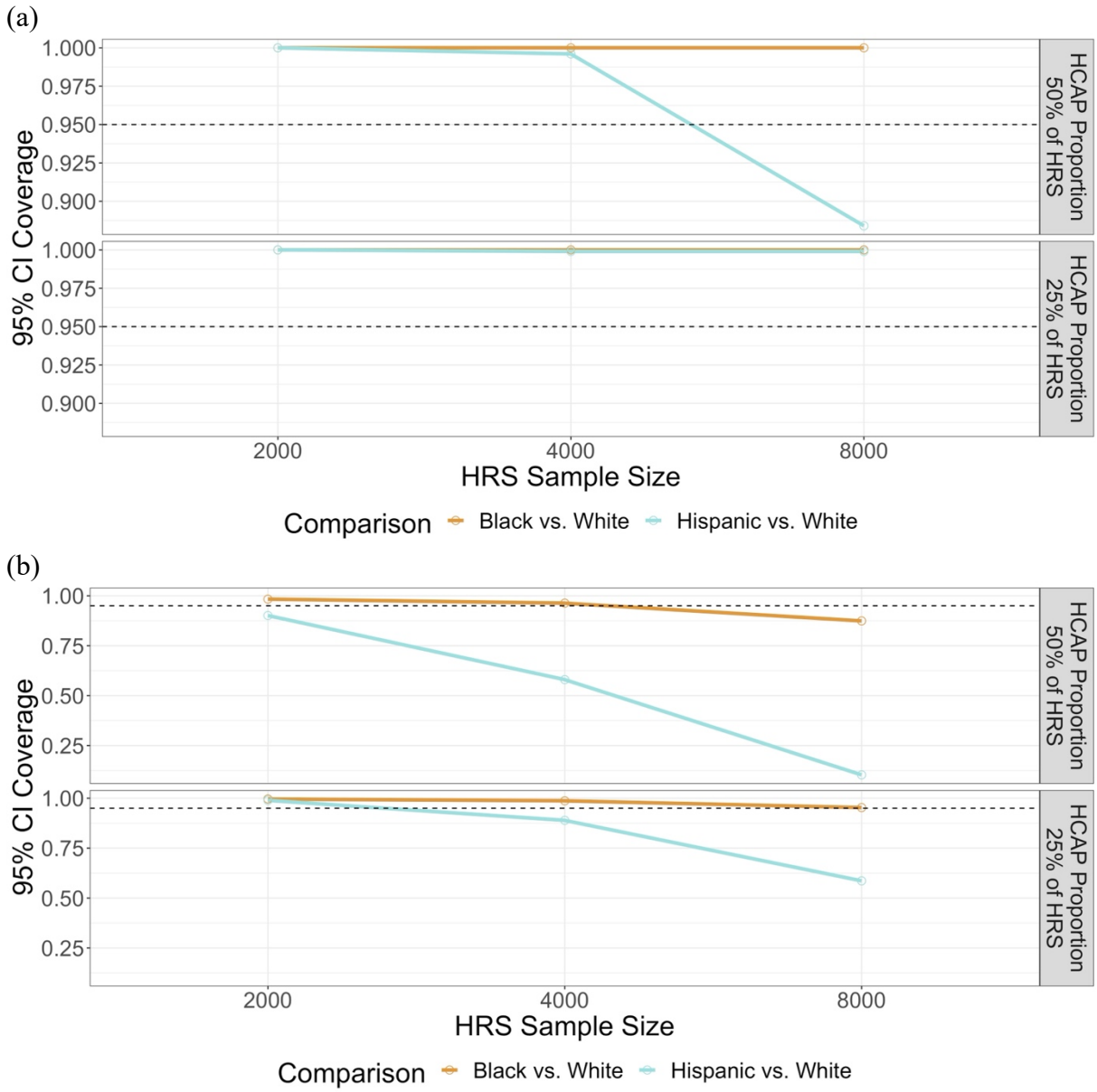


Figure 4.23: 95% interval coverage of (a) true prevalence ratio (PR) and (b) true prevalence difference in dementia for Black vs. White and Hispanic vs. White participants by HRS sample size and HCAP sampling proportion across 1000 simulation runs. Dashed horizontal line denotes nominal coverage of 95%.

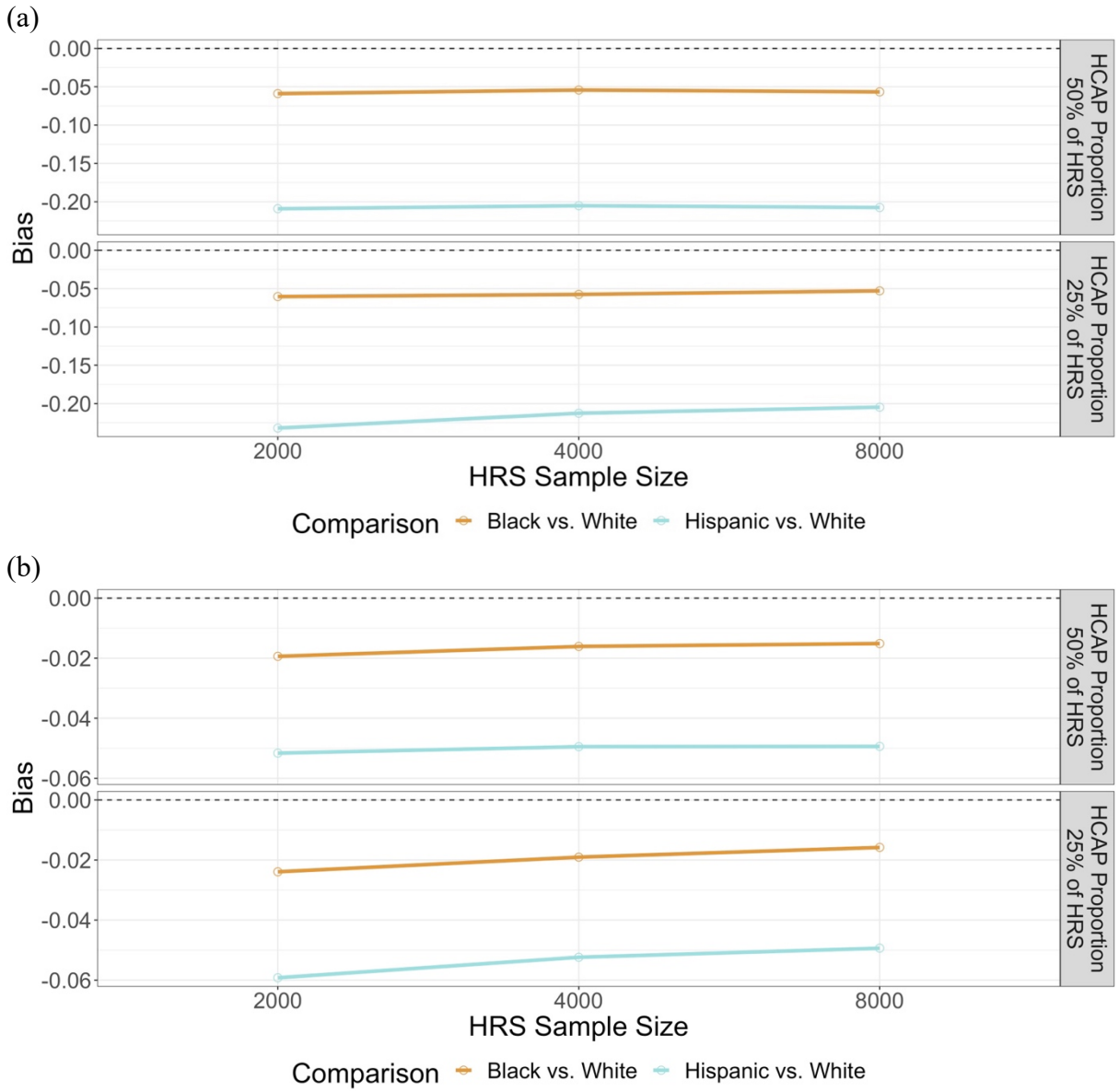


Figure 4.24: Bias in estimated (a) prevalence ratio (PR) and (b) prevalence difference of dementia for Black vs. White and Hispanic vs. White participants by HRS sample size and HCAP sampling proportion across 1000 simulation runs. Dashed horizontal line denotes no bias.

Figure 4.23a shows 100% coverage for all dementia PR 95% interval estimates across HRS sample sizes and HCAP sampling proportions except for PR Hispanic vs. White participants in the scenario with $n_{HRS} = 8000$ with 50% HCAP sampling proportion. On the difference scale, there is nominal coverage for PD Black vs. White but not for PD Hispanic vs. White (**Figure**

4.23b). Bias for PR and PD estimates were relatively stable (**Figure 4.24**) and RMSE decreased slightly (**Figure 4.25**) as sample sizes increased.

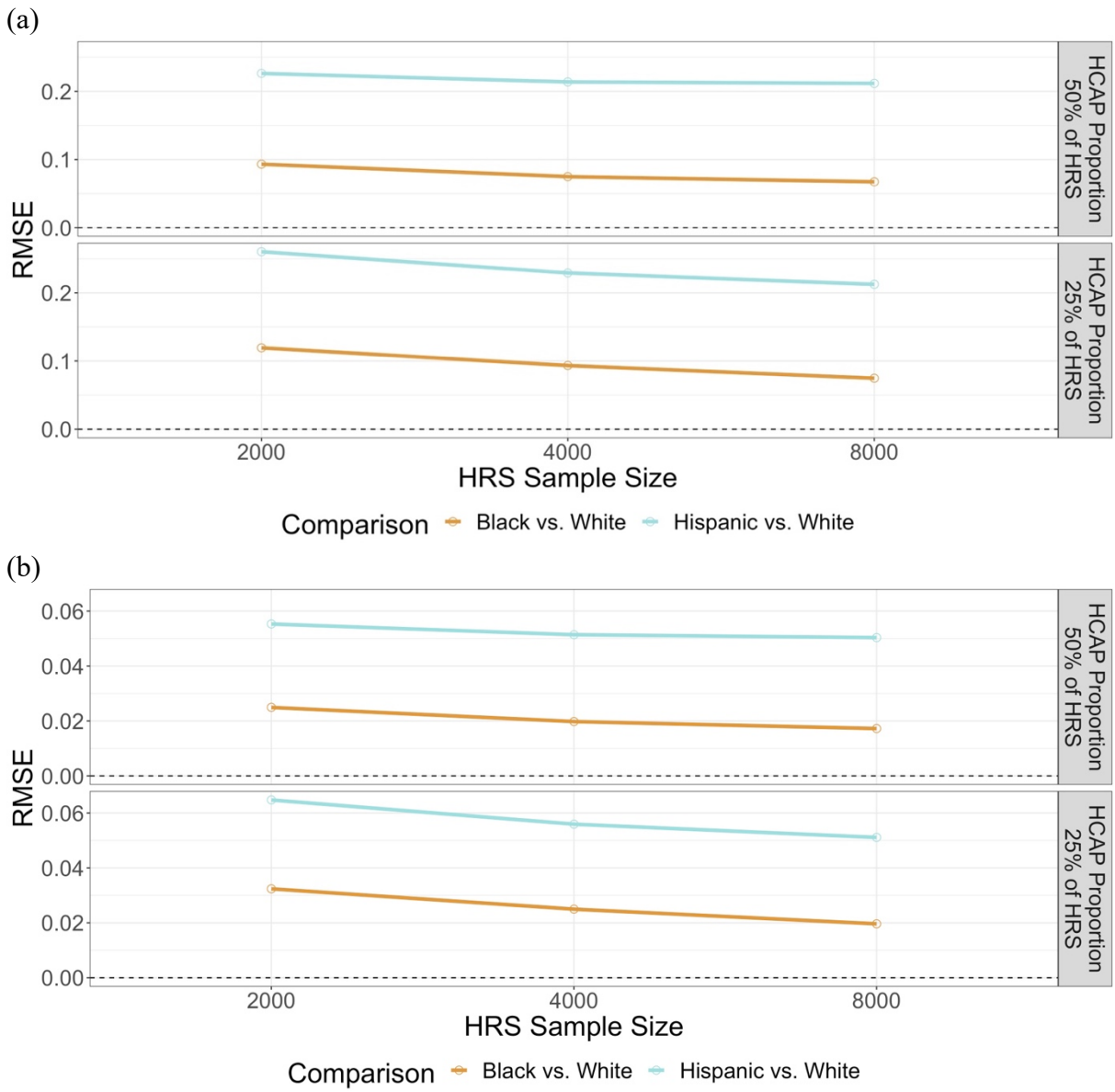


Figure 4.25: RMSE for estimated (a) prevalence ratio (PR) and (b) prevalence difference (PD) in dementia for Black vs. White and Hispanic vs. White participants by HRS sample size and HCAP sampling proportion across 1000 simulation runs. Dashed horizontal line denotes no bias.

4.7: Discussion

Results from this simulation study demonstrate good results overall for algorithmic dementia classification using the Bayesian latent class mixture model with priors specified based on the ADAMS sample. Race/ethnicity-specific analyses demonstrate that downstream analyses in population-representative samples would likely be more accurate for White and Black participants as true standardized dementia prevalence estimates and prevalence ratios were recovered for these groups in population-representative samples. This is likely due to the very small subset of Hispanic participants in the ADAMS sample (n=83) on which prior distributions for this group are based. Accurate algorithmic dementia classification for Hispanic participants in HRS is a persistent challenge because algorithms tend to provide unstable estimates for this subgroup due to small sample sizes. As a result, algorithms are occasionally developed in samples without Hispanic participants and are not recommended for use within this subgroup (Gianattasio, Ciarleglio, and Power 2020; Wu et al. 2013).

Compatibility between the prior (ADAMS) and the data (HCAP) is particularly important in this work because ADAMS is the only source of information for classifying participants into cognitive impairment classes since HCAP does not currently perform gold-standard dementia adjudication. Since ADAMS and HCAP participants were sampled from HRS using different sampling frames and inclusion criteria and because these participants were recruited nearly 20 years apart, there are differences in sample characteristics between the studies.

The overall results of algorithmic dementia classification and subsequent population-level inferences were satisfactory for Black and White participants, but pre-simulation posterior predictive checks showed some evidence for lack of model fit in the synthetic HCAP datasets, especially in race by stroke distributions which only captured true counts in the tails of some of

the distributions (**Figure 5.5**). Medians for continuous variables in synthetic datasets were mostly centered around true medians or were off by a small amount (**Appendix Figure D.1**), but the degree of skewness in observed data for the most part was not well-captured in synthetic datasets (**Appendix Figure D.2**). Whether or not the quality of synthetic datasets is satisfactory depends on the intended downstream use. The simulation study showed that the resulting algorithmic dementia classification and subsequent analysis of racial/ethnic differences in prevalent dementia for Black vs. White participants were consistent with the truth in the superpopulation, despite some lack of fit.

An important issue raised by recent work in algorithmic dementia classification is that existing algorithms differentially misclassify participants by race, which makes them unsuitable for use in racial/ethnic disparities research (Gianattasio et al. 2019). Revised algorithms with similar sensitivity and specificity by race have been developed, but they require specifying different neuropsychological exam score cutoffs by race/ethnicity (Gianattasio, Ciarleglio, and Power 2020). The race/ethnicity-specific results for algorithmic dementia classification presented in this chapter demonstrated differential misclassification since impairment tended to be underestimated in White participants and overestimated in Black and Hispanic participants on average.

Differences between ADAMS and HCAP sample characteristics, the potential for cohort effects from (i.e., ADAMS being 20 years older than HCAP) to impact associations between cognitive tests and other measures and cognitive impairment classification, results related to differential misclassification by race, and the lack of fit demonstrated for some variables in the synthetic HCAP datasets motivated a follow-up simulation study. In the next chapter, I describe

a simulation study assessing the impact of adjudicating subsets of the HCAP study and using information from those subsets as priors in the Bayesian latent class mixture model instead.

Chapter 5 Simulation Study: Refining Priors in the Bayesian Latent Class Mixture Model using Calibration Samples

The simulation studies in **Chapter 4** showed promising results for algorithmic dementia classification using the Bayesian latent class mixture model. Algorithmic dementia classification results closely replicated the truth in simulated datasets, and I was able to recover population-level inferences for dementia prevalence and prevalence ratios for White and Black participants. The model was not able to recover inferences for Hispanic participants, however, and there was some evidence for lack of fit for categorical and continuous variables overall.

I hypothesized that the lack of fit may be due to incompatibility between priors based on the ADAMS study and the observed HCAP data since the datasets were sampled differently and the studies were conducted nearly 20 years apart. Currently, there are no updated clinically adjudicated subsets of HRS to update prior distributions in the Bayesian latent class mixture model. Thus, I designed a simulation study to assess the impact of adjudicating subsets of HCAP to specify priors that would be better calibrated to the observed data. Gold-standard clinical adjudication is an expensive process, so I was interested in how much of the HCAP sample would need to be adjudicated to improve upon the results from Chapter 4 and whether different sampling strategies made an impact on results.

5.1: Simulation study outline

This simulation scenarios in this chapter are extensions to the simulation study design presented in Chapter 4. The updated simulation study flow diagram is presented in **Figure 5.1**.

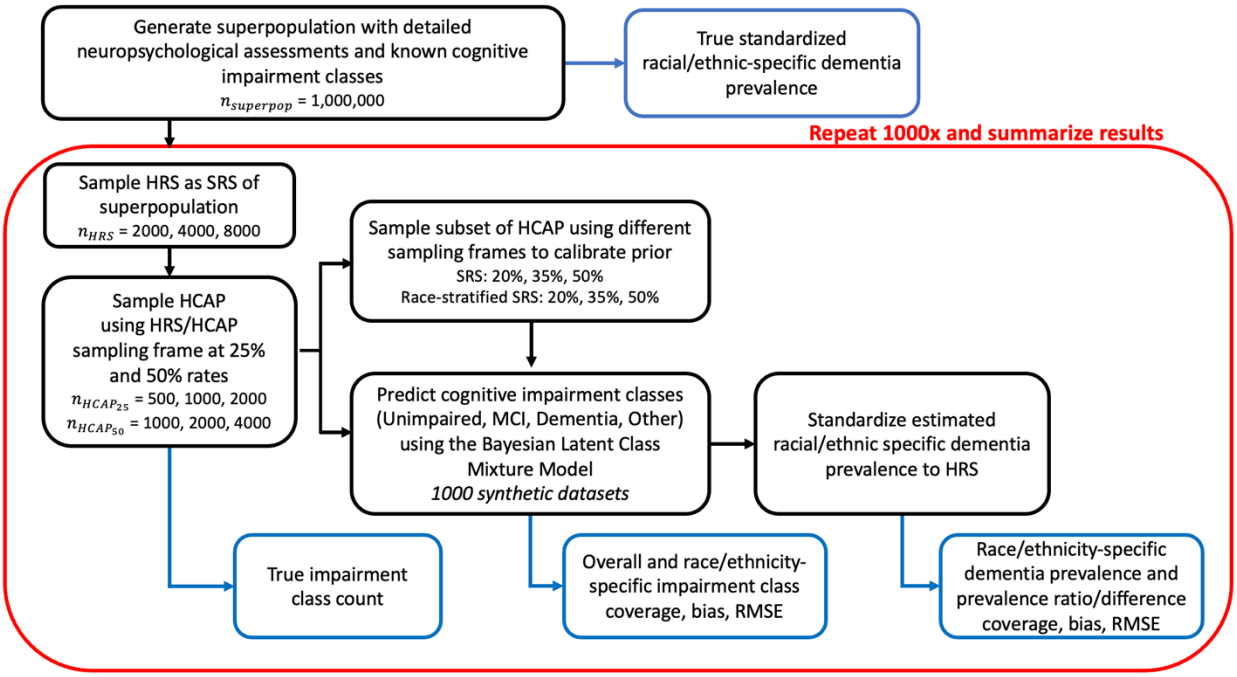


Figure 5.1: Simulation study flow diagram. Black boxes indicate analysis steps and blue boxes indicate calculation steps. The red box denotes the sequence of steps that are repeated 1000 times in the simulation study.

I used the same superpopulation that was constructed for the simulation studies in Chapter 4. Details for creating the superpopulation are presented in **Section 4.3:**. Prevalence of cognitive impairment overall and by race/ethnicity and racial/ethnic differences in prevalent dementia in the superpopulation were considered the truth in this simulation study as well.

The difference between the simulation study in this chapter and the one presented in Chapter 4 is that priors in this simulation will be specified based on information from clinically adjudicated subsets of HCAP instead of the ADAMS study. For each simulation run, I again sampled HRS studies as a simple random sample (SRS) from the superpopulation using three different sample sizes n_{HRS} = 2000, 4000, 8000. Then, I sampled HCAP studies using the sampling frame implemented in the real HRS/HCAP study—stratified random sampling by married/partnered status (Langa et al. 2020). I sampled HCAP studies at 25% and 50% rates

from each HRS study. After sampling HCAP, I flagged a subset of the HCAP sample as a clinically adjudicated subsample of participants for which cognitive impairment classes were known (referred to as the "HCAP calibration sample"). I flagged participants for the calibration sample at a rate of 20%, 35%, and 50% using SRS and race-stratified SRS for a total of 36 simulation scenarios (3 HRS sample sizes x 2 HCAP sampling proportions x 3 calibration sample proportions x 2 calibration sampling strategies). In SRS calibration samples, subsets of HCAP were flagged at the calibration sample rate regardless of the observations' race/ethnicity. In race-stratified SRS calibration samples, 60% of Black and Hispanic participants in HCAP were selected for the calibration sample and White participants were sampled to attain desired calibration sample sizes.

Then I proceeded with the same steps as the simulation study in Chapter 4: for each HCAP sample in each simulation run, I algorithmically classified cognitive impairment status and quantified estimation uncertainty by using the Bayesian latent class mixture model to create 1000 synthetic HCAP datasets and taking the mean and 95% credible interval of posterior predicted distributions of cognitive impairment class counts. Then, I age and sex-standardized race/ethnic specific estimates of dementia prevalence to HRS studies from which the HCAP study was sampled.

I assessed model performance in the same way as Chapter 4 simulation studies: I assessed interval estimate coverage, bias, and RMSE of predicted cognitive impairment class proportions overall and by race/ethnicity across 1000 simulation runs At a population-representative level, I assessed interval coverage, bias, and RMSE of standardized race/ethnic specific estimates of dementia prevalence in HRS; dementia prevalence ratios and differences for Black vs. White and Hispanic vs. White participants in HRS; and interval coverage, bias, and RMSE of dementia

prevalence ratios and differences for Black vs. White and Hispanic vs. White participants in HRS across 1000 simulation runs.

5.2: Dataset preparation

I used the same HCAP and HRS datasets prepared for the simulation study in Chapter 4. Detailed data preparation steps are described in **Sections 4.2.1:** and **4.2.3:**, respectively.

5.3: Specifying prior distributions

Like the simulation study in Chapter 4, the general strategy for specifying prior distributions in this simulation study is identical to the strategies described in Chapter 3 (see **Section 3.4.1:**) However, parameters for prior distributions in this simulation were estimated in HCAP calibration samples instead of in ADAMS. Further, parameters stored for latent cognitive impairment class prediction were based on fitting models **(4.1)-(4.5)** in the HCAP calibration subsample. The updated overview of parameter storage steps for the priors in this simulation study is (1) sample a calibration subset from the HCAP sample (2) bootstrap the HCAP calibration sample, (3) store parameter estimates characterizing effects of covariates on cognitive impairment class membership, contingency cell counts, and effects of contingency cell membership on continuous covariates, and (4) repeat the process 1,000 times (once for each synthetic dataset created in the Bayesian latent class mixture model) to represent both sampling variability and estimation uncertainty in model parameters. Details for specifying prior distributions for each component of the model are described in **Section 3.4.1:**.

5.4: Pre-simulation study tuning of the Bayesian latent class mixture model

As discussed in Chapter 4, all Bayesian workflow steps (prior predictive checks, convergence diagnostics, prior predictive checks) would be infeasible to perform for each of the 1000 iterations of each simulation study scenario. I again performed all Bayesian workflow steps for one iteration of each simulation scenario and set tuning parameters in the simulation study based on results from those runs. Selected results for Bayesian workflow steps are presented below.

5.4.1: Prior predictive checks

I performed prior predictive checks for distributions of contingency cell counts and continuous variables stratified by cognitive impairment classes. For each simulation scenario, I generated 1000 synthetic HCAP datasets by drawing from prior distributions only. Prior predictive distributions of contingency cell counts and normalized MMSE for the scenario with $n_{HRS} = 8000$, 50% HCAP sampling proportion, and 50% race-stratified SRS calibration sample ($n_{calibration} = 2000$) are presented in **Figure 5.2** and **Figure 5.3**, respectively.

Prior predictive distributions for the 1000 synthetic contingency cell counts were mostly centered around true counts, with some lack of fit for Black participants without stroke history in the Unimpaired, MCI, and Other cognitive impairment groups and for White participants without stroke history in the Unimpaired and MCI groups. Prior predictive distributions of normalized MMSE were slightly wider (more variable) than true distributions, ensuring that the full range of values was captured by features encoded in the priors. Note that different from prior predictive checks in the previous chapter, these continuous variable densities are not smooth because half of the prior predictive distribution is based on draws from normal distributions and the other half was considered adjudicated, so we know the true values. Prior predictive distributions for the

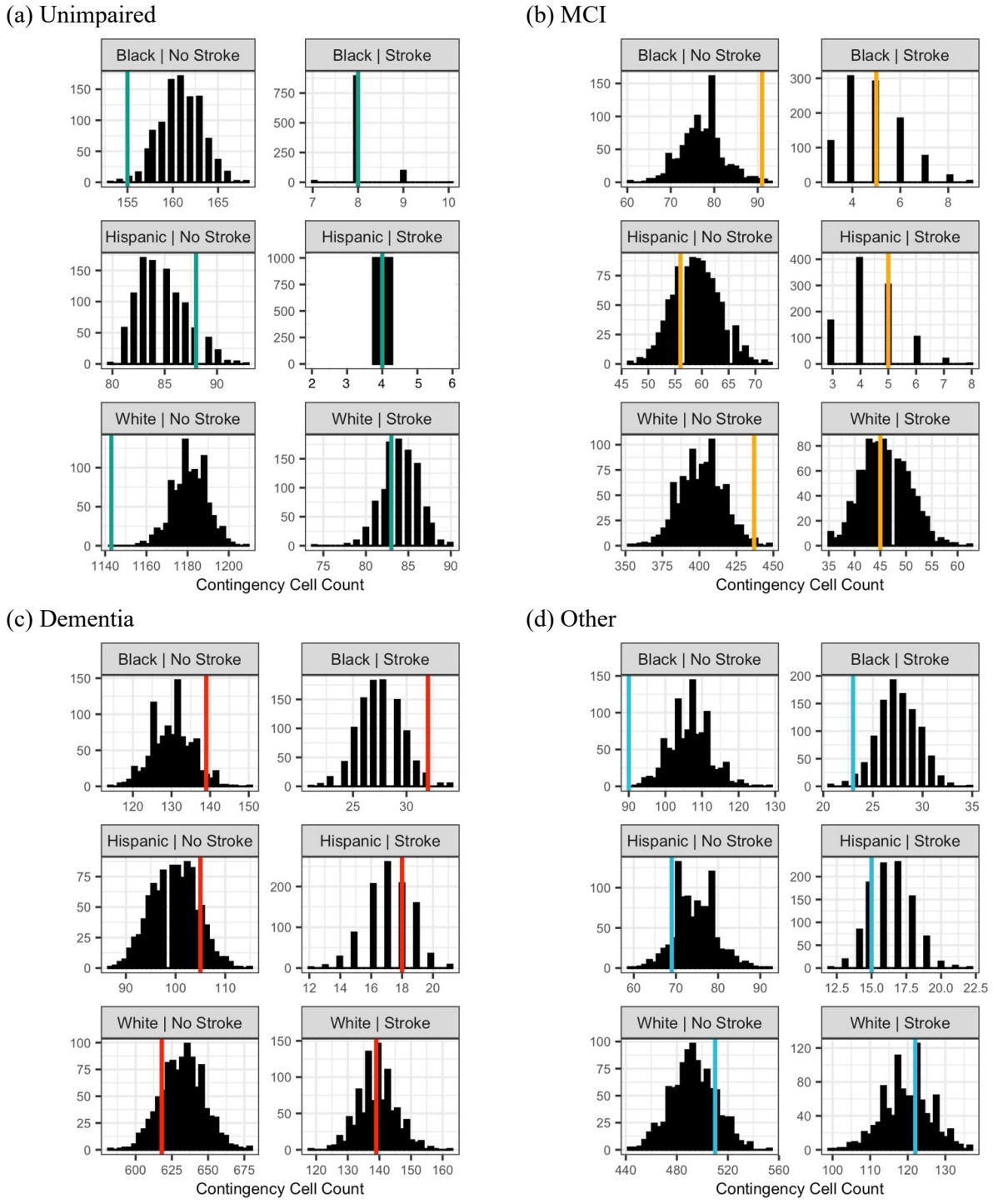


Figure 5.2: Prior predictive distributions of contingency cell counts for the simulation scenario with $n_{HRS} = 8000$, 50% HCAP sampling proportion, and 50% race-stratified SRS calibration sample used for the prior stratified by impairment group: (a) Unimpaired, (b) MCI, (c) Dementia, (d) Other. Distributions are based on 1000 synthetic HCAP datasets of size $n_{HCAP_{50}} = 4000$. Colored vertical lines in each panel represent true impairment group-specific contingency cell counts.

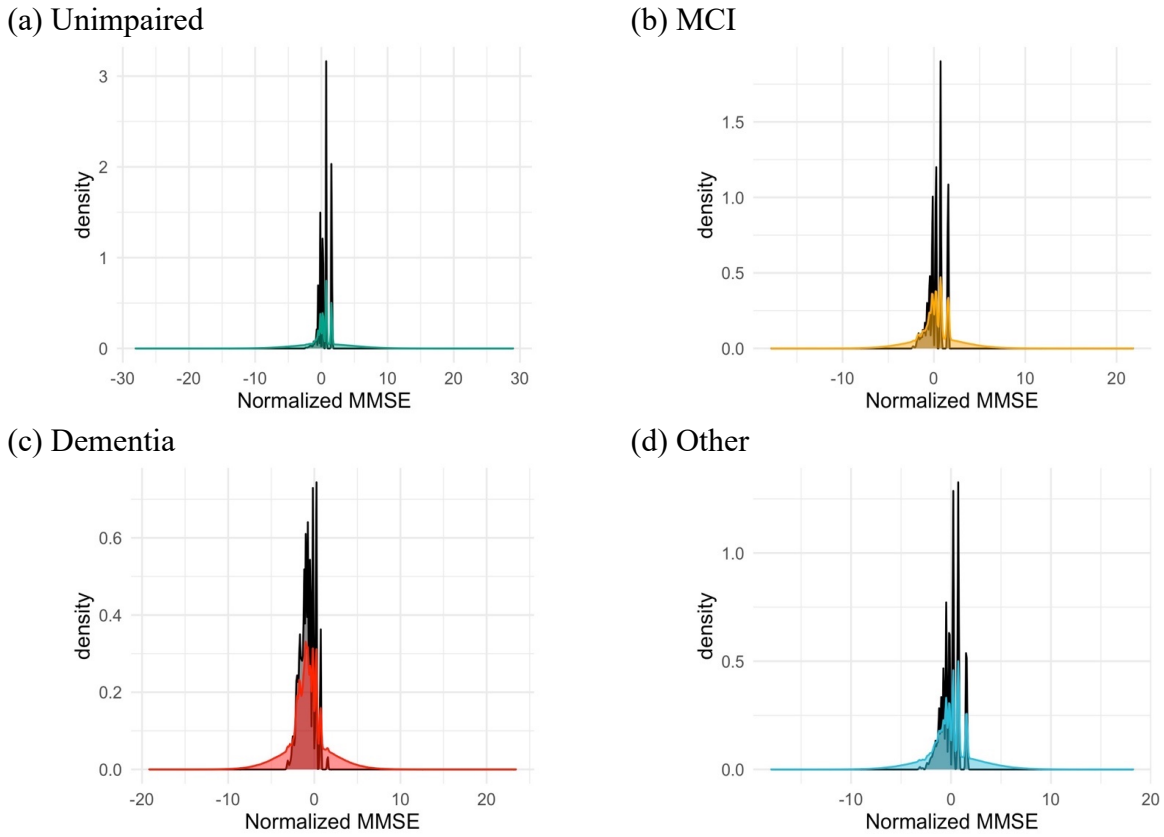


Figure 5.3: Prior predictive distributions of normalized MMSE (colored densities) for the simulation scenario with $n_{HRS} = 8000$, 50% HCAP sampling proportion, and 50% race-stratified SRS calibration sample used for the prior stratified by impairment group: (a) Unimpaired, (b) MCI, (c) Dementia, (d) Other. Distributions are based on 1000 synthetic HCAP datasets of size $n_{HCAP_{50}} = 4000$. Gray densities are true impairment group-specific distributions of MMSE.

remaining continuous variables and all other simulation scenarios were similarly more variable than true distributions. Code for producing prior predictive checks including .gif files for all simulation scenarios can be found in the associated GitHub repository.

5.4.2: Assessing model convergence

I produced MCMC chains for each parameter in this analysis and for each simulation scenario, but I primarily monitored cognitive impairment class proportion chains and impairment group-specific variances for continuous variables. All chains converged in all simulation scenarios.

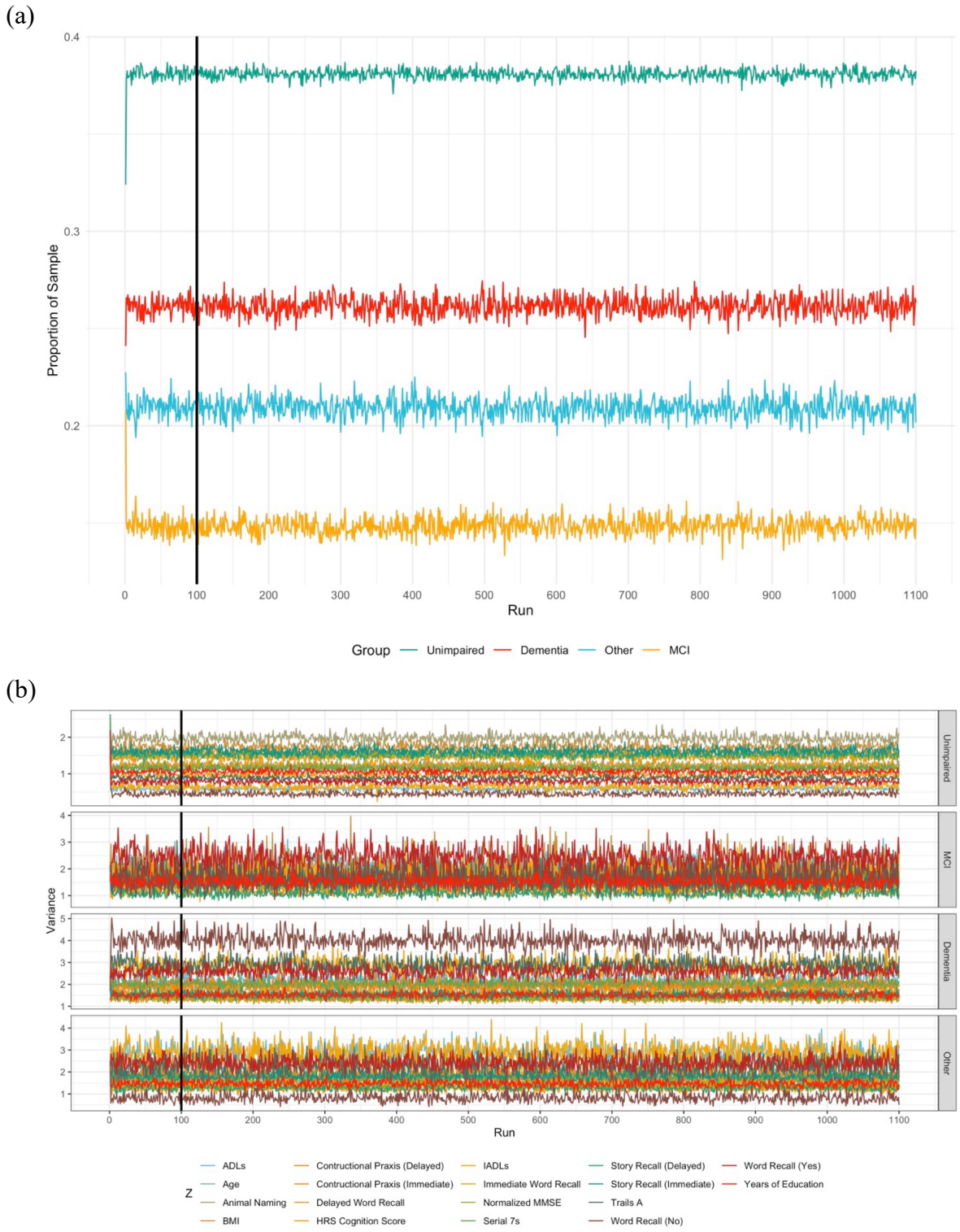


Figure 5.4: MCMC chains of proportions of (a) impairment class membership and (b) impairment group-specific variances of continuous variables for the simulation scenario with $n_{HRS} = 8000$, 50% HCAP sampling proportion, and 50% race-stratified SRS calibration sample used for the prior based on 1000 synthetic HCAP datasets of size $n_{HCAP_{50}} = 4000$. Black vertical lines mark the end of the burn-in period (100 runs).

MCMC chains for the simulation scenario with $n_{HRS} = 8000$, 50% HCAP sampling proportion, and 50% race-stratified SRS calibration sample are presented in **Figure 5.4**. Code for producing MCMC chains for the other simulation scenarios can be found in the associated GitHub repository.

In this simulation study, all chains were initiated in the "random" state with equal proportions of group membership for all cognitive impairment classes (25% Unimpaired, 25% MCI, 25% Dementia, 25% Other) since model stability was established in the illustrative example in **Section 3.4.2.2:**.

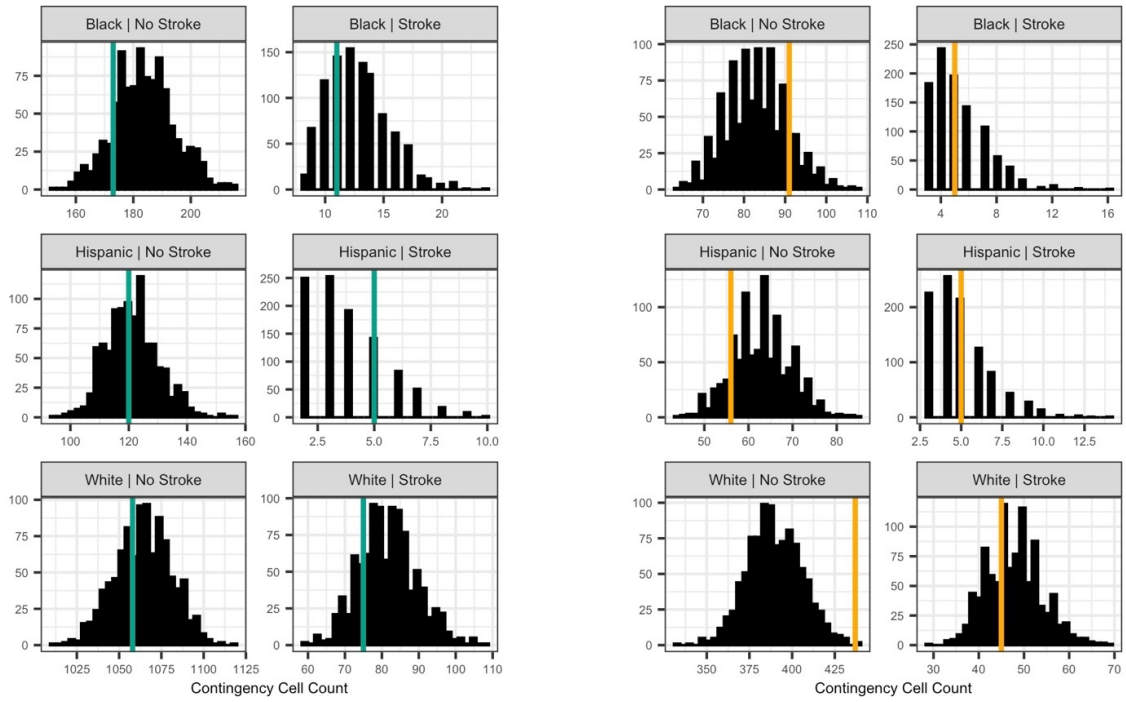
5.4.3: Posterior predictive checks

I assessed posterior distributions of contingency cell counts and median and skew for continuous variables for all simulation scenarios. All posterior statistics were stratified by predicted cognitive impairment class. Posterior predictive distributions for cell counts and median and skew of normalized MMSE in the simulation scenario with $n_{HRS} = 8000$, 50% HCAP sampling proportion, and 50% race-stratified SRS calibration sample are presented in **Figure 5.5-Figure 5.7.** Posterior predictive distributions of contingency cell counts were centered around true counts for every cell in every cognitive impairment class except White participants without stroke history which only capture the true count in the tail of the distribution (**Figure 5.5**).

Posterior distributions of median normalized MMSE were centered around observed medians except in the Unimpaired group which was only off by about 2 points (on a scale of 0-100) for Unimpaired and Other groups (**Figure 5.6**). Posterior distributions of normalized MMSE skew were roughly centered at true measures of skew except in the Unimpaired group,

(a) Unimpaired

(b) MCI



(c) Dementia

(d) Other

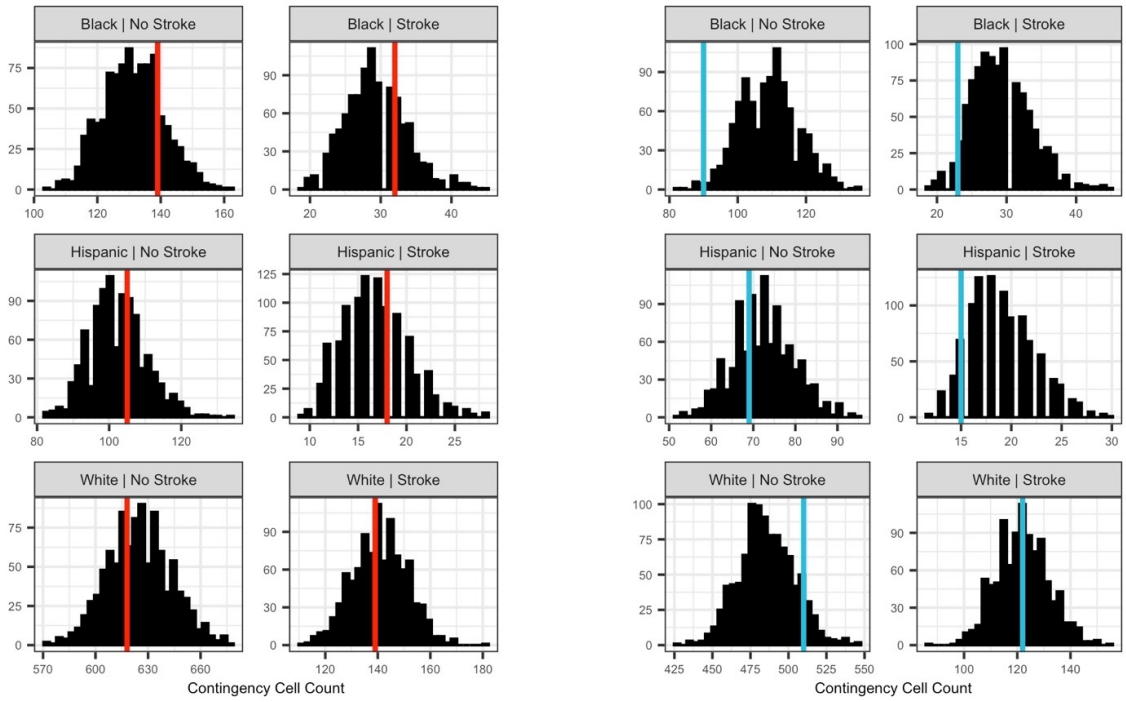


Figure 5.5: Posterior predictive distributions of contingency cell counts for the simulation scenario with $n_{HRS} = 8000$, 50% HCAP sampling proportion, and 50% race-stratified SRS calibration sample used for the prior stratified by impairment group: (a) Unimpaired, (b) MCI, (c) Dementia, (d) Other. Distributions are based on 1000 synthetic HCAP datasets of size $n_{HCAP_{50}} = 4000$. Colored vertical lines in each panel represent true impairment group-specific contingency cell counts.

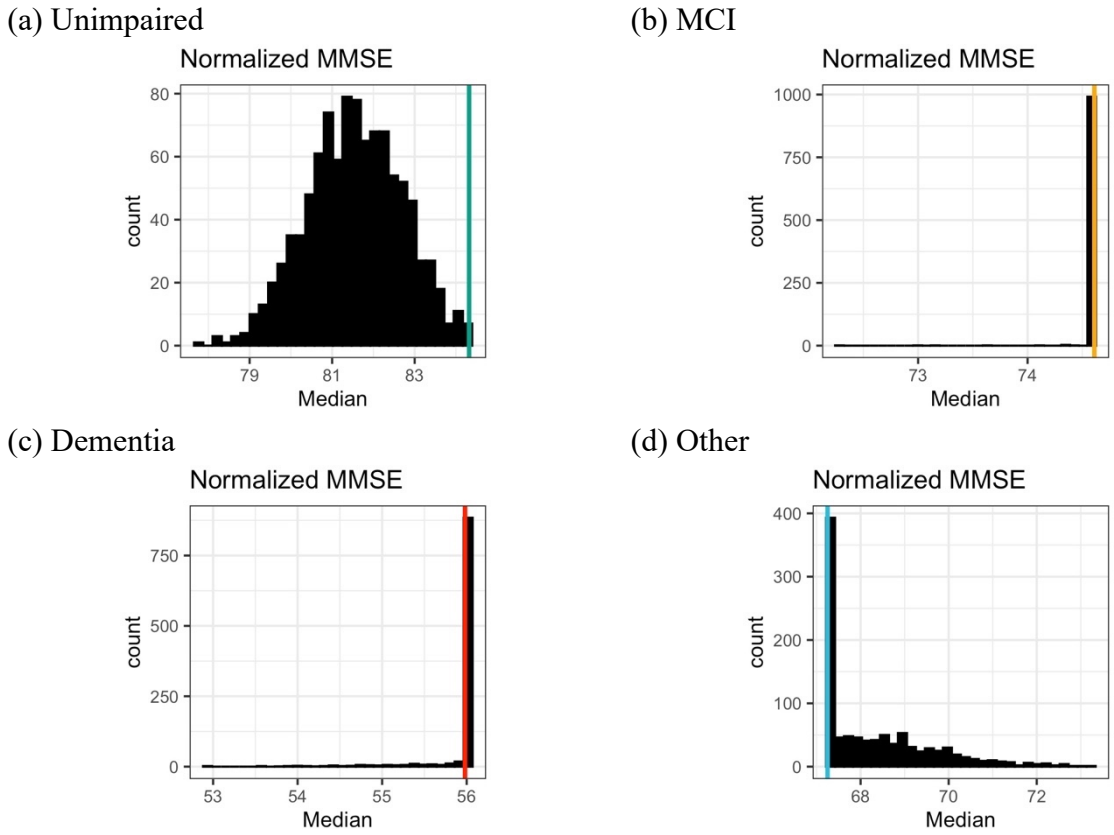
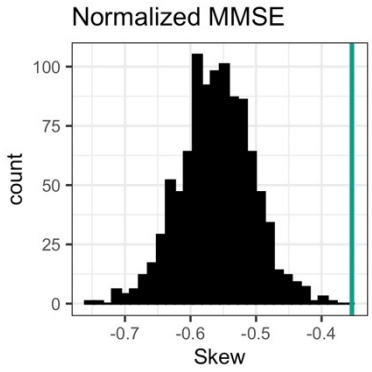


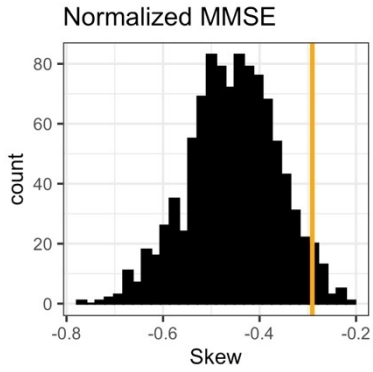
Figure 5.6: Posterior predictive distributions of median normalized MMSE for the simulation scenario with $n_{HRS} = 8000$, 50% HCAP sampling proportion, and 50% race-stratified SRS calibration sample used for the prior stratified by impairment group: (a) Unimpaired, (b) MCI, (c) Dementia, (d) Other. Distributions are based on 1000 synthetic HCAP datasets of size $n_{HCAP_{50}} = 4000$. Colored vertical lines in each panel represent true impairment group-specific medians of normalized MMSE.

demonstrating that synthetic variables were able to reproduce skewness in these measures (**Figure 5.7**). Posterior predictive distributions of medians and skewness for the remaining continuous variables for this simulation scenario are displayed in **Appendix Figure D.3** and **Appendix Figure D.4**, respectively. Posterior predictive distributions for other simulation scenarios were similar and code for producing figures is available on the associated GitHub Repository.

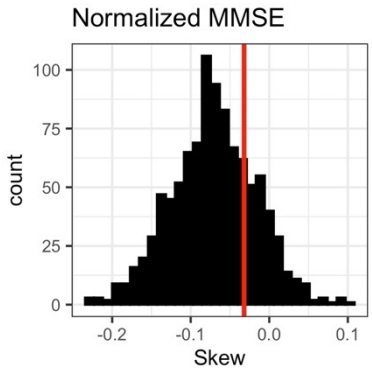
(a) Unimpaired



(b) MCI



(c) Dementia



(d) Other

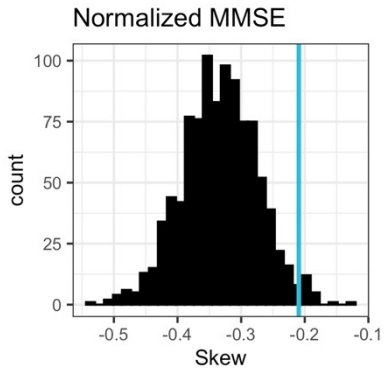


Figure 5.7: Posterior predictive distributions of normalized MMSE skew for the simulation scenario with $n_{HRS} = 8000$, 50% HCAP sampling proportion, and 50% race-stratified SRS calibration sample used for the prior stratified by impairment group: (a) Unimpaired, (b) MCI, (c) Dementia, (d) Other. Distributions are based on 1000 synthetic HCAP datasets of size $n_{HCAP_{50}} = 4000$. Colored vertical lines in each panel represent true impairment group-specific skew of normalized MMSE.

5.4.4: Algorithmic dementia classification

Figure 5.8 shows 95% credible intervals of participant (a) counts and (b) proportions in each cognitive impairment class across 1000 synthetic HCAP samples for the simulation scenario with $n_{HRS} = 8000$, 50% HCAP sampling proportion, and 50% race-stratified SRS calibration sample. Estimates are very tight, likely due to the large sample used for the prior ($n_{calibration} = 2000$). Credible intervals for Dementia and Other groups captured the true cognitive impairment class count/proportion, and point estimates were identical to true counts/proportions.

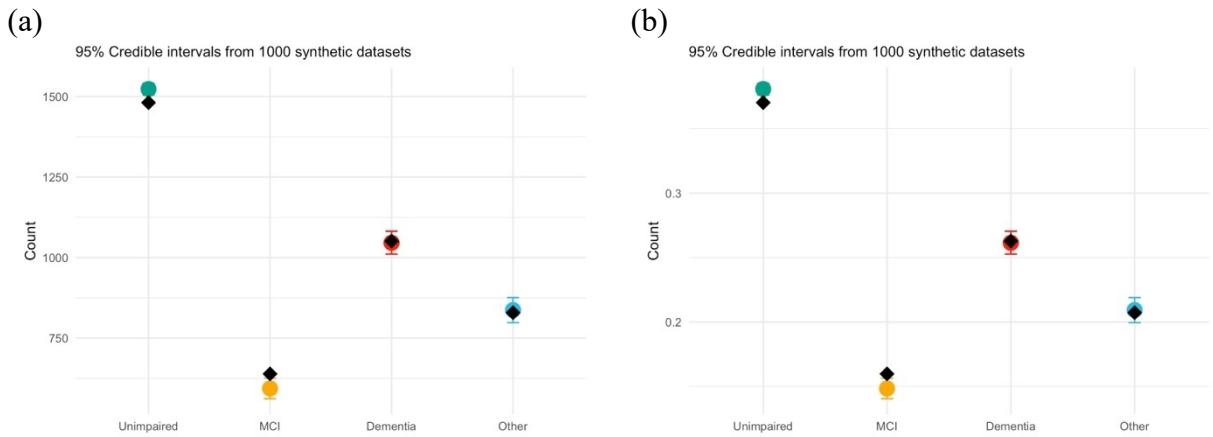


Figure 5.8: 95% interval estimates of participant counts (a) and proportions (b) within each impairment group across 1000 synthetic HCAP datasets for the simulation scenario with $n_{HRS} = 8000$, 50% HCAP sampling proportion, and 50% race-stratified SRS calibration sample used for the prior. Black diamonds are group-specific true counts/proportions.

Intervals for Unimpaired and MCI groups did not capture true counts/proportions, however, the largest discrepancy in mean count was in the Unimpaired group where the model underestimated the count by only 60 people on average (about a 1.5% error).

Results for model tuning were satisfactory, so I proceeded with the simulation study to assess performance of the Bayesian latent class mixture model in these simulation scenarios across repeated runs in the various simulation scenarios.

5.5: Simulation study results

To conveniently compared results from using different prior distributions, all result plots for this simulation study include results from Chapter 4 simulations where ADAMS was used as a prior.

5.5.1: Algorithmic dementia classification

Mean and 95% interval estimates for cognitive impairment class proportions summarized across 1000 simulation runs are presented in **Figure 5.9** by HRS sample size, HCAP sampling

proportion, and sample used for the prior. Like Chapter 4 summaries, mean cognitive impairment class proportions were obtained by averaging means of posterior predictive distributions of participant counts in each cognitive impairment class across simulation runs and dividing by HCAP sample size. Upper and lower limits of cognitive impairment class proportions were obtained by averaging 97.5% and 2.5% percentiles of posterior predictive distributions of participant counts, respectively, across simulation runs and dividing by HCAP sample size. Since this simulation study assesses the impact of adjudicating subsets of HCAP to better calibrate prior distributions, results for adjudicating 100% of the HCAP sample are presented for comparison.

Adjudicating 100% of HCAP replicated true impairment class proportions in the superpopulation, as expected due to properties of SRS sampling. In contrast to results using ADAMS priors which increased in precision with increasing sample size, increased sample sizes improved point estimates and precision when priors were specified based on HCAP calibration subsets. This is expected because the size HCAP calibration subset was a proportion of the HCAP sample. Thus, smaller HCAP samples had smaller prior samples that produced imprecise estimates. Point estimates were furthest away from true impairment class proportions in the Dementia and Other impairment groups for the scenario with the smallest sample sizes, 25% HCAP sampling proportion with $n_{HRS} = 2000$ ($n_{HCAP} = 500$). Otherwise, point estimates are close to true impairment class proportions in all impairment groups and improved the most when increasing from $n_{HRS} = 2000$ to $n_{HRS} = 4000$ but were similar between $n_{HRS} = 4000$ and $n_{HRS} = 8000$.

Average interval estimates included the truth in all scenarios except in the Unimpaired group with the largest sample sizes where estimates were very tight. Estimated precision for the

Unimpaired group increased drastically with sample size which is expected because the Unimpaired group is the largest subgroup. Interval estimates shrunk to such a degree for the Unimpaired group, though, that in the scenario with the largest samples sizes, 50% HCAP sampling proportion with $n_{HRS} = 8000$ ($n_{HCAP} = 4000$), average interval estimates no longer included the true proportion. In the larger samples, point estimates were similar to results using ADAMS as the prior.

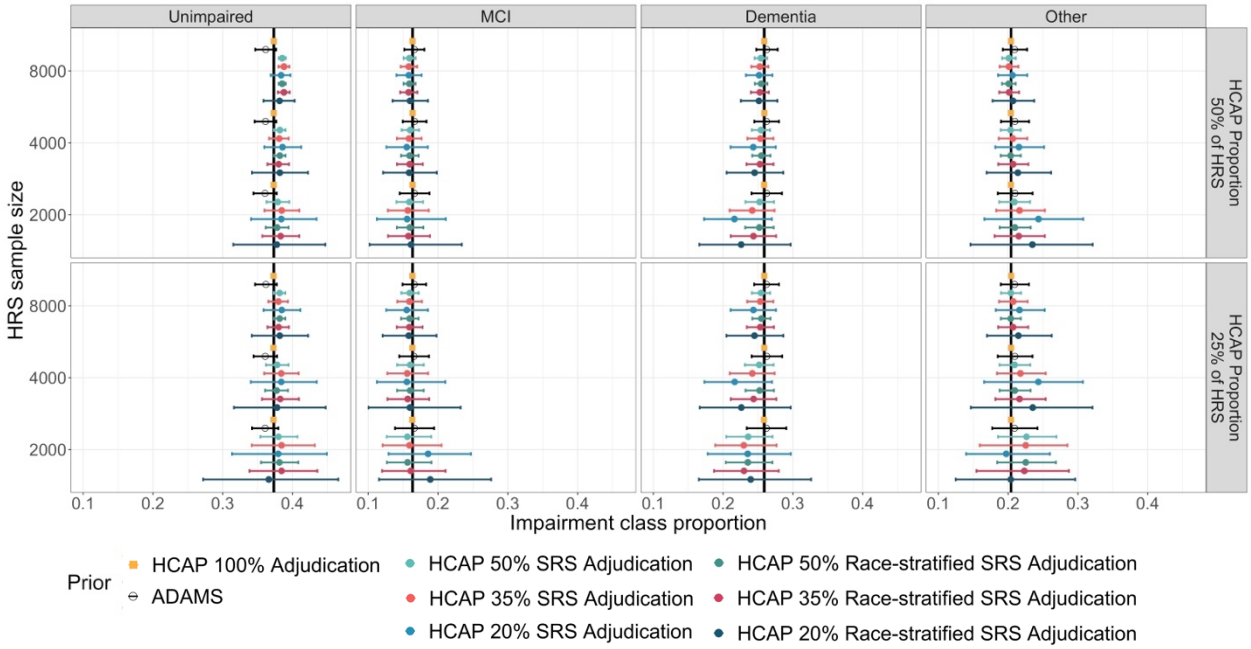


Figure 5.9: Mean and 95% interval estimate for cognitive impairment class proportions by HRS sample size, HCAP sampling proportions, and sample used for the prior averaged across 1000 simulation runs. Black vertical lines denote true impairment class proportions in the superpopulation. Yellow squares denote impairment class proportions obtained from adjudicating 100% of the HCAP sample.

Figure 5.10 shows 95% interval coverage by HRS sample size, HCAP sampling proportion, and sample used for the prior. In most simulation scenarios and cognitive impairment groups, interval estimates achieved at least 75% coverage, and coverage was less than simulations using the ADAMS prior in all scenarios. The lack of coverage was most pronounced in the Unimpaired

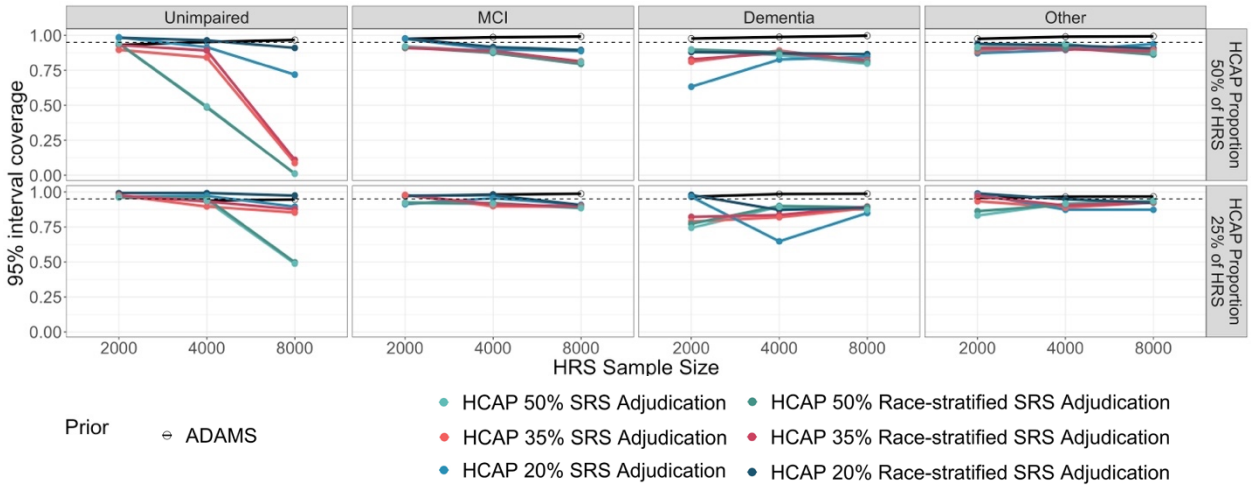


Figure 5.10: 95% interval coverage of true cognitive impairment class proportion by cognitive impairment class, HRS sample size, HCAP sampling proportion, and sample used for the prior across 1000 simulation runs. Dashed horizontal line denotes nominal coverage of 95%.

group in scenarios with the largest HCAP calibration samples. This lack of nominal coverage is a result of the increased precision in the prior sample at larger sample sizes. Bias and RMSE provide additional context for this result.

Bias and percent bias in cognitive impairment class proportions across 1000 simulation runs are presented in

Figure 5.11**(a)** and

Figure 5.11**(b)**. Like Chapter 4 summaries, bias was calculated by averaging mean posterior cognitive impairment class proportions across runs and subtracting the cognitive impairment class proportion in the superpopulation (e.g., for the Unimpaired group, $\text{bias}_{Unimpaired} = \bar{\hat{\phi}}_{Unimpaired} - \phi_{Unimpaired}$, where $\hat{\phi}$ is the mean of the posterior predictive distribution of proportion of Unimpaired participants). Percent bias is bias divided by the true cognitive impairment class proportion in the superpopulation. We see that bias improved with increasing HRS samples sizes and HCAP sampling proportions and was minimal in simulation scenarios with the largest sample sizes. In scenarios with the largest

sample sizes, bias in simulations that used HCAP calibration priors was less than in simulations with ADAMS priors. Models with HCAP calibration priors consistently underestimated the

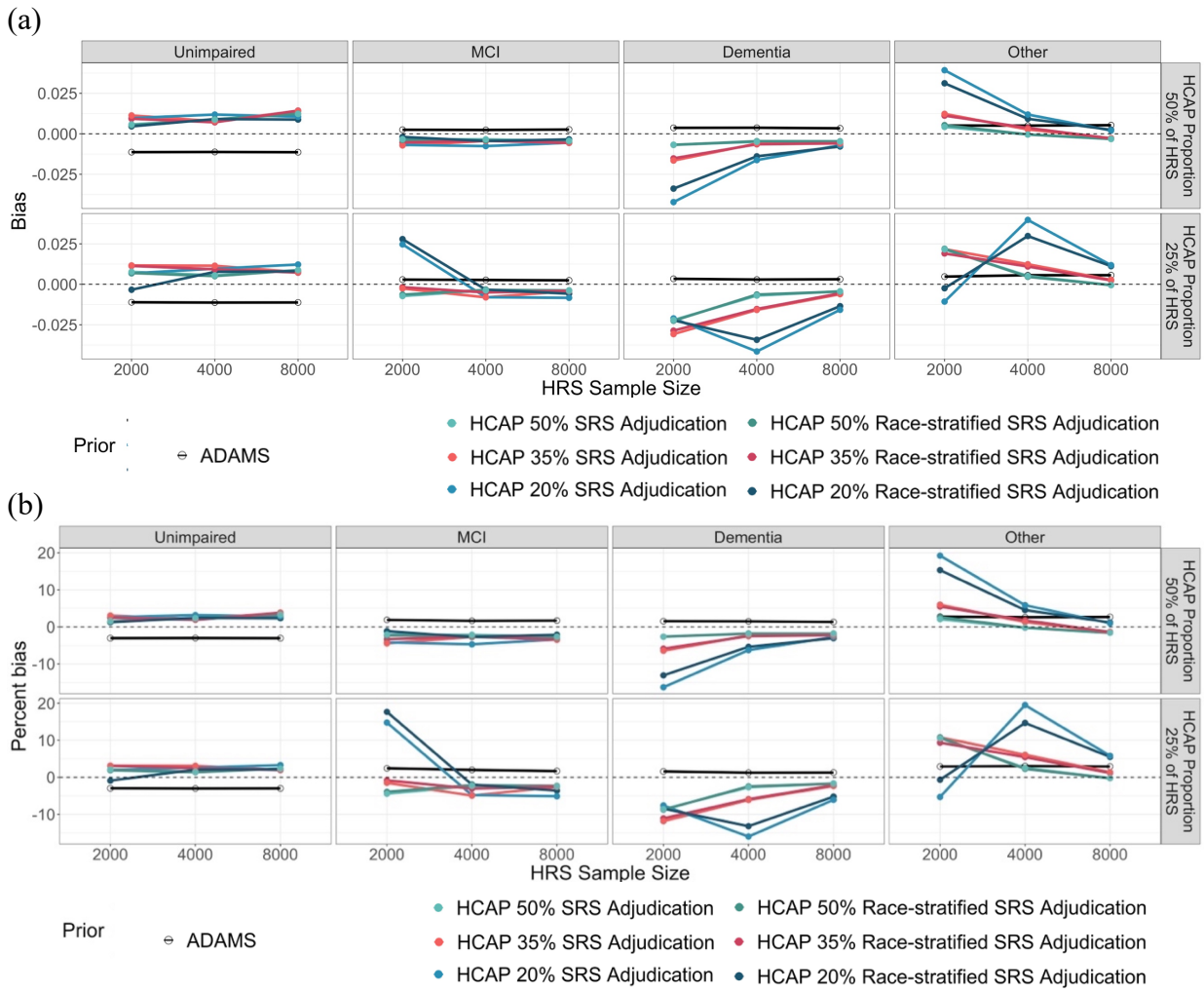


Figure 5.11: Bias (a) and percent bias (b) in cognitive impairment class proportions by HRS sample size, HCAP sampling proportion, and sample used for the prior across 1000 simulation runs. Dashed horizontal line denotes no bias.

proportion of participants in the Unimpaired and Other groups and overestimated proportions of participants in the MCI and Dementia groups.

Figure 5.12 shows RMSE for cognitive impairment class proportions across simulation runs. RMSE decreased for all cognitive impairment groups across simulation scenarios but was larger than for scenarios that used ADAMS as a prior. Since we observed that bias was less for

these scenarios compared to simulations with ADAMS priors, this indicates that estimates using HCAP calibration samples as the prior are more variable. RMSE for scenarios with 50% HCAP

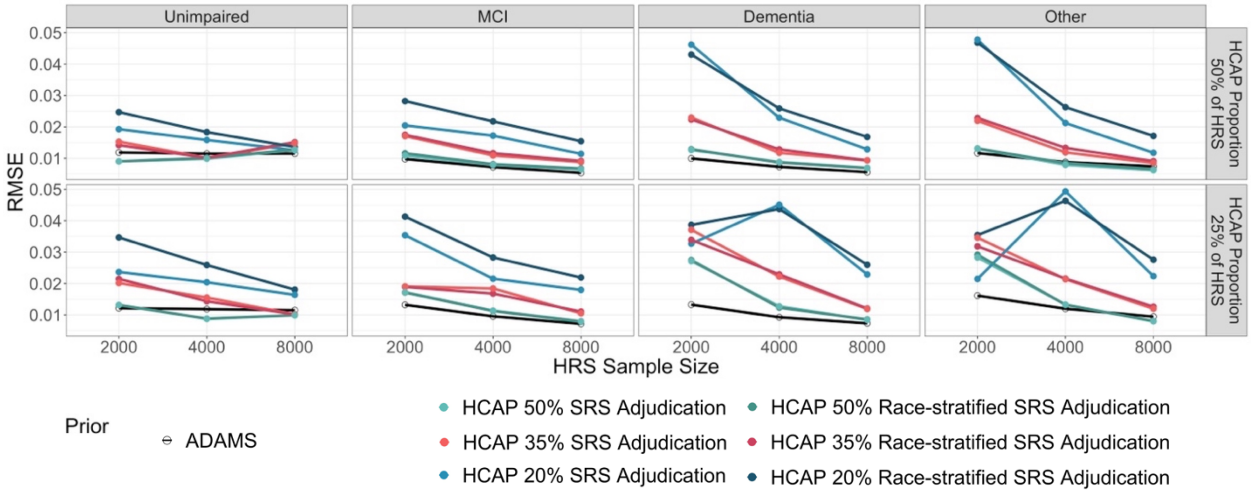


Figure 5.12: Root mean square error (RMSE) for estimated cognitive impairment class proportions by HRS sample size, HCAP sampling proportion, and sample used for the prior across 1000 simulation runs.

calibration samples were similar to scenarios with ADAMS priors.

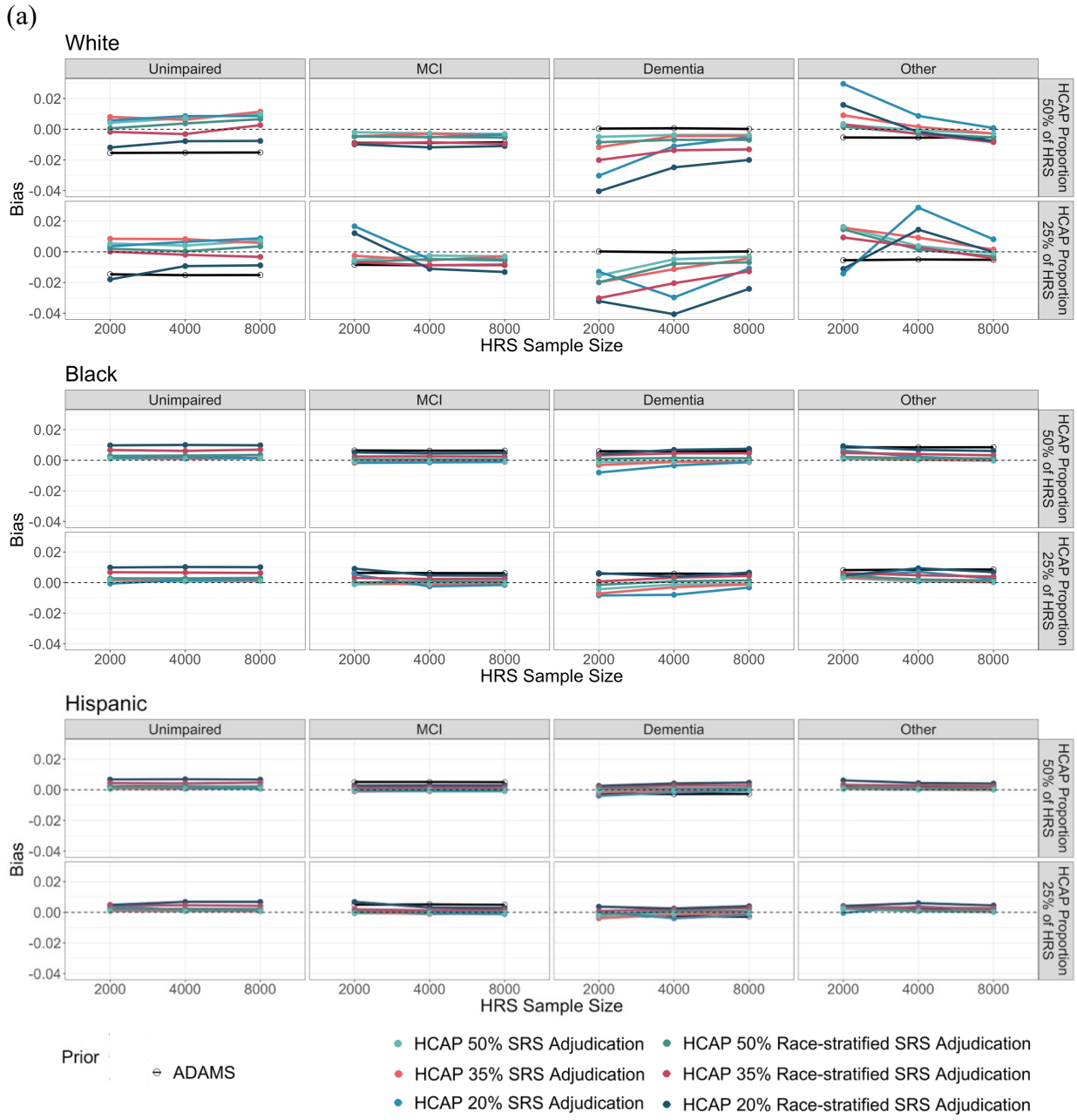
Race-stratified bias (

Figure 5.13**a**) and percent bias (

Figure 5.13**b**) plots show less bias for Black and Hispanic participants compared to White participants when HCAP calibration samples are used to specify priors. Bias was relatively stable with increasing sample size for Black and Hispanic participants but improved more noticeably for White participants. Compared to simulations using ADAMS priors, there was less bias for White participants in the Unimpaired group and Black and Hispanic participants in MCI, Dementia, and Other groups when using HCAP calibration samples to specify priors. There is virtually no bias across race/ethnicities and cognitive impairment classes in the scenarios using 50% SRS or race-stratified SRS HCAP calibration samples for the prior.

RMSE was also lower for Black and Hispanic participants compared to White participants in this simulation study. Race/ethnicity specific RMSE was lowest for Black and Hispanic participants across cognitive impairment groups in the scenarios using 50% SRS or race-stratified SRS HCAP calibration samples for the prior (

Figure 5.14).



(b)

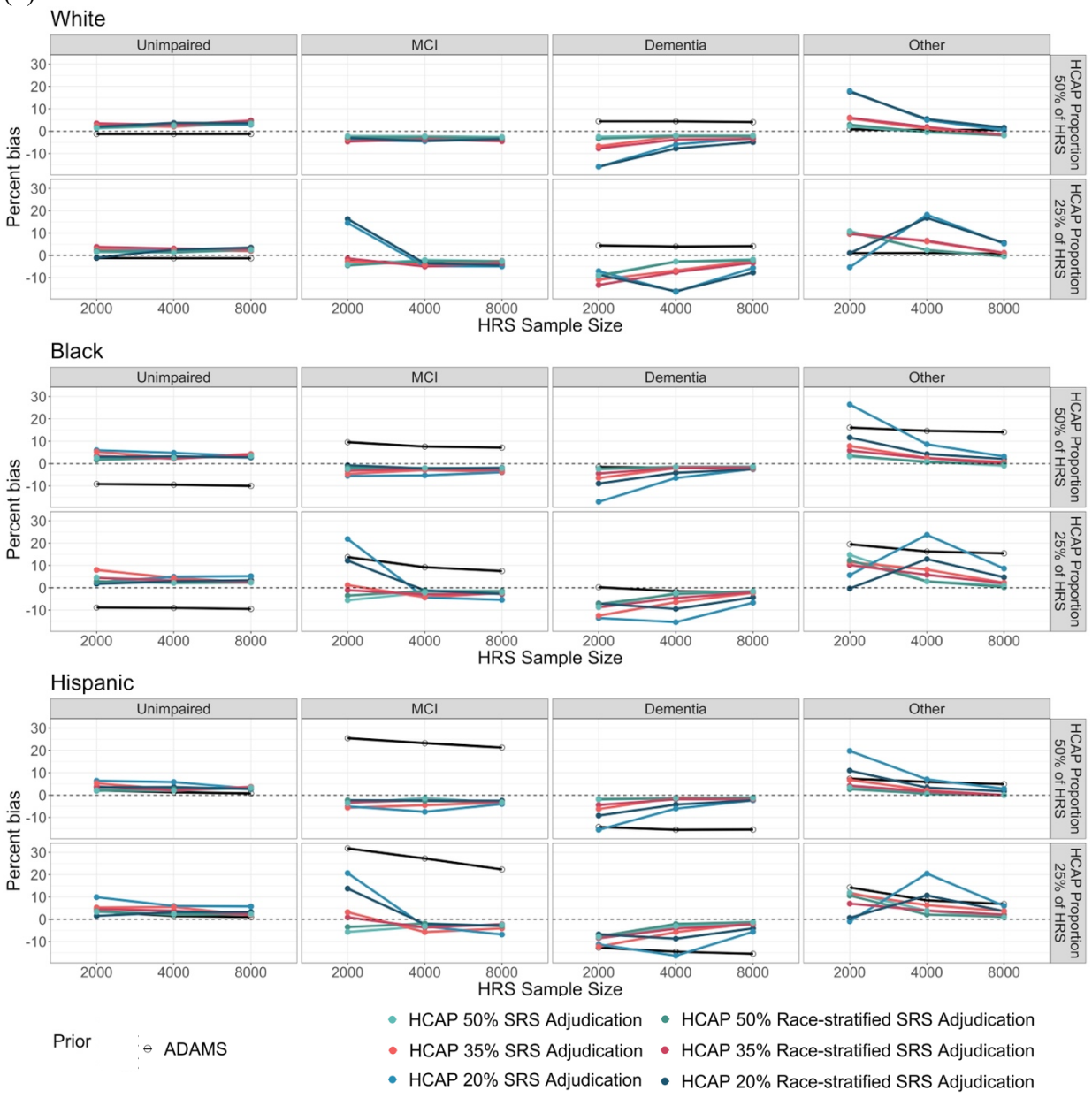

White
Black
Hispanic
Unimpaired
MCI
Dementia
Other
HCAP Proportion 50% of HRS
HCAP Proportion 25% of HRS
Percent bias
HRS Sample Size
2000
4000
8000
Prior
ADAMS
HCAP 50% SRS Adjudication
HCAP 35% SRS Adjudication
HCAP 20% SRS Adjudication
HCAP 50% Race-stratified SRS Adjudication
HCAP 35% Race-stratified SRS Adjudication
HCAP 20% Race-stratified SRS Adjudication

Figure 5.13: Race/ethnicity-specific bias (a) and percent bias (b) in cognitive impairment class proportions by HRS sample size, HCAP sampling proportion, and sample used for the prior across 1000 simulation runs. Dashed horizontal line denotes no bias.

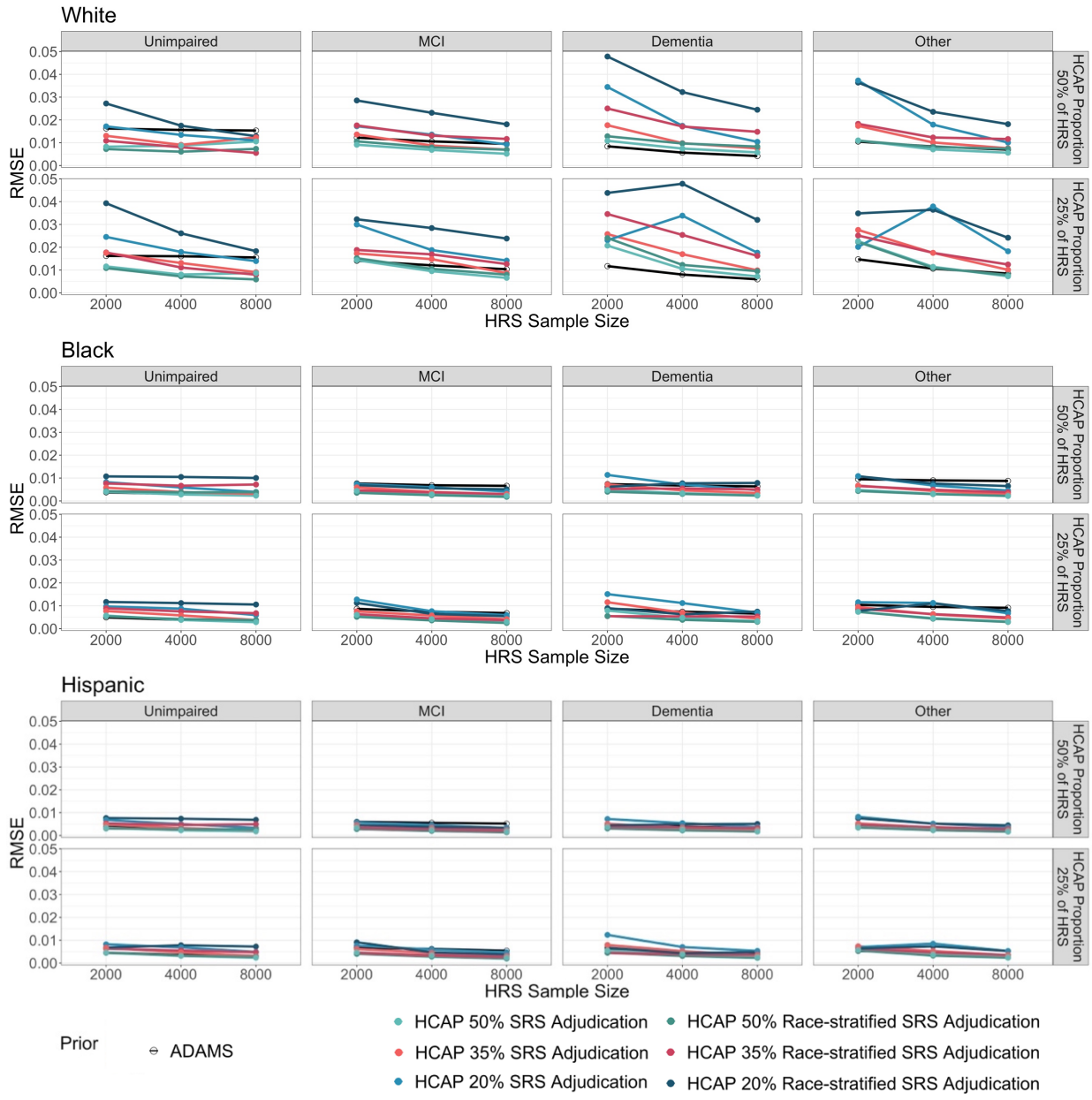


Figure 5.14: Race/ethnicity-specific RMSE for estimated cognitive impairment class proportions by HRS sample size, HCAP sampling proportion, and sample used for the prior across 1000 simulation runs.

5.5.2: Inference in a population-representative study

Algorithmic dementia classification results showed reduced bias compared to analyses using ADAMS priors for Black and Hispanic participants. The most noticeable improvements, however, were in population-level inferences for Hispanic participants.

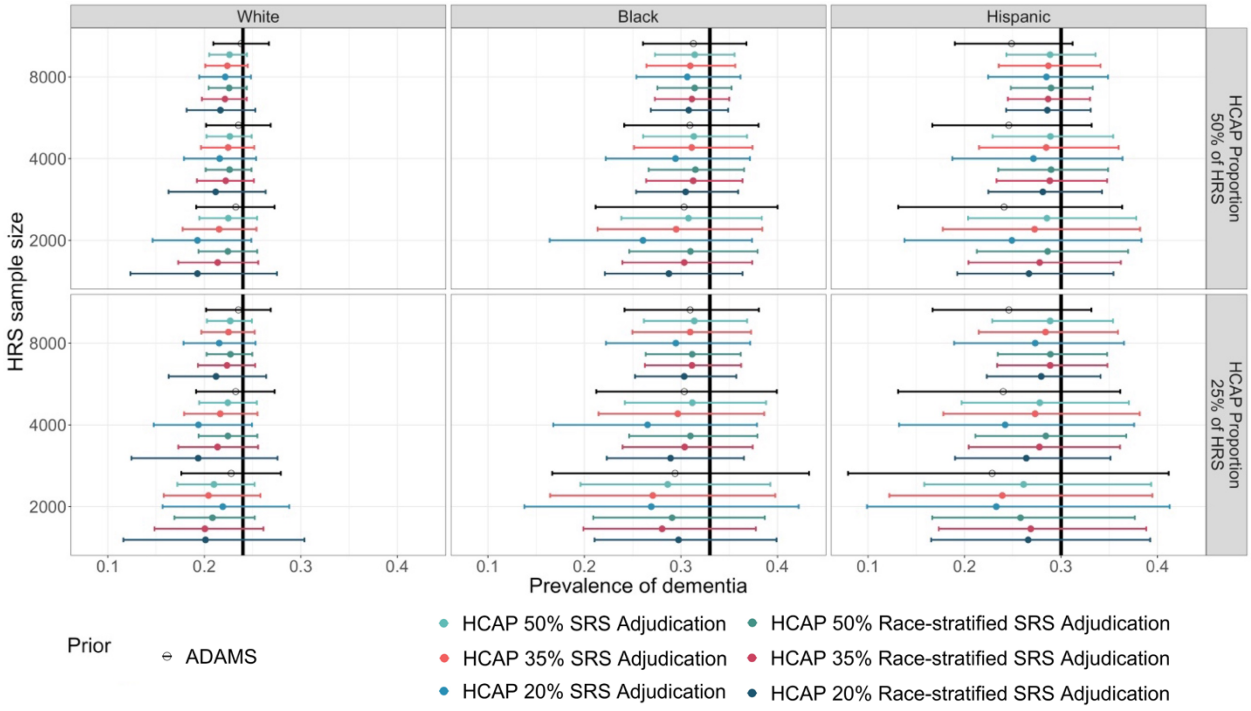


Figure 5.15: Mean and 95% interval estimate for race-specific age and sex-standardized dementia prevalence by HRS sample size, HCAP sampling proportion, and sample used for the prior across 1000 simulation runs. Prevalence was standardized to synthetic HRS samples at each simulation run. Vertical lines denote true race-specific dementia prevalence in the superpopulation.

Figure 5.15 shows mean and 95% interval estimates for age- and sex-standardized race/ethnicity-specific dementia prevalence by HRS sample size, HCAP sampling proportion, and sample used for the prior across 1000 simulation runs. Point estimates are closer to the truth with increased sample sizes for all race/ethnicities. Results for White and Black participants are similar to simulations that used ADAMS as a prior. For Hispanic participants, however, results were noticeably more accurate when HCAP calibration samples were used for the prior instead. In the scenario with the largest sample sizes where the prevalence of dementia among Hispanics

using the model with ADAMS priors was estimated as 25.0% when the truth in the superpopulation was 30.2%, the model using a prior based on a 50% race-stratified SRS calibration sample estimated 29.0% prevalence of dementia.

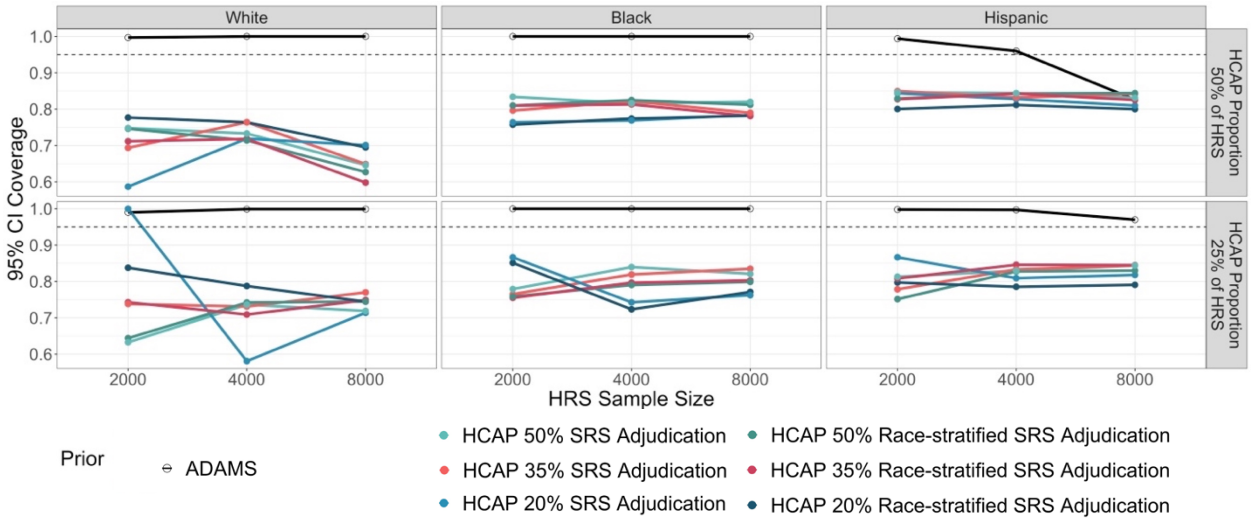


Figure 5.16: 95% interval coverage of true prevalence of dementia by race/ethnicity, HRS sample size, HCAP sampling proportion, and sample used for the prior across 1000 simulation runs. Dashed horizontal line denotes nominal coverage of 95%.

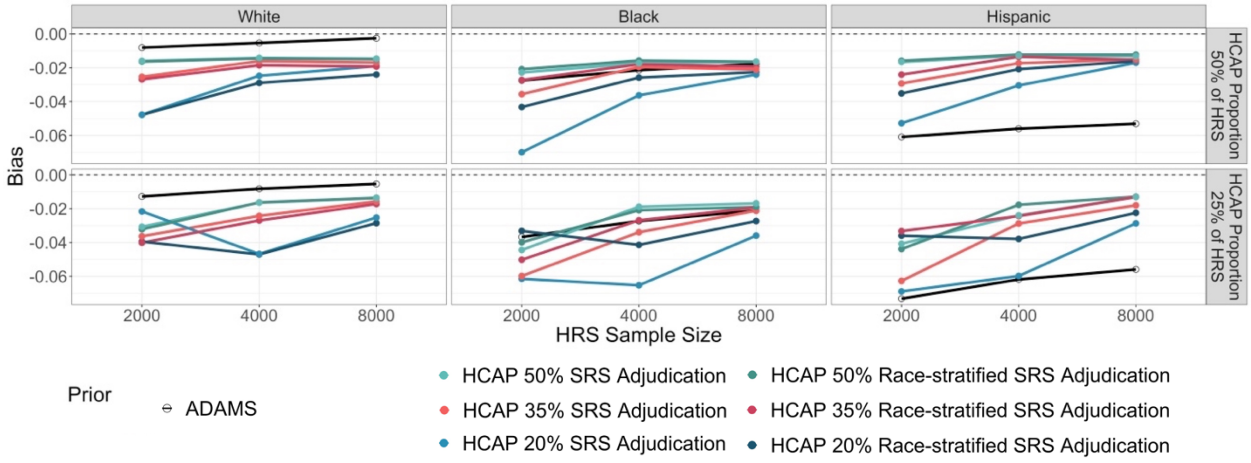


Figure 5.17: Bias in race-specific age and sex-standardized dementia prevalence by HRS sample size, HCAP sampling proportion, and sample used for the prior across 1000 simulation runs. Dashed horizontal line denotes no bias.

Figure 5.16 shows less than nominal coverage for race/ethnicity-specific 95% interval estimates of dementia prevalence. Coverage was at least 60% across simulation scenarios with the least

coverage for estimates of dementia prevalence for White participants. Bias and RMSE improve with sample size for all simulation scenarios. In scenarios with 50% HCAP calibration samples,

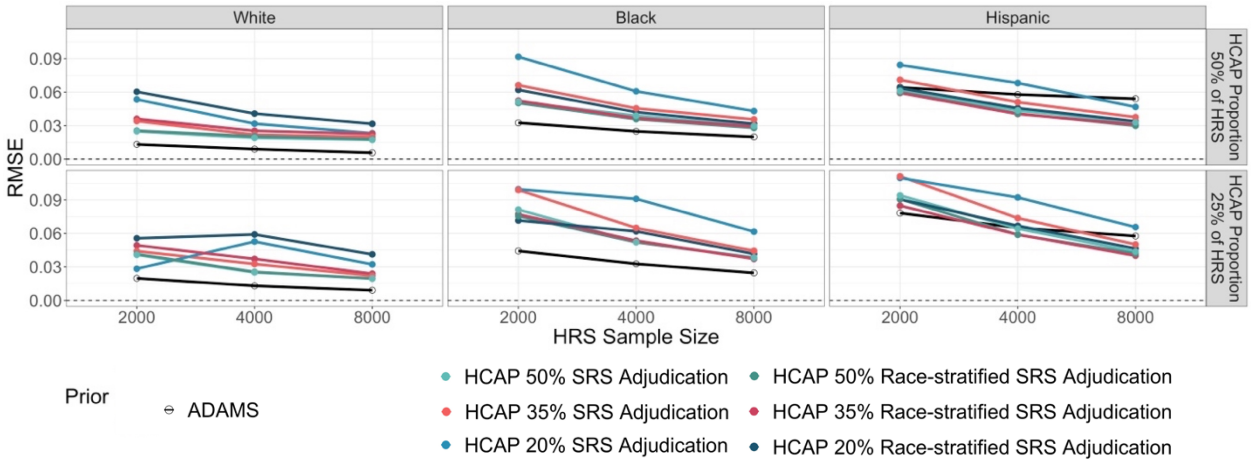


Figure 5.18: RMSE for race-specific age and sex-standardized dementia prevalence by HRS sample size, HCAP sampling proportion, and sample used for the prior across 1000 simulation runs. Dashed horizontal line denotes no bias and no variance.

bias (**Figure 5.17**) and RMSE (**Figure 5.18**) are uniformly better for Hispanic and Black participants but worse for White participants. Taken together with coverage results, these indicate that estimates for dementia prevalence from models using HCAP calibration samples as the prior are more variable across simulation runs than estimates that use ADAMS as the prior.

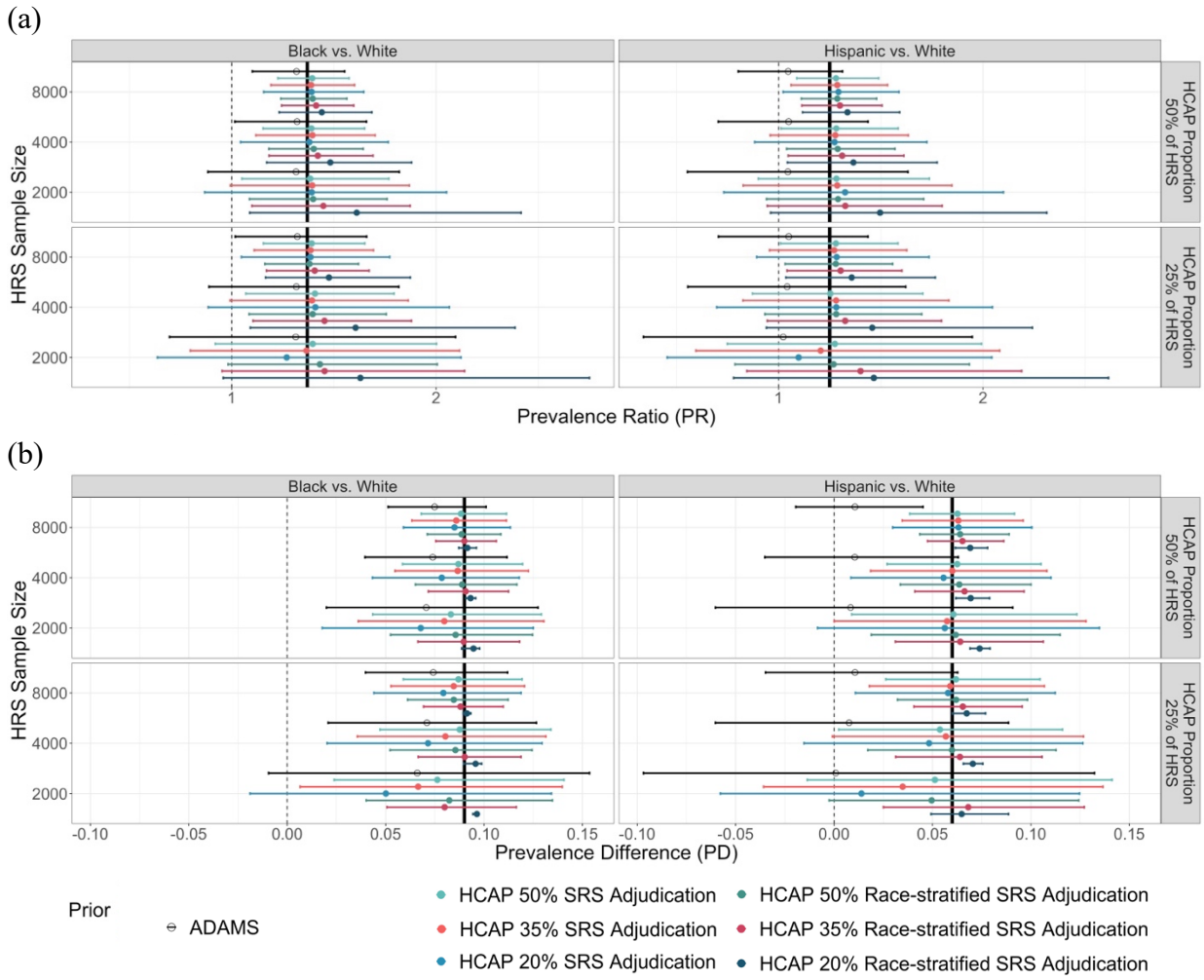


Figure 5.19: Mean and 95% interval estimate for (a) prevalence ratio (PR) and (b) prevalence difference (PD) in dementia for Black vs. White and Hispanic vs. White participants by HRS sample size, HCAP sampling proportion, and sample used for the prior across 1000 simulation runs. Black solid vertical lines denote true PR/PD of dementia in the superpopulation. Black dashed vertical line denotes no racial ethnic differences (PR=1; PD=0).

Nevertheless, racial/ethnic difference in prevalent dementia were better replicated and estimates more precise when using models with HCAP calibration priors as shown in **Figure 5.19**. Point estimates were close to the superpopulation truth across HRS sample sizes, HCAP sampling proportions, and priors used for the sample. Where the models using ADAMS priors underestimated increased prevalence in Hispanic participants to be 5% in the scenario with $n_{HRS} = 8000$ with 50% HCAP sampling proportion, the estimate using a 50% race-stratified

SRS HCAP calibration sample was much closer to the truth in the superpopulation (25% increased prevalence) with an estimate of 28% (**Figure 5.19a**).

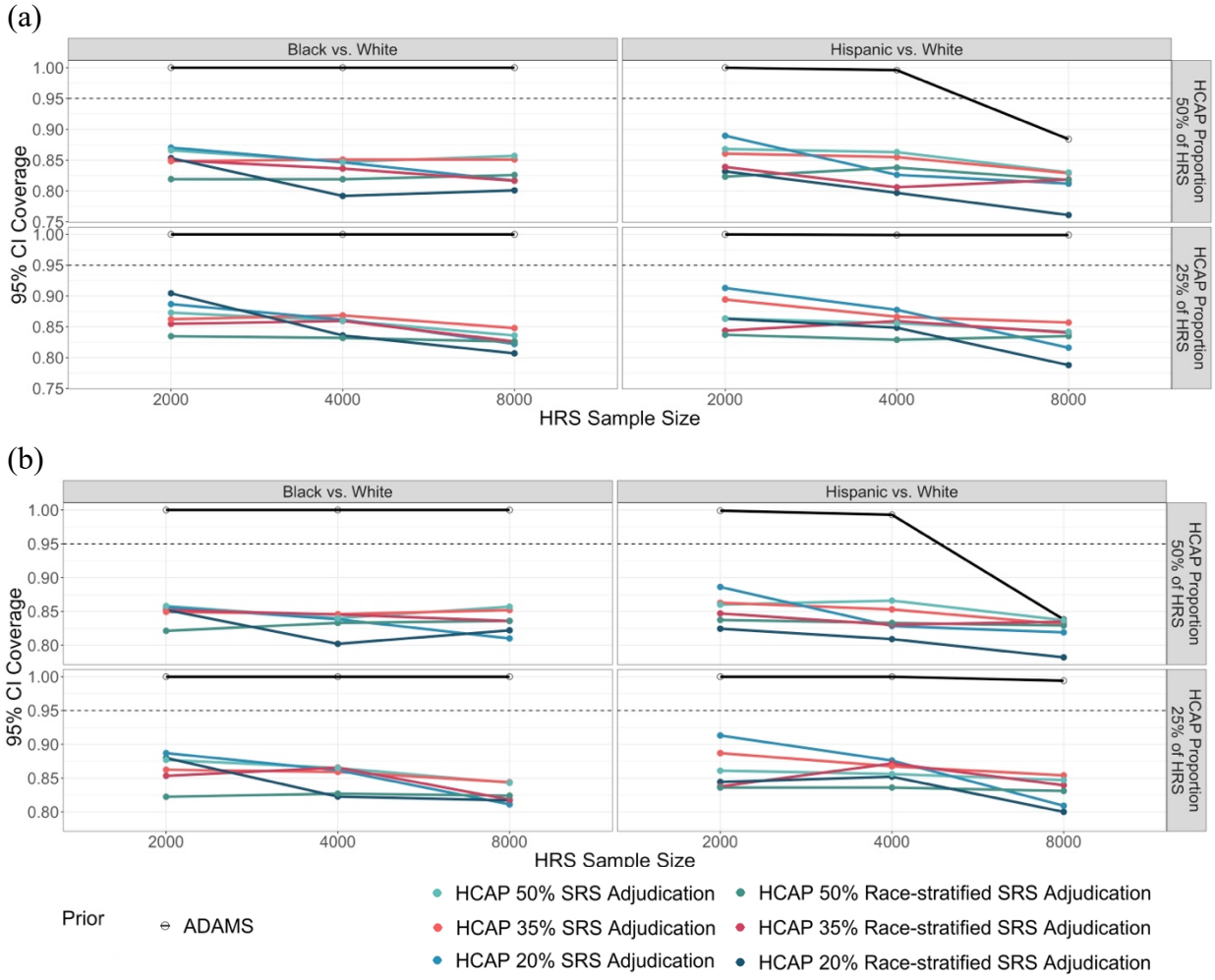


Figure 5.20: 95% interval coverage of true (a) prevalence ratio and (b) prevalence difference in dementia for Black vs. White and Hispanic vs. White participants by HRS sample size, HCAP sampling proportion, and sample used for the prior across 1000 simulation runs. Dashed horizontal line denotes nominal coverage of 95%.

There was at least 80% coverage across simulation scenarios for 95% interval estimates of PRs and PDs for Black vs. White and Hispanic vs. White participants (**Figure 5.20**). Bias improved with increased sample sizes and was nearly zero for scenarios with 50% HCAP calibration samples (**Figure 5.21**).

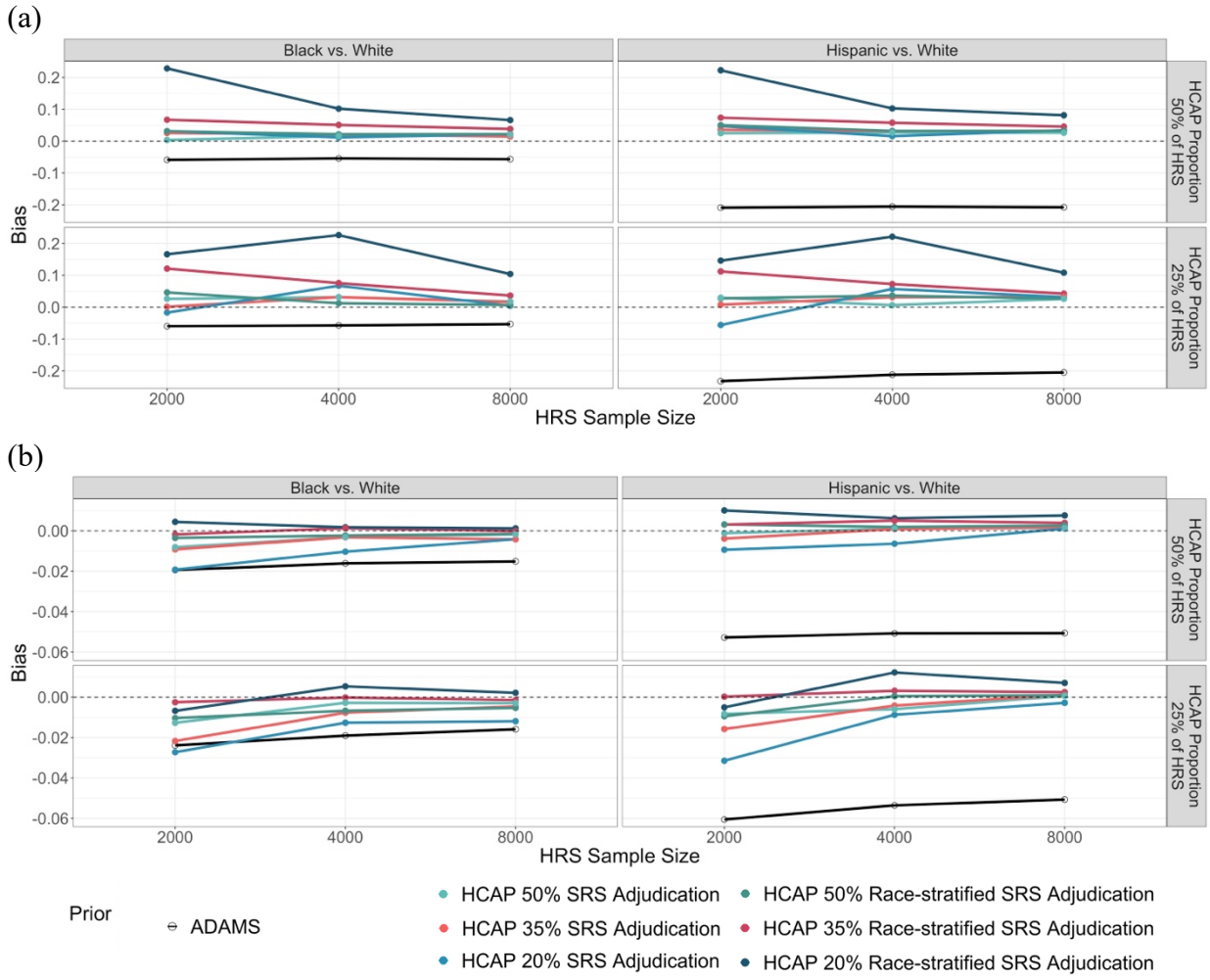


Figure 5.21: Bias in estimated (a) prevalence ratio and (b) prevalence difference in dementia for Black vs. White and Hispanic vs. White participants by HRS sample size, HCAP sampling proportion, and sample used for the prior across 1000 simulation runs. Dashed horizontal line denotes no bias.

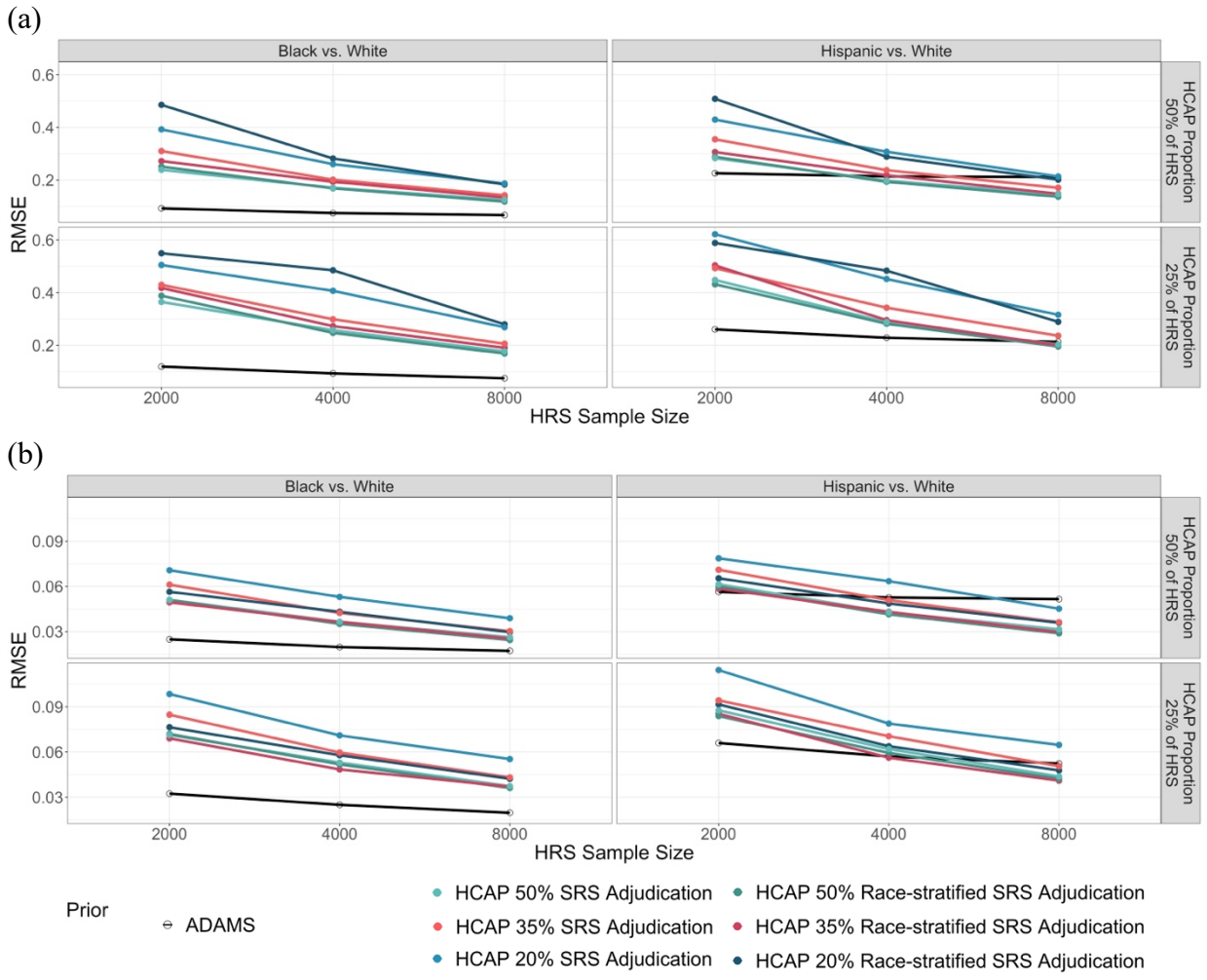


Figure 5.22: RMSE for estimated (a) prevalence ratio and (b) prevalence difference in dementia for Black vs. White and Hispanic vs. White participants by HRS sample size, HCAP sampling proportion, and sample used for the prior across 1000 simulation runs. Dashed horizontal line denotes no bias.

RMSE decreased with increasing sample sizes across simulation scenarios but was still greater in most scenarios compared to simulations that used ADAMS to specify priors, demonstrating that estimates from these models were more variable compared to models that used ADAMS priors (**Figure** 5.22).

5.6: Combining results across analyses that use different priors

Results from the simulation study discussed above showed improved results in algorithmic dementia classification and subsequent population-level inferences for Black and White participants in HCAP when priors were specified based on adjudicated subsamples of the HCAP study. With the ADAMS study being conducted over 20 years ago, it seems reasonable to update priors using newly adjudicated samples of participants. Information from the ADAMS study is still useful, however, and overall results for algorithmic dementia classification based on ADAMS priors were acceptable.

In this section I present combined results from analyses using the ADAMS prior and priors specified based on each of the HCAP calibration samples used in the simulation study above. There are several ways to combine results across different analyses; I chose to combine results using Rubin's Rules, which are typically used for combining multiple imputations (see **Appendix Section B.1:**. Thus, variance estimates for combined results account for variance within the analysis and variance between analyses.

Plots for combined results follow. Average point estimates for cognitive impairment class proportions aligned with the truth in the superpopulation for all classes, but variance was larger (**Figure 5.23**). As a result, all interval estimates for combined analyses had above nominal coverage (**Figure 5.24**). There was virtually no bias in Unimpaired and MCI groups across sample sizes (**Figure 5.25**) and RMSE decreased with increasing sample size (**Figure 5.26**). Race/ethnicity-specific results for bias and RMSE more closely resembled results using ADAMS priors with proportions in cognitive impairment groups overestimated for Black and Hispanic participants (**Figure 5.27**-**Figure 5.28**). Population-level inference for dementia prevalence from combined estimates were not as accurate for Hispanic participants because they were being

pulled towards results from ADAMS priors (**Figure 5.29**). There was much more variance in estimates, however, so coverage resembles simulations using ADAMS priors with above-nominal coverage except for Hispanic participants in scenarios with the largest sample size (**Figure 5.30**).

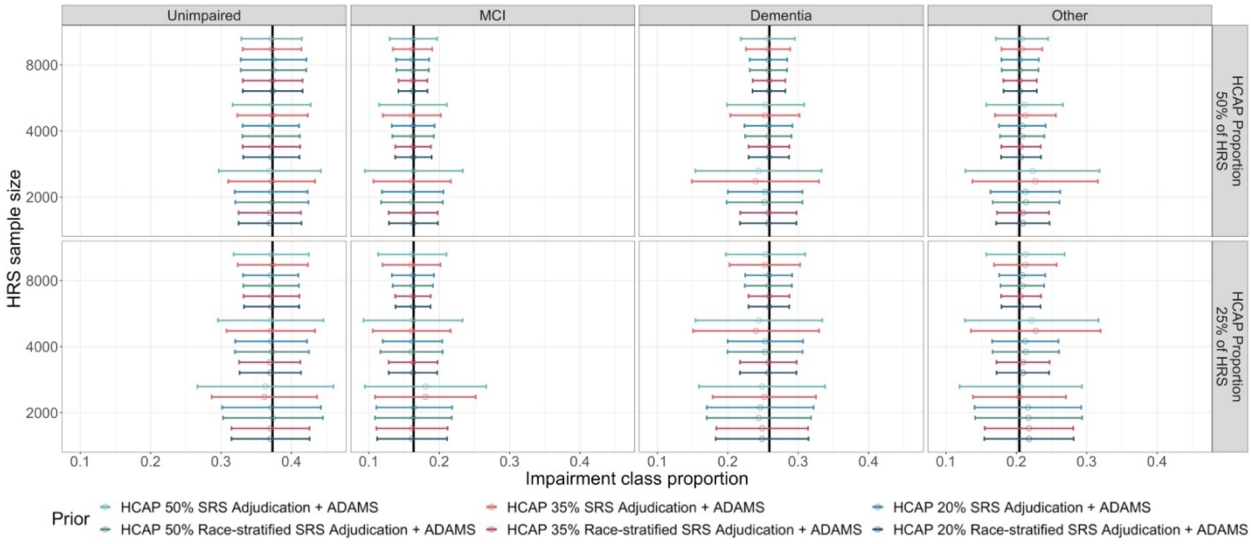


Figure 5.23: Mean and 95% interval of combined estimates (results using ADAMS priors + results using specified prior) for cognitive impairment class proportions by HRS sample size, HCAP sampling proportions, and priors used for the sample. Black vertical lines denote true impairment class proportions in the superpopulation.

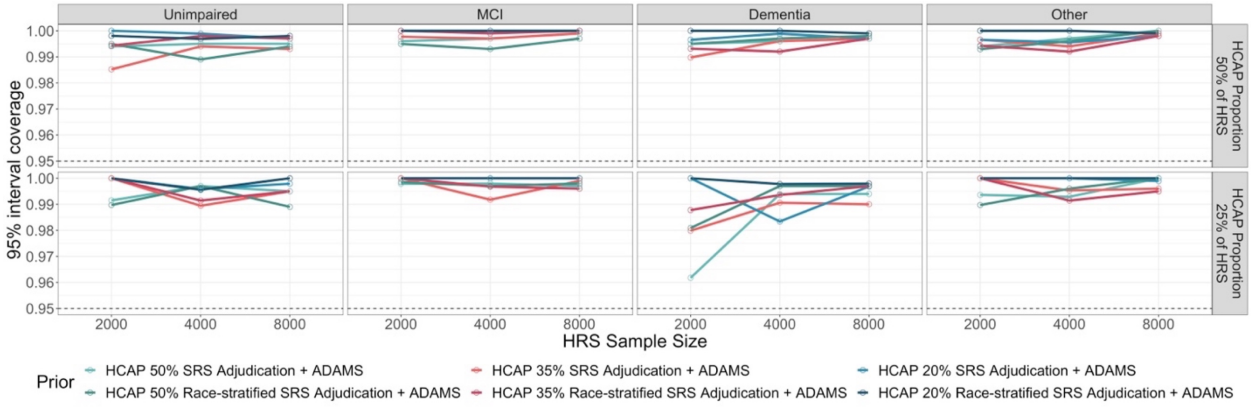


Figure 5.24: 95% interval coverage of true impairment class proportion for combined estimates (results using ADAMS priors + results using specified prior) by cognitive impairment class, HRS sample size, HCAP sampling proportion, and sample used for the prior. Dashed horizontal line denotes nominal coverage of 95%.

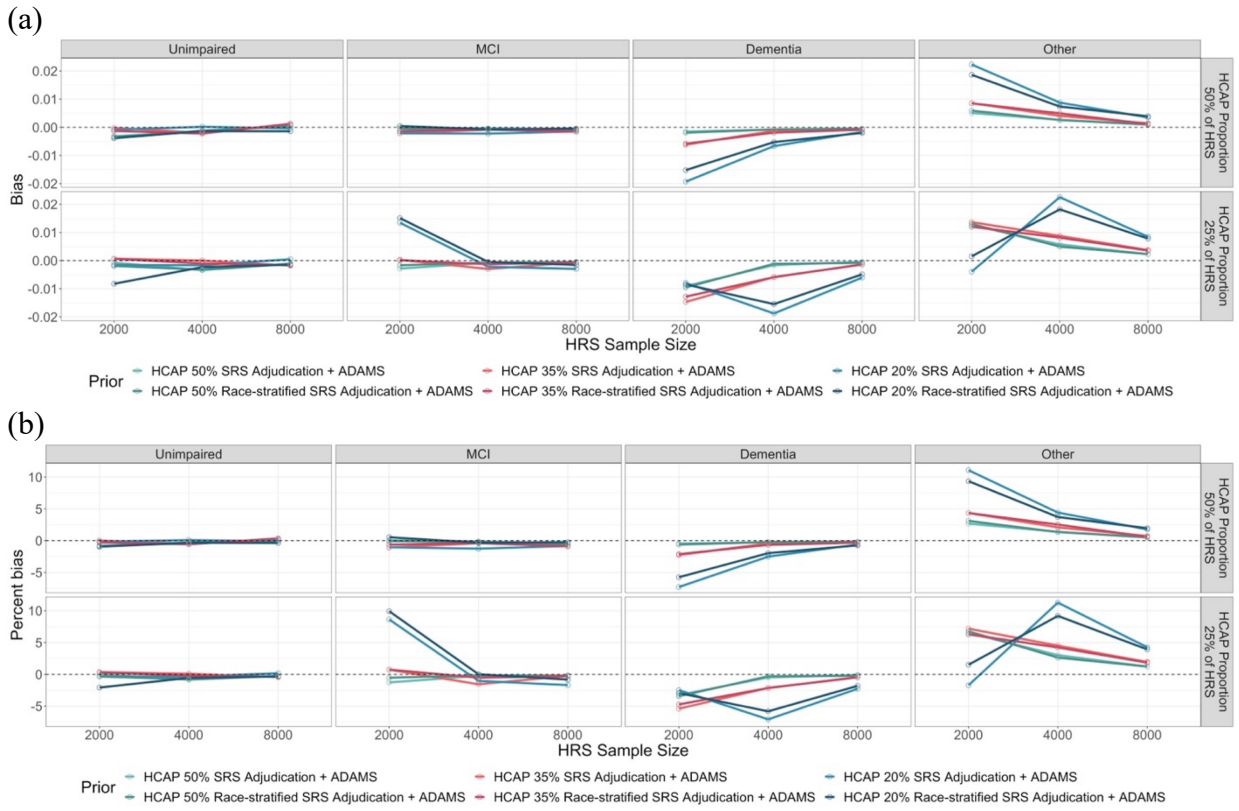


Figure 5.25: Bias (a) and percent bias (b) in combined estimates of cognitive impairment class proportions (results using ADAMS priors + results using specified prior) by HRS sample size, HCAP sampling proportion, and sample used for the prior. Dashed horizontal line denotes no bias.

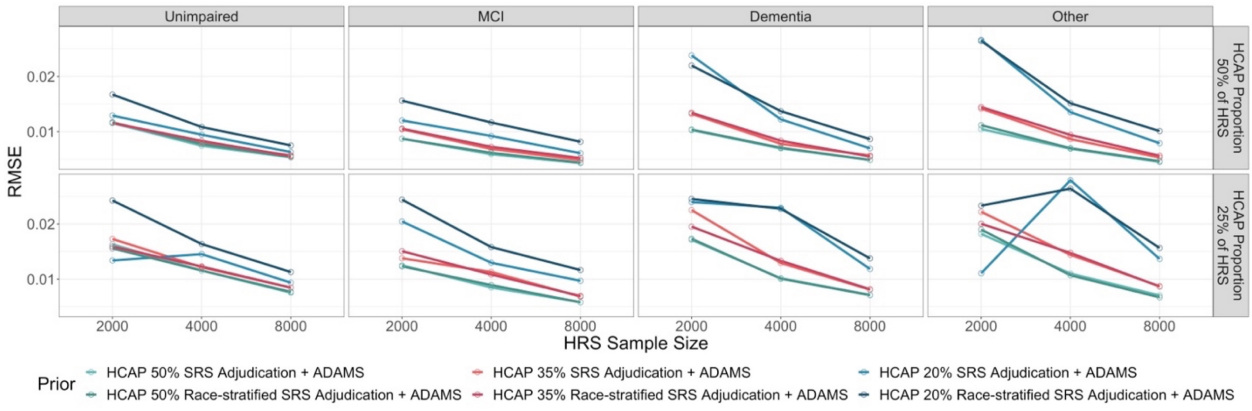
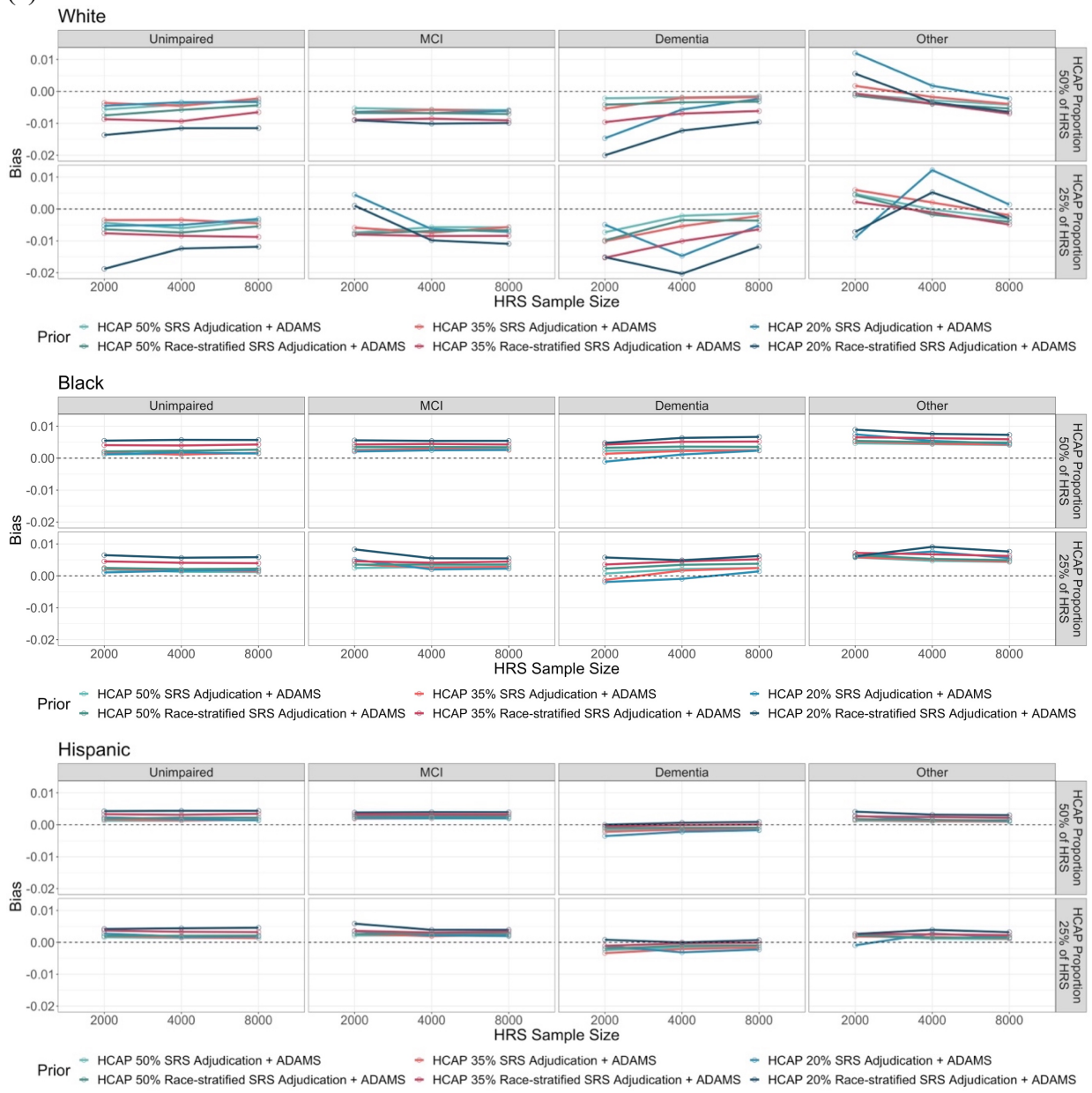


Figure 5.26: Root mean square error (RMSE) for combined estimated cognitive impairment class proportions (results using ADAMS priors + results using specified prior) by HRS sample size, HCAP sampling proportion, and sample used for the prior.

(a)


White
Unimpaired
MCI
Dementia
Other
HCAP Proportion 50% of HRS
HCAP Proportion 25% of HRS
Bias
0.01
0.00
-0.01
-0.02
2000
4000
8000
HRS Sample Size
Prior
HCAP 50% SRS Adjudication + ADAMS
HCAP 35% SRS Adjudication + ADAMS
HCAP 20% SRS Adjudication + ADAMS
HCAP 50% Race-stratified SRS Adjudication + ADAMS
HCAP 35% Race-stratified SRS Adjudication + ADAMS
HCAP 20% Race-stratified SRS Adjudication + ADAMS
Black
Unimpaired
MCI
Dementia
Other
HCAP Proportion 50% of HRS
HCAP Proportion 25% of HRS
Bias
0.01
0.00
-0.01
-0.02
2000
4000
8000
HRS Sample Size
Prior
HCAP 50% SRS Adjudication + ADAMS
HCAP 35% SRS Adjudication + ADAMS
HCAP 20% SRS Adjudication + ADAMS
HCAP 50% Race-stratified SRS Adjudication + ADAMS
HCAP 35% Race-stratified SRS Adjudication + ADAMS
HCAP 20% Race-stratified SRS Adjudication + ADAMS
Hispanic
Unimpaired
MCI
Dementia
Other
HCAP Proportion 50% of HRS
HCAP Proportion 25% of HRS
Bias
0.01
0.00
-0.01
-0.02
2000
4000
8000
HRS Sample Size
Prior
HCAP 50% SRS Adjudication + ADAMS
HCAP 35% SRS Adjudication + ADAMS
HCAP 20% SRS Adjudication + ADAMS
HCAP 50% Race-stratified SRS Adjudication + ADAMS
HCAP 35% Race-stratified SRS Adjudication + ADAMS
HCAP 20% Race-stratified SRS Adjudication + ADAMS

(b)

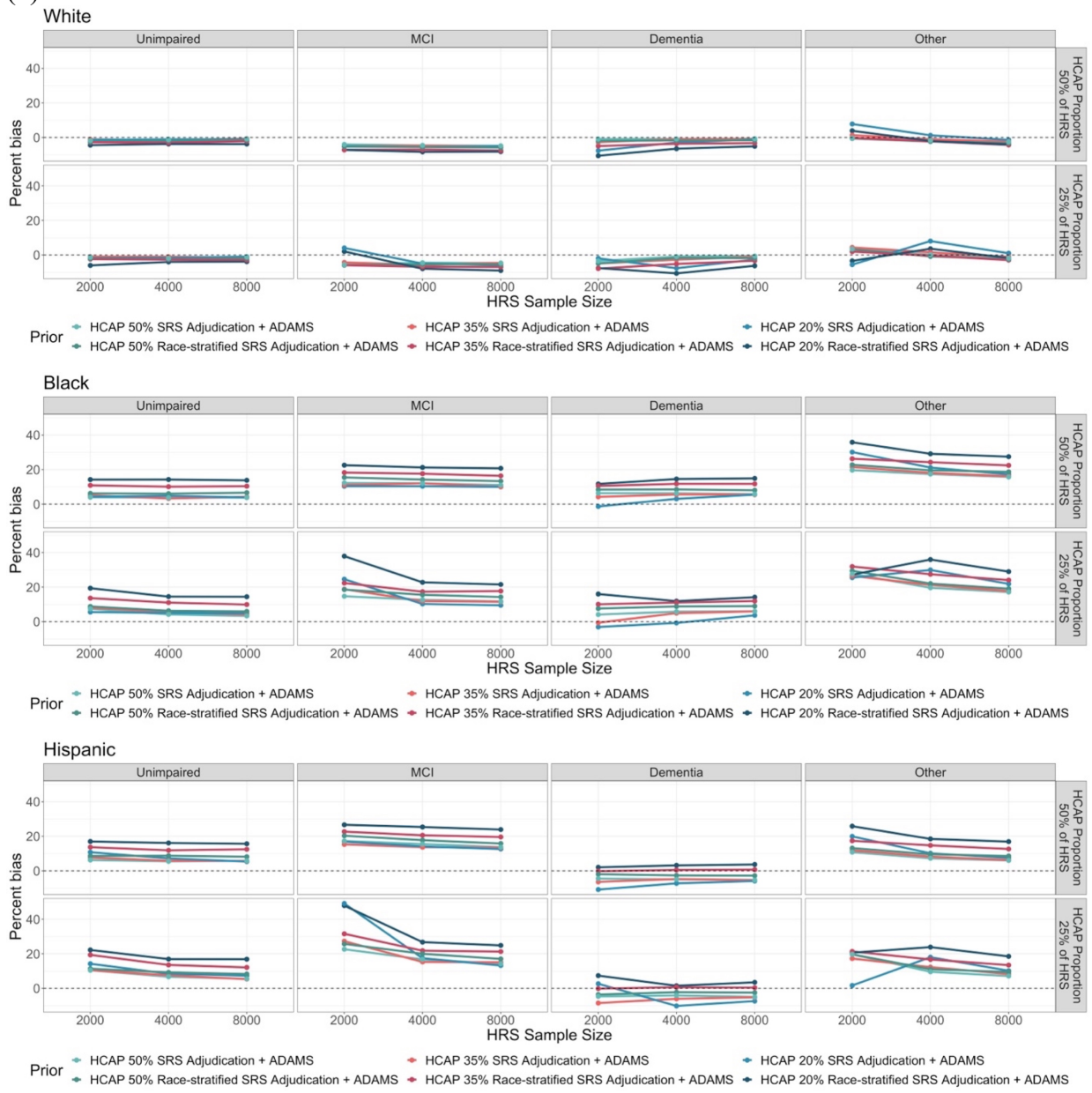


Figure 5.27: Race/ethnicity-specific bias (a) and percent bias (b) in combined estimated cognitive impairment class proportions (results using ADAMS priors + results using specified prior) by HRS sample size, HCAP sampling proportion, and sample used for the prior. Dashed horizontal line denotes no bias.

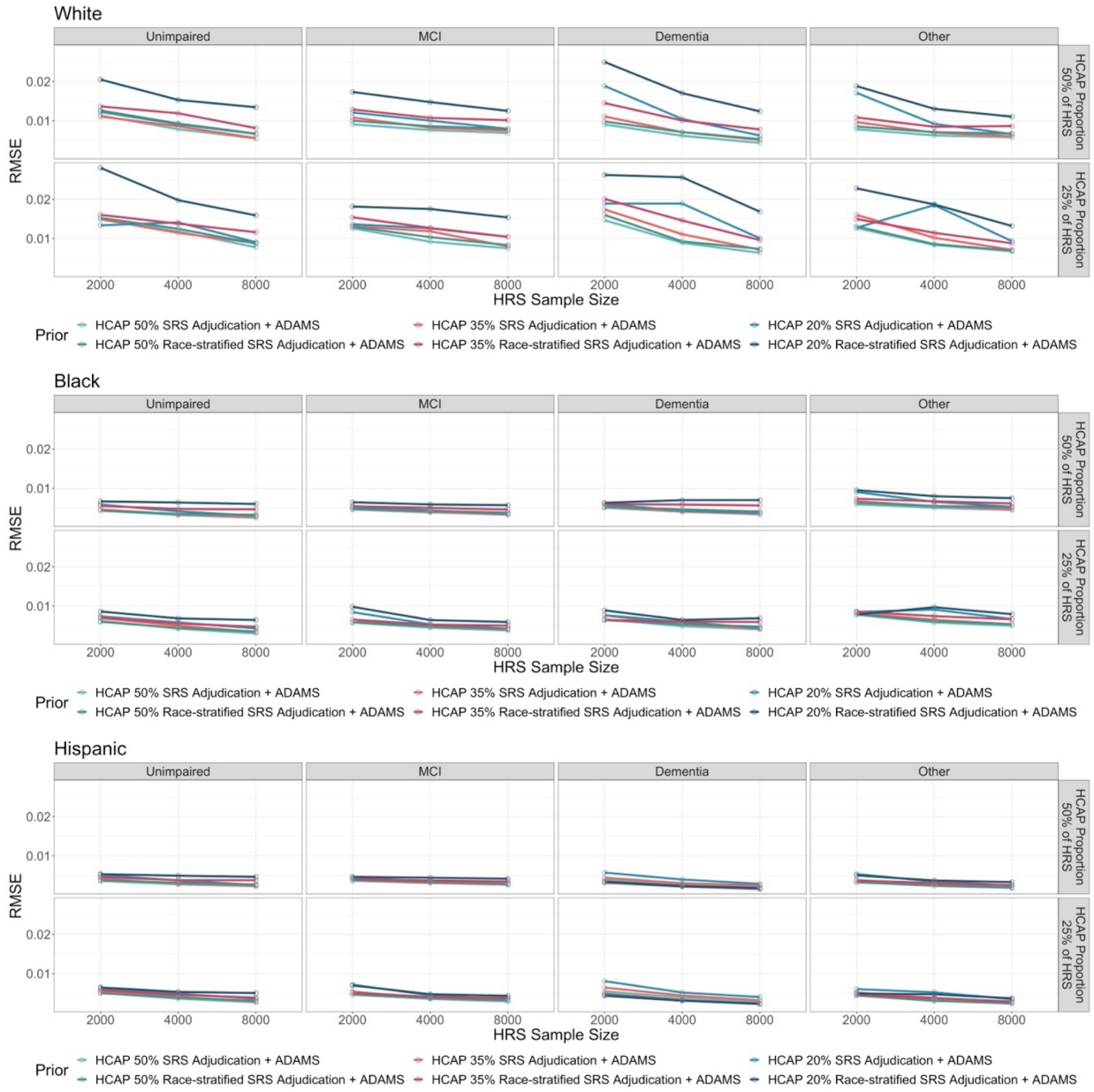


Figure 5.28: Race/ethnicity-specific RMSE for combined estimated cognitive impairment class proportions (results using ADAMS priors + results using specified prior) by HRS sample size, HCAP sampling proportion, and sample used for the prior.

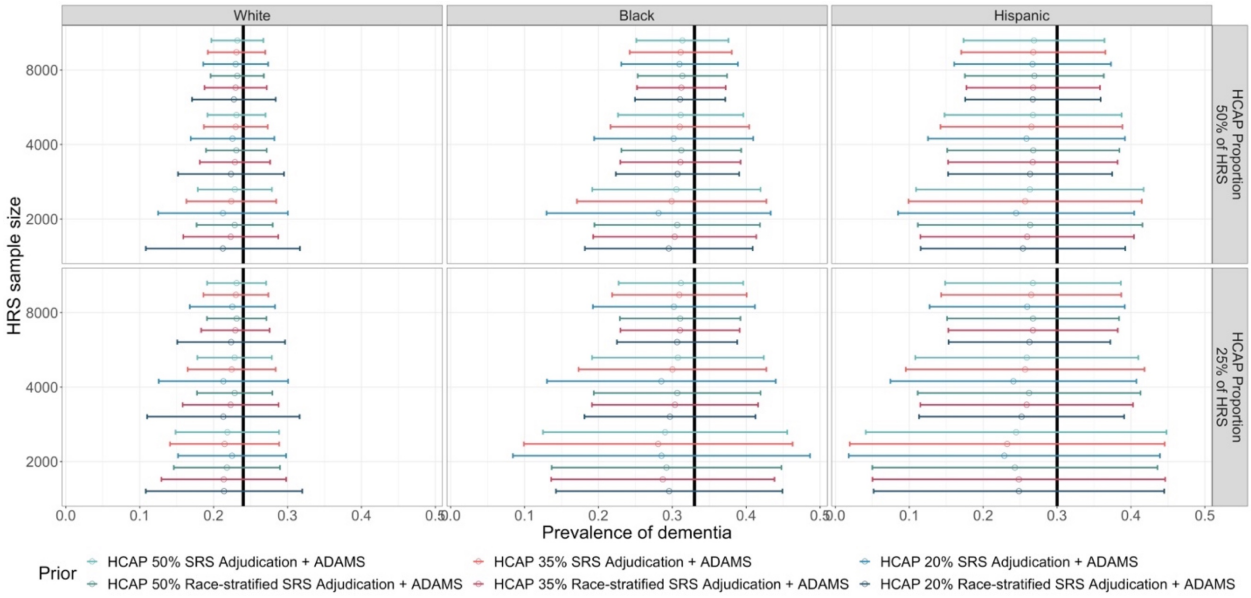


Figure 5.29: Mean and 95% interval estimate for combined estimated race-specific age and sex-standardized dementia prevalence (results using ADAMS priors + results using specified prior) by HRS sample size, HCAP sampling proportion, and sample used for the prior. Vertical lines denote true race-specific dementia prevalence in the superpopulation.

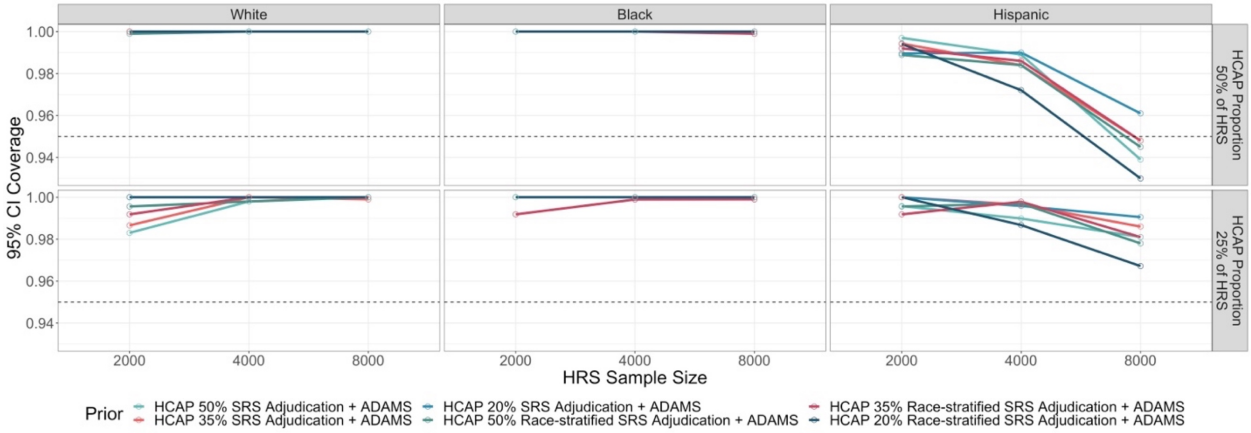


Figure 5.30: 95% interval coverage of combined estimated prevalence of dementia (results using ADAMS priors + results using specified prior) by race/ethnicity, HRS sample size, HCAP sampling proportion, and sample used for the prior across 1000 simulation runs. Dashed horizontal line denotes nominal coverage of 95%.

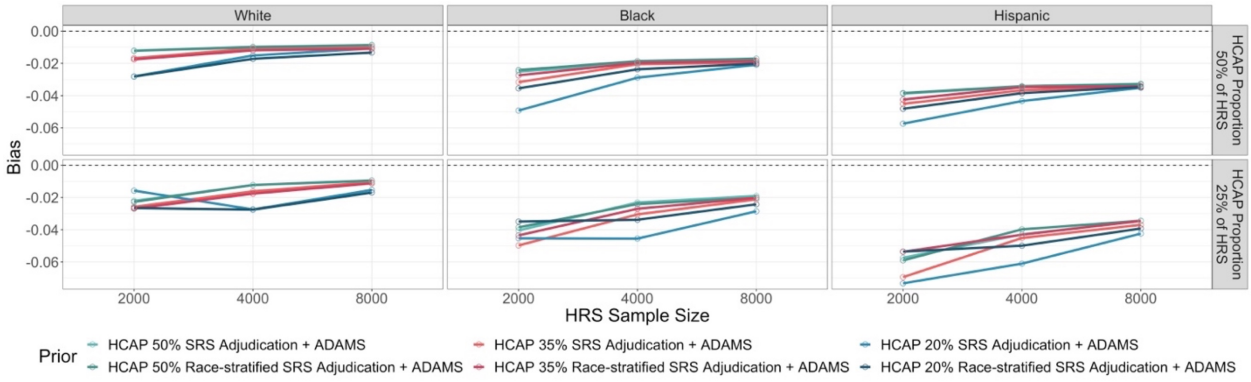


Figure 5.31: Bias in combined estimated race-specific age and sex-standardized dementia prevalence (results using ADAMS priors + results using specified prior) by HRS sample size, HCAP sampling proportion, and sample used for the prior. Dashed horizontal line denotes no bias.

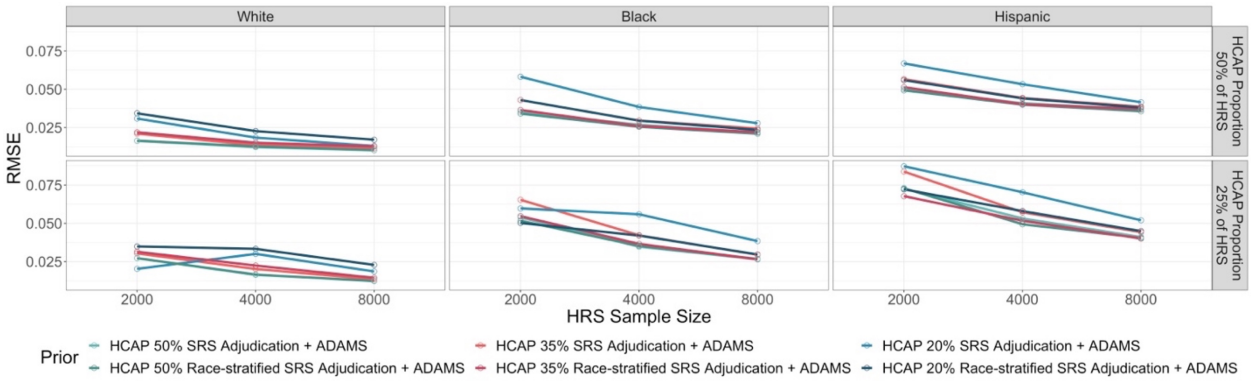


Figure 5.32: RMSE for combined estimated race-specific age and sex-standardized dementia prevalence (results using ADAMS priors + results using specified prior) by HRS sample size, HCAP sampling proportion, and sample used for the prior.

Bias and RMSE for combined estimates of standardized dementia prevalence improved with sample size (**Figure 5.31**-**Figure 5.32**).

Combined point estimates for racial ethnic differences in dementia were closer to the truth for Black vs. White but not for Hispanic vs. White estimates compared to the simulation results that specified priors based off HCAP calibration samples (**Figure 5.33**). Estimates were very imprecise which meant that all 95% interval estimates had above nominal coverage (**Figure**

5.34). Bias and RMSE of estimates improved with increased sample size (**Figure 5.35**-**Figure 5.36**).

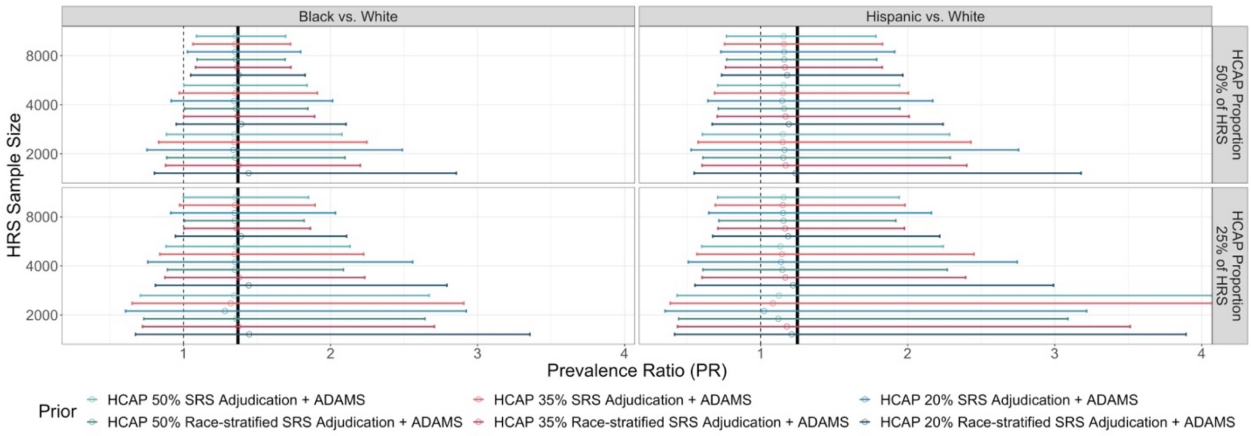


Figure 5.33: Mean and 95% interval for combined estimated prevalence ratio of dementia for Black vs. White and Hispanic vs. White participants (results using ADAMS priors + results using specified prior) by HRS sample size, HCAP sampling proportion, and sample used for the prior. Black solid vertical lines denote true PR of dementia in the superpopulation. Black dashed vertical line denotes no racial ethnic differences (PR = 1).

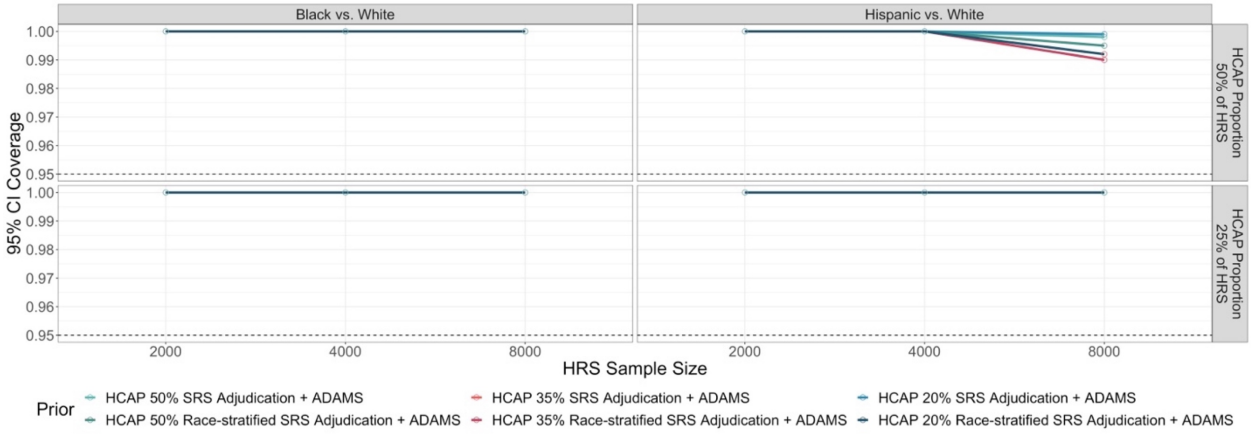


Figure 5.34: 95% interval coverage of combined estimated prevalence ratio of dementia for Black vs. White and Hispanic vs. White participants (results using ADAMS priors + results using specified prior) by HRS sample size, HCAP sampling proportion, and sample used for the prior. Dashed horizontal line denotes nominal coverage of 95%.

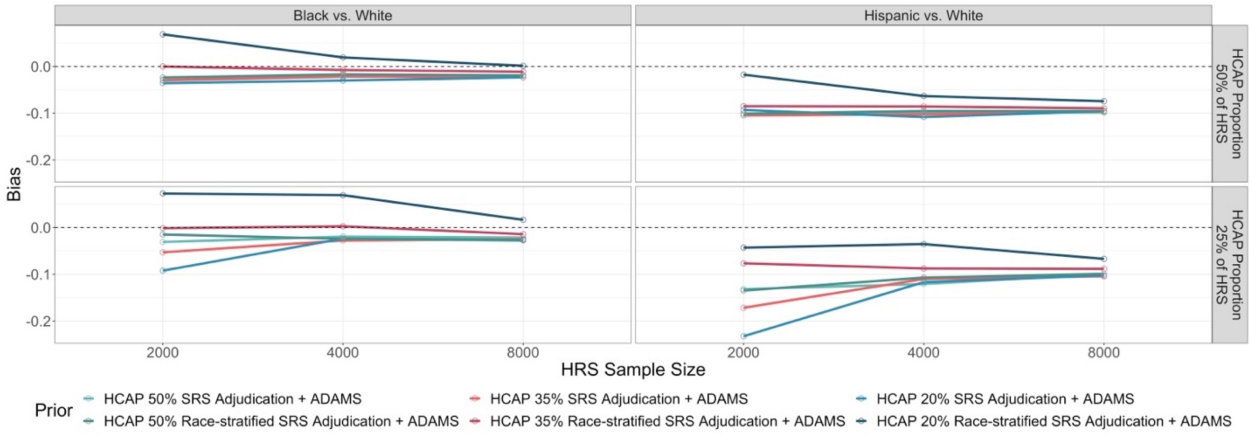


Figure 5.35: Bias in combined estimated prevalence ratio of dementia for Black vs. White and Hispanic vs. White participants (results using ADAMS priors + results using specified prior) by HRS sample size, HCAP sampling proportion, and sample used for the prior. Dashed horizontal line denotes no bias.

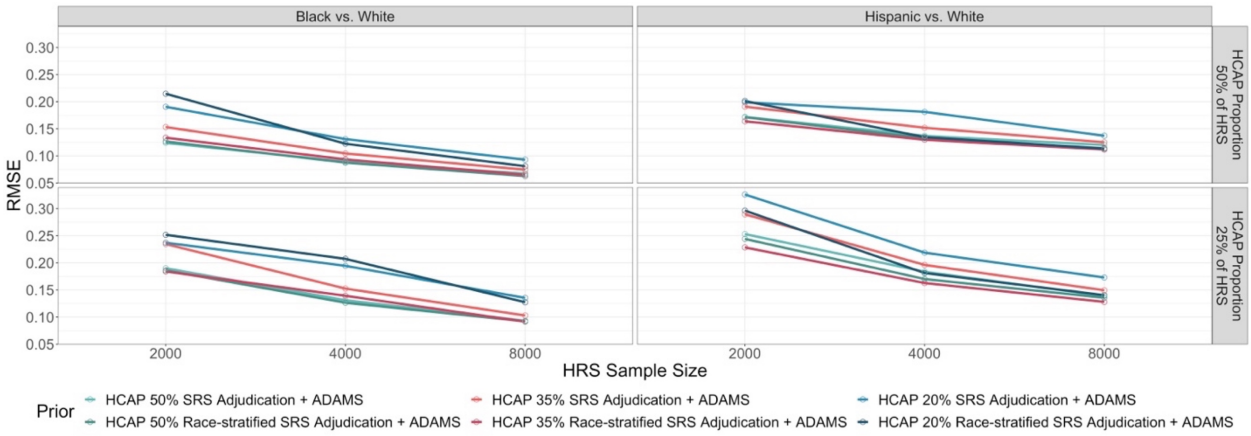


Figure 5.36: RMSE for combined estimated prevalence ratio of dementia for Black vs. White and Hispanic vs. White participants (results using ADAMS priors + results using specified prior) by HRS sample size, HCAP sampling proportion, and sample used for the prior.

5.7: Discussion

Results from the simulation study presented in this chapter demonstrate that specifying priors in the Bayesian latent class mixture model based off of adjudicated subsets of the HCAP sample improve estimates for Black and Hispanic participants and yielded correct inferences for racial/ethnic differences in the population. This was especially noticeable for Hispanic

participant for whom inferences were not correct when priors were specified using the ADAMS sample.

I assessed adjudicating subsets at the rate of 20%, 35%, and 50% and using SRS vs. race-stratified SRS sampling. Race-stratified sampling improved results the most in simulation scenarios with smaller sample sizes. Results for 35% vs. 50% calibration sample rates were similar. In the largest sample sizes, even 20% sampling was sufficient for recovering correct inferences. Taken together, these results imply that a calibration subset with sufficient representation among racial/ethnic groups and across the distribution of cognitive performance would greatly improve algorithmic dementia classification using the Bayesian latent class mixture model and subsequent analyses. For larger studies, smaller proportions of the study could be adjudicated for these calibration subsets which could amount to large cost savings for study investigators.

The quality of individual synthetic datasets produced by the Bayesian latent class mixture model becomes increasingly important as the intended use for the datasets is more general. This simulation study demonstrated that the Bayesian latent class mixture model with priors specified using adjudicated subsets of HCAP produced accurate dementia classification and was able to reproduce racial/ethnic differences in the superpopulation. Assessments of categorical variable distributions and medians and skew for continuous variables showed that synthetic data are more like observed data when priors are specified in this way compared to priors specified using ADAMS data. This provides increased confidence that additional analyses using these datasets would be valid.

Recognizing that whether the prior is "correct" cannot be verified in practice and that prior distributions placed on cognitive impairment class membership are particularly important

in this model because there is no information in the data with which to update the prior, I presented combined results from analyses with priors specified based on ADAMS and priors specified based on the HCAP calibration samples assessed in this simulation. Using Rubin's Rules to combine analyses is equivalent to an unweighted random-effects meta-analysis (Borenstein et al. 2010) and with only two studies to combine, variance estimates were very large. Thus, combined estimates had better coverage properties but with a large amount of lost precision. Additionally, results for Hispanic participants were pulled toward results from analyses that used ADAMS priors, so combined results were worse for that group.

Results from this simulation show promising ways forward for algorithmic dementia classification using the Bayesian latent class model, especially for racial/ethnic disparities research. The improved estimation of cognitive impairment for Hispanic participants is especially exciting as estimates for this group have to date been a persistent challenge.

Chapter 6 Real Data Application: Algorithmic Dementia Classification in the Harmonized Cognitive Assessment Protocol

The simulation studies in Chapters 4-5 provided insight to the statistical properties of the Bayesian latent class mixture model for algorithmic dementia classification and whether we can expect to recover correct inference when generalizing results to population-representative studies. Because results from the simulation study were promising, I used the Bayesian latent class mixture model to algorithmically classify cognitive impairment in the HCAP 2016 study. The HCAP study does not include clinical dementia adjudication for its participants, however, HCAP investigators have published results for algorithmic dementia classification in HCAP using an algorithm developed in-house that used a series of decision rules and normative score cut-offs to classify cognitive impairment for participants (Manly et al. 2022).

Implementation details and results for the Bayesian latent class mixture model applied to the HCAP 2016 study are presented in this chapter. Results will be compared to published results for algorithmic dementia classification in HCAP 2016.

6.1: Data preparation

The HCAP 2016 data was prepared and cleaned in the same way described in Chapter 4 (see **Section 4.2.3:**). Thus, this analysis only included HCAP 2016 participants aged 70+. I hotdeck imputed the HCAP data using the steps described in **Section 4.2.3.2:**; however, for this analysis, I imputed the HCAP data 10 ten times to account for uncertainty in imputed values. I chose 10 imputations to exceed that maximum amount of missing data in any HCAP 2016 variable (4.6% for Trails A task; **Table 4.1**)

6.2: Specifying prior distributions

I specified priors in this analysis using information from the ADAMS study because there are no gold-standard clinically adjudicated subsets of HCAP. Priors on all model parameters were specified in the same way as described in Chapter 4 (see **Section 4.4:**).

6.3: Results

I used the same Bayesian modeling diagnostics described in previous chapters to assess model fit— prior predictive checks, model convergence diagnostics, and posterior predictive checks. It would be infeasible to perform these checks for each of the 10 imputed HCAP 2016 datasets, thus, checks were only performed for one of the imputed datasets. The same tuning parameters were then specified for all imputed HCAP datasets based on model diagnostics.

A key difference between performing these checks in the real HCAP 2016 studied compared to simulation studies is that I did not have clinically adjudicated cognitive impairment statuses to stratify prior predictive and posterior predictive checks by to compare subsample characteristics. Thus, prior and posterior predictive checks were visualized for the full HCAP sample to check whether salient features of the overall sample were captured by the synthetic datasets.

6.3.1: Prior predictive checks

Prior predictive checks for overall contingency cell counts and normalized MMSE are presented in **Figure 6.1**. Prior predictive distributions of contingency cell counts are all centered at observed counts in the HCAP 2016 sample (**Figure 6.1a**) and the prior predictive distribution for normalized MMSE is wider than the observed distribution, as desired. Prior predictive checks for

remaining continuous variables were similarly satisfactory and code for producing the checks is provided on the associated GitHub repository.

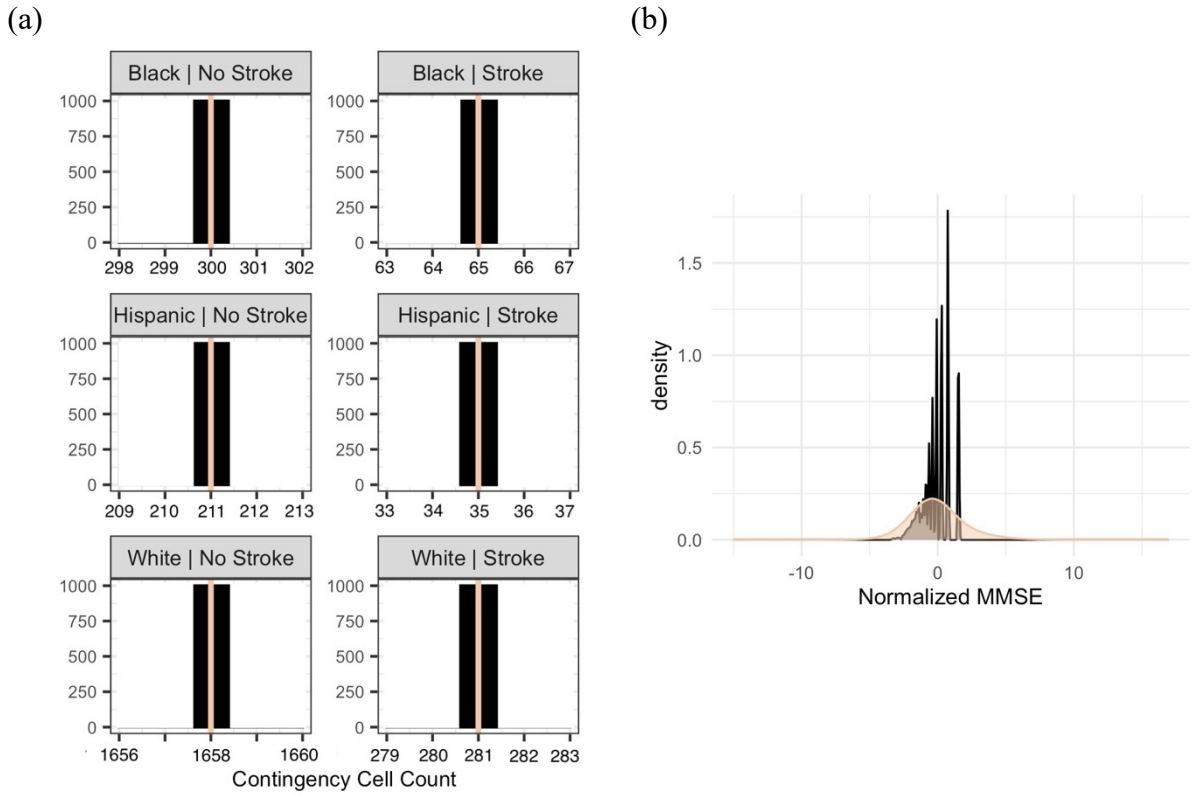


Figure 6.1: Prior predictive distributions of (a) contingency cell counts and (b) normalized MMSE for the HCAP sample overall. Distributions are based on 1000 synthetic HCAP datasets. Colored vertical lines in panel (a) denote observed counts in HCAP and gray density in panel (b) represents observed distribution of normalized MMSE in HCAP.

6.3.2: Assessing model convergence

I produced MCMC chains for each parameter in this analysis, but I primarily monitored cognitive impairment class proportion chains and impairment group-specific variances for continuous variables (**Figure 6.2**). All parameter chains demonstrated convergence, and code for producing MCMC chains for the other model parameters can be found in the associated GitHub repository.

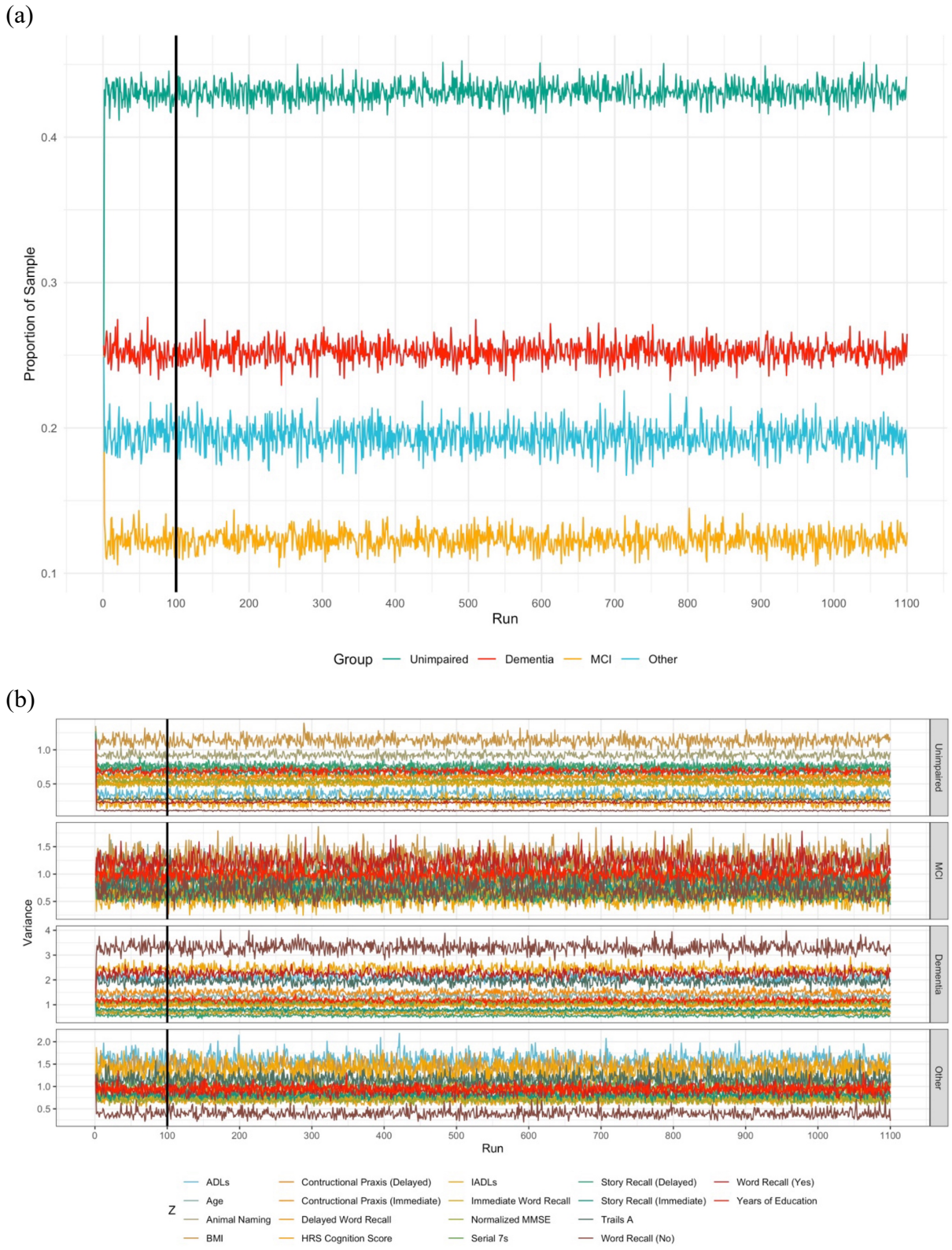


Figure 6.2: MCMC chains of proportions of (a) impairment class membership and (b) impairment group-specific variances of continuous variables based on 1000 synthetic HCAP datasets. Black vertical lines mark the end of the burn-in period (100 runs).

6.3.3: Posterior predictive checks

I assessed posterior distributions of contingency cell counts and median and skew for continuous variables overall. Posterior distributions of contingency cell counts were mostly satisfactory with some lack of fit for White participants without stroke history. The model underestimated the count in that category by about 80 participants (**Figure 6.3a**). The posterior distribution of median normalized MMSE was not centered at the observed median, however, the model tended to underestimate normalized MMSE by only 3.5 points (on a scale of 0-100). Overall skewness for normalized MMSE was reproduced by the model. Posterior distributions of median and skew for the remaining continuous variables are presented in **Appendix Figure D.5-Appendix Figure D.6**.

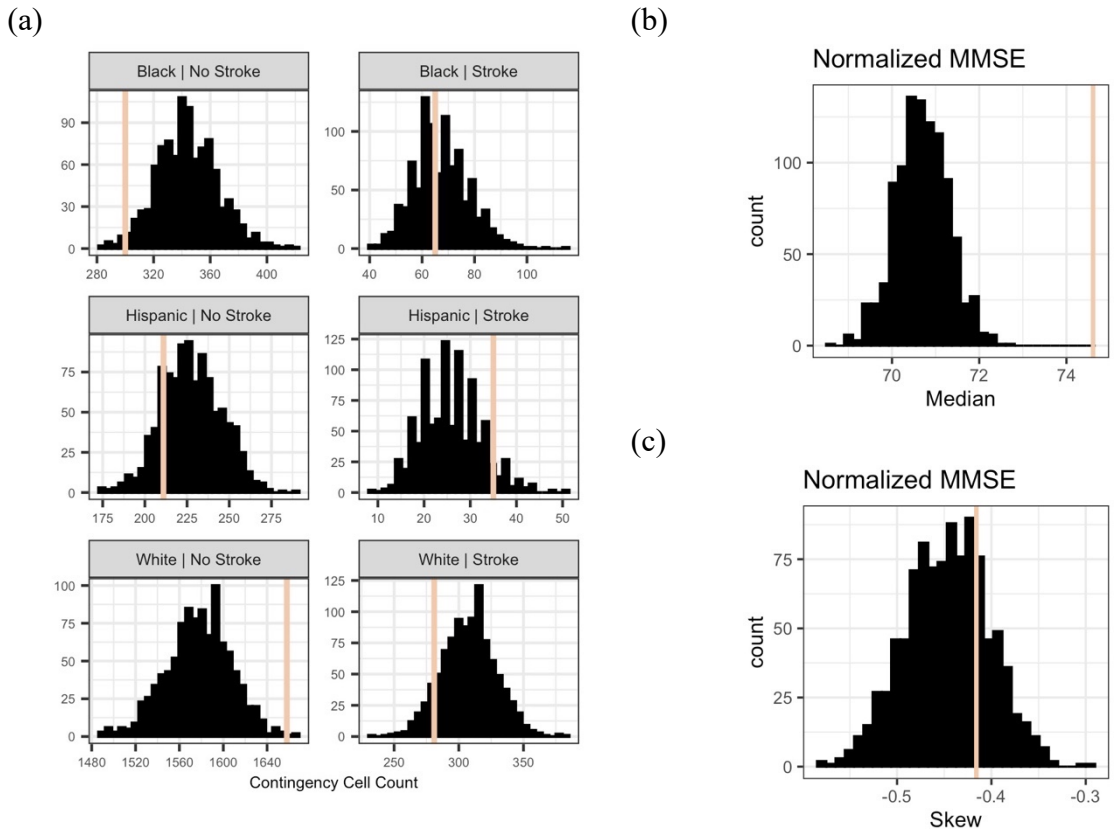


Figure 6.3: Posterior predictive distributions of (a) contingency cell counts (b) median normalized MMSE, and (c) normalized MMSE skew. Distributions are based on 1000 synthetic HCAP datasets. Colored vertical lines in each panel represent observed values in the HCAP data.

6.3.4: Algorithmic dementia classification

Algorithmic dementia classification results for each imputed HCAP 2016 dataset were combined using Rubin's rules and are presented in **Figure 6.4** along with estimates reported by HCAP investigators. Estimates from the Bayesian latent class mixture model were precise for all cognitive impairment groups. The model estimated that in HCAP 70+, Unimpaired participants comprised the largest group (43.1% [41.8%, 44.4%]) and participants with MCI the smallest (12.3% [11.0%, 13.6%]). The model estimated that participants with Dementia comprised 25.1% of the sample (95% CI: [23.7%, 26.5%]).

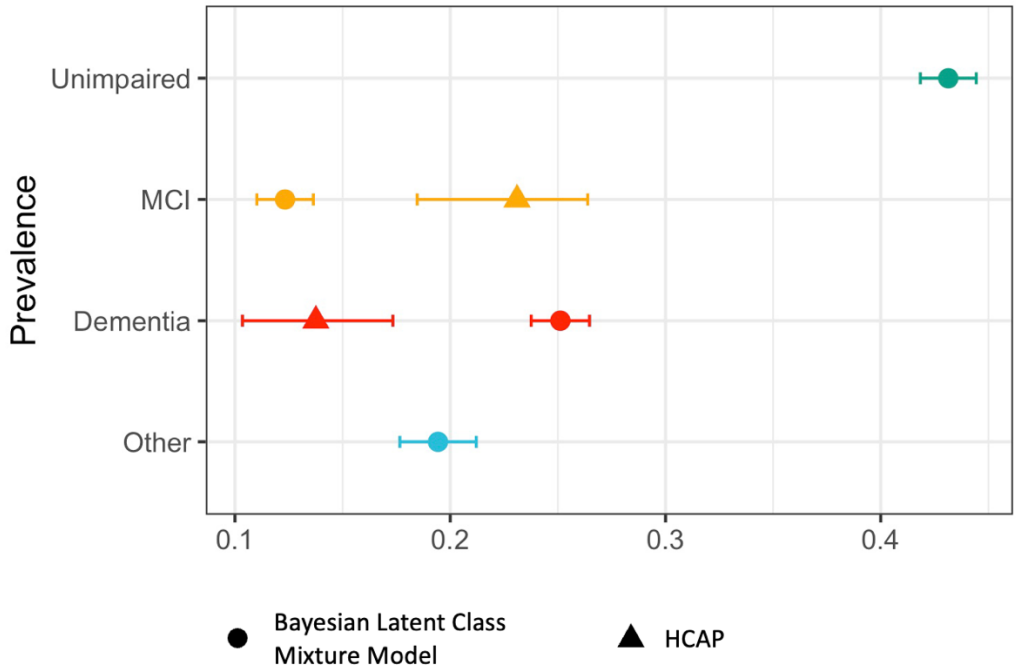


Figure 6.4: Algorithmic dementia classification results in HCAP 2016 70+ using the Bayesian latent class mixture model compared with results reported by HCAP investigators.

HCAP investigators estimated that in HCAP participants aged 70+ (n=2675), participants with Dementia comprised 13.8% (95% CI: [10.3%, 17.3%]) and participants with MCI 23.1% (95% CI: [18.5%, 26.4%]) of the sample. Both methods estimated any cognitive impairment (Dementia + MCI) in the 70+ HCAP 2016 sample to be about 36%.

6.3.5: Risk factor associations with dementia and MCI

HCAP investigators reported results for associations between several risk factors of dementia and MCI based on their classification algorithm. Recall that HCAP analyses were not conducted in the same sample as analyses for the Bayesian latent class mixture model because I restricted analyses in HCAP to participants aged 70+ and removed participants missing data on variables other than neuropsychological assessments. In contrast, HCAP investigators algorithmically classified cognitive impairment for the complete HCAP sample. Thus, HCAP results include younger participants compared to the sample I used in the Bayesian latent class mixture model. The results comparison serves as a sanity check of the plausibility of synthetic HCAP datasets generated from the Bayesian latent class mixture model and is not intended as a comparison between equivalent samples.

That said, I attempted to make estimates as comparable as possible between the two analyses. Since HCAP investigators reported odds ratios (ORs), I fit logistic regression models to estimate ORs of dementia and MCI for each risk factor in each of the 1000 synthetic datasets created by the Bayesian latent class mixture model for each imputed HCAP dataset. Mean log(OR) and standard errors (SE) for each imputed dataset were obtained by averaging point estimates and standard errors across the 1000 synthetic datasets created for each imputed dataset. Point estimates and standard errors were then combined across imputed datasets using Rubin's rules and exponentiated to obtain ORs and 95% interval estimates. Results are presented in **Table 6.1**.

Overall, risk factor associations based on algorithmic dementia classifications from the Bayesian latent class mixture models were within range of estimates based on the HCAP algorithm. In line with HCAP estimates, I estimated no sex/gender differences in dementia in the

HCAP 70+ sample. In contrast to HCAP estimates, I estimated protective associations between identifying as female and MCI. In line with HCAP estimates, I estimated higher risk of dementia and MCI for Black and Hispanic participants. Protective associations between educational attainment and dementia and MCI were replicated nearly identically in both studies.

	Bayesian Latent Class Mixture Model[1]		HCAP Algorithm[2]	
	MCI OR (95% CI)	**Dementia OR (95% CI)**	**MCI OR (95% CI)**	**Dementia OR (95% CI)**
Age (5 years)	1.32 (1.21, 1.45)	1.60 (1.48, 1.72)	1.17 (1.09, 1.26)	1.95 (1.77, 2.14)
Sex/gender				
Female	0.65 (0.51, 0.83)	0.97 (0.80, 1.16)	0.9 (0.80, 1.20)	1.1 (0.80, 1.40)
Male	Ref	Ref	Ref	Ref
Race/Ethnicity				
Black	1.49 (1.10, 2.02)	1.4 (1.11, 1.77)	1.0 (0.80, 1.40)	1.81 (1.20, 2.75)
Hispanic	1.71 (1.19, 2.45)	1.09 (0.80, 1.48)	1.42 (1.03, 1.96)	1.10 (0.70, 1.70)
White	Ref	Ref	Ref	Ref
Education (1 year)	0.96 (0.93, 0.99)	0.92 (0.90, 0.95)	0.94 (0.91, 0.97)	0.93 (0.89, 0.97)

[1]Obtained by fitting crude logistic regression models in the HCAP 70+ sample
[2]Obtained by fitting crude weighted logistic regression models in the complete HCAP sample

Table 6.1: Odds ratios (OR) and 95% interval estimates for the association between dementia and MCI risk factors resulting from algorithmic dementia classification using the Bayesian latent class mixture model compared to the HCAP algorithm.

6.4: Discussion

Estimates of dementia prevalence in the 70+ HCAP 2016 sample using the Bayesian latent class mixture model were higher and estimates for MCI lower compared to the algorithm developed by HCAP investigators. Estimates for impaired (MCI + Dementia) vs. unimpaired were consistent between both models, however.

Dementia is a progressive condition, and it is difficult to make sharp distinctions between MCI and dementia in study participants. Thus, the cutoff between classifying a participant as having MCI versus having dementia can be fuzzy and is often up to clinician's judgement. The

gold-standard clinical dementia adjudication procedure used in cohort studies is an attempt at triangulating evidence and arriving at a consensus diagnosis based on the assessment from several medical professionals (e.g., geropsychiarists, neurologists, and cognitive neuroscientists). The switch observed between estimated proportions of dementia and MCI using the Bayesian latent class mixture model vs. the HCAP algorithm may reflect the 2011 updated criteria for dementia diagnosis which incorporated updated scientific evidence for earlier stages of the disease and recognized that cognitive impairment in domains other than memory may signal onset of the condition (Jack et al. 2011; Albert et al. 2011; McKhann et al. 2011). Differences between results from the Bayesian latent class mixture model and the HCAP algorithm could also represent changing trends in clinician's perspectives on the condition and how they choose to adjudicate cases over time. For example, if the HCAP team adjudicated ADAMS participants, they may have classified participants differently given updated findings and a deeper understanding of progression to dementia.

Prior distributions in the Bayesian latent class mixture model were specified using information from the ADAMS study which was conducted nearly 20 years prior to HCAP (and prior to the updated criteria) but is currently the best resource for gold-standard clinical dementia adjudication because. Any characteristics of the diagnostic process used in the study from which priors are specified would be carried through to synthetic datasets created by the Bayesian latent class mixture model.

I have increased confidence in the fidelity of synthetic datasets generated by the Bayesian latent class mixture model since risk factor associations with dementia and MCI mostly reflected patterns reported by HCAP investigators. However, the results presented regarding some evidence of lack of fit in posterior predictive distributions for continuous and categorical

variables together with the results in Chapter 5 that showed increased dementia classification accuracy for Black and Hispanic participants when subset of HCAP were adjudicated and used as priors make a compelling case for adjudicating subsets of HCAP in the near future. Based on Chapter 5 results, adjudicating around 35% of the HCAP study should be sufficient to create synthetic datasets that reflect true cognitive impairment class proportions and lead to valid population-level inferences.

Chapter 7 Future Directions

The area of algorithmic dementia classification has abundant opportunities for conducting immediately relevant and highly impactful research. Methods developed for this dissertation project are directly applicable to existing studies on AD/ADRD. Specifically, the HRS HCAP study was designed to be harmonizable with studies including the Rush Memory and Aging Project (Bennett et al. 2012), the 10/66 studies focused on dementia epidemiology in low- and middle-income countries (Prina et al. 2019), and the HRS ADAMS Study (Langa et al. 2005). Versions of HCAP are also being conducted in Mexico (Mex-Cog, (Mejia-Arango et al. 2020)), India (LASI-DAD, (Lee and Dey 2020)), England (ELSA, (Steptoe et al. 2013)), China (CHARLS, (Zhao et al. 2014)), and South Africa (HAALSI, (Gómez-Olivé et al. 2018)), thus, methods for strengthening algorithmic dementia classification in US studies like HRS have the potential to positively impact methods used in international studies of AD/ADRD.

The ADAMS study is currently the best source of gold-standard clinical dementia adjudication from which to specify prior distributions in the Bayesian latent class mixture model because it is the first study of its kind to perform gold-standard dementia adjudication in population-representative subsample of a study with participants from different regions of the US and using standardized protocol (Langa et al. 2005). Since ADAMS was conducted 20 years ago, however, it is reasonable to assume that there may be potential cohort effects (shifting distributions in key characteristics over time or changes in diagnostic practices) that would impact the validity of using the ADAMS prior in analyses of more updated studies. Further, ADAMS was sampled using a different sampling frame than the HCAP study and there was some evidence of lack of fit for categorical and continuous variables in the application of the Bayesian latent class mixture model to the HCAP study (**Chapter 6**).

Though the results for algorithmic dementia classification in HCAP using the Bayesian latent class mixture model seemed reasonable, the simulation results in **Chapter 5** related to adjudicating subsets of HCAP for use as better-calibrated prior distributions makes a compelling case for adjudicating subsets of the HCAP study in the near future to improve the quality of inferences, especially for Black and Hispanic participants.

An important area of research in algorithmic dementia classification is the development of algorithms that do not differentially misclassify participants by race/ethnicity and are thus appropriate for use in disparities research (Gianattasio et al. 2019; Gianattasio, Ciarleglio, and Power 2020). An especially difficult challenge in algorithmic dementia classification is the estimation of dementia prevalence for Hispanic participants who typically comprise a very small subset of cohort studies. Simulation studies that specified prior distributions using the ADAMS study showed increased bias in estimates for Black and Hispanic participants compared to White participants and inferences on racial/ethnic differences in dementia were only correct for Black vs. White participants (**Chapter 5**). Adjudicating subsets of HCAP, however, significantly reduced bias in estimations of dementia prevalence and racial/ethnic differences in dementia for Black and Hispanic participants. In those simulation studies, inferences on racial/ethnic differences for Black vs. White and Hispanic vs. White participants were correct, which shows a promising way forward for improving this framework for use in disparities research (**Chapter 6**).

The strategy for algorithmic dementia classification outlined in this dissertation apply to cross-sectional analyses, but longitudinal trends in functional ability and cognition are essential for diagnosing dementia with increased confidence (Langa et al. 2005). A valuable next step for this research would be incorporating longitudinal data in the Bayesian latent class mixture modeling framework. As discussed in **Section 2.3.1:**, the appeal of embedding the algorithmic

dementia classification model in a Bayesian analysis framework is the ability of Bayesian frameworks to handle complicated modeling tasks. Adding a longitudinal aspect to the model will inevitably increase computational challenges, but there is a wealth of research on methods for handling situations where sampling from posterior distributions is less straight-forward than the sampling implemented in this proposal (Gelman et al. 2014).

Implications of findings from this dissertation project extend beyond algorithmic dementia classification in HRS. First, the methods in this project could be extended to perform algorithmic dementia classification in non-HRS studies like the National Health and Aging Trends Study (NHATS) which does not contain a substudy of clinically adjudicated dementia cases (Freedman and Kasper 2019). Implementing these methods in an external HRS dataset may require additional model flexibility to capture features of the new dataset. Strategies may include relaxing covariance assumptions such that covariances of continuous variables are allowed to vary not only latent by classes but also by contingency cell membership (Liu and Rubin 1998). Using prior distributions for latent class membership based solely on the ADAMS may not be desirable when moving to external datasets, depending on the characteristics of the target dataset compared to ADAMS. Alternatively, prior distributions could be specified based on existing studies or the combination of multiple studies through data fusion (Saporta 2002), integrative analysis (Bazeley 2011), or meta-analysis methods (Hedges 1992). Second, the creation of fully synthetic datasets with detailed neuropsychological measures and assigned impairment classes expands the potential uses for products from the Bayesian latent class mixture model. For example, investigators conducting studies with data unavailable for public use due to privacy concerns could use this framework to create fully synthetic versions of their studies. Further,

rich, high quality synthetic dataset enable researchers to explore a broader set of questions related to sociodemographic characteristics, cognition, and impairment.

In conclusion, methods developed in this project lay important groundwork for improving methods for algorithmic dementia classification, which are important for overcoming the barriers that gold-standard methods have created for understanding trends in AD/ADRD risk factors, incidence, and prevalence in population-representative studies. Outcomes from this work create exciting opportunities for using existing data sources that to date have not been primarily used to conduct AD/ADRD research and strengthen inferences based on algorithmic dementia classification in studies that currently rely on models for classifying cognitive impairment in participants.

Appendix A Details on Neuropsychological Assessments and ADAMS Cognitive Impairment Categories

Measure	Dataset	Range	Prompt; Scoring
Backwards count 20 (Brandt, Spencer, and Folstein 1988)	HRS	[0, 2]	Count backwards as quickly as possible for 10 continuous numbers starting at 20 (2 trials; 2 pts=correct on first try; 1pt=correct on second try)
Serial 7s (Brandt, Spencer, and Folstein 1988)	HRS	[0, 5]	Subtract 7 from 100 and continue for 5 trials; 1 pt for each correct subtraction
Item naming (scissors) (Brandt, Spencer, and Folstein 1988)	HRS	[0, 1]	"What do you usually use to cut paper?"; 1pt for correct
Item naming (cactus) (Brandt, Spencer, and Folstein 1988)	HRS	[0, 1]	"What do you call the kind of prickly plant that grows in the desert?"; 1pt for correct
President naming (Brandt, Spencer, and Folstein 1988)	HRS	[0, 1]	"Name the current president of the US"; 1 pt for correct
Vice President naming (Brandt, Spencer, and Folstein 1988)	HRS	[0, 1]	"Name the current vice president of the US"; 1 pt for correct
Word list recall (Immediate)	HRS	[0, 10]	Adapted from CERAD word list (Morris et al. 1989); Participant is visually shown 10 high-imagery words for 2 seconds each; 1pt for each correctly recalled word
Word list recall (Delayed)	HRS	[0, 10]	Participant was asked to recall list of 10 words from recall task after answering other survey questions; 1pt for each correctly recalled word
Backwards count 86 (Brandt, Spencer, and Folstein 1988)	ADAMS	[0, 2]	Count backwards as quickly as possible for 10 consecutive numbers starting at 86 (2 trials; 2 pts=correct on first try; 1pt=correct on second try)
Mini-mental State Exam (Folstein, Folstein, and McHugh 1975)	ADAMS HCAP	[0, 30]	20-item test of global cognition in five domains: memory, calculation, space and time orientation, language, and word recognition
Normalized Mini-mental State Exam (Philipps et al. 2014)	ADAMS HCAP	[0, 100]	Normalized Mini-Mental State Exam scores
Animal naming (Morris et al. 1989)	ADAMS HCAP	[0, 33]	"Name as many animals as you can in 1 minute"; 1pt for each animal named
Boston naming test (Morris et al. 1989)	ADAMS	[0, 15]	Participant was shown line drawings of common man-made and naturally occurring objects; 1pt for each correctly named item

Word list recognition (yes)	ADAMS HCAP	[0, 10]	1pt for each correctly identified word that was included in the word list from the recall task
Word list recognition (no)	ADAMS HCAP	[0, 10]	1pt for each correctly identified word that was not included in the word list from the recall task
Story recall (Immediate) (Elwood 1991)	ADAMS HCAP*	[0, 37] [0, 35]	Participants were read two brief stories and asked to recall as much of the story as possible; 1pt for each correctly recalled detail
Story recall (Delayed)	ADAMS HCAP*	[0, 37] [0, 35]	Participant was asked to recall story details after answering other survey questions; 1pt for each correctly recalled detail
Constructional praxis (Immediate)	ADAMS HCAP	[0, 11]	Adapted from Rosen et al. (1984); Participant is asked to draw four geometric forms of varying difficulty (circle=2pts; overlapping rectangles=3pts; diamond=2pts; cube=4 pts); 1 pt for each correctly drawn feature
Constructional praxis (Delayed)	ADAMS HCAP	[0, 11]	After answering other survey questions, participant is asked to draw four geometric forms of varying difficulty from the previous task from memory (circle=2pts; overlapping rectangles=3pts; diamond=2pts; cube=4 pts); 1 pt for each correctly drawn feature
Symbol/Digit substitution (Smith 1968)	ADAMS HCAP	[0, 63]	Participants are asked to substitute geometric figures with digits 1-9; a key is provided with each digit symbol pairing; 1 pt for each correct substitution
Trails A (Reitan and Wolfson 1992)	ADAMS HCAP	[0, 373]	Participants are asked to draw a line connecting consecutively numbered circles on a sheet; score is number of seconds for task completion
Trails B (Reitan and Wolfson 1992)	ADAMS HCAP	[0, 727]	Participants are asked to draw a line connecting consecutively numbered and lettered circles, alternating between letters and numbers; score is number of seconds for task completion

*One of the brief stories used in HCAP differs from ADAMS, so the total possible score is 35 in HCAP instead of 37.

Table A.1: Descriptions of detailed neuropsychological and general cognitive assessment items available in the HRS, ADAMS, and HCAP studies. HRS items area available for HCAP and ADAMS participants as well.

Collapsed Category	Wave A n	ADAMS Adjudicated Cognitive Impairment Category (% of category)
Unimpaired	211	normal/non-case (100%)
MCI	65	mild cognitive impairment (100%)
Dementia	158	possible AD (41.1%), probable AD (36.1%), probable vascular dementia (8.9%), possible vascular dementia (8.2%), dementia of undetermined etiology (5.1%), alcoholic dementia (0.6%),
Other	86	other medical conditions (38.4%), stroke (24.4%), impairment secondary to vascular disease (15.1%), other neurological conditions (7.0%), depression (5.8%), alcohol abuse (past) (3.5%), alcohol abuse (current) (1.2%), mental retardation (2.3%), Parkinson's (1.2%), normal pressure hydrocephalus (1.2%)

Table A.2: Mapping from ADAMS adjudicated cognitive impairment category to collapsed cognitive impairment categories (Unimpaired, MCI, Dementia, Other).

Sociodemographic Characteristics	Neuropsychological Exam and Cognition	Health and Health Behaviors
Age	Total MMSE Score	Stroke History (yes/no)
Sex/Gender	Backwards Count (20, 86)	Hypertension (yes/no)
Race/Ethnicity	Serial 7s	Diabetes (yes/no)
Education	Item Naming (scissors, cactus)	Heart disease (yes/no)
Marital Status	President/VP Naming	BMI
Retirement Status	Animal Naming	IADLs
	Boston Naming Test	ADLs
	Word Recall (Immediate, Delayed)	Depression
	Word list recall (Yes, No)	Smoking
	Story Recall (Immediate, Delayed)	Alcohol Use
	Constructional Praxis (Immediate, Delayed)	
	Symbol/Digit Substitution	
	Trails (A, B)	
	Subjective Change in Cognition	
	Average Proxy Cognition	

Table A.3: Candidate variables for inclusion in multi-part models of predicted ADAMS cognitive impairment classes (Unimpaired, Other, MCI, Dementia).

A.1: Normalized Mini Mental State Exam scores

MMSE is notorious for having ceiling and floor effects (Philipps et al. 2014), and the ceiling effect can be seen in ADAMS Wave A— 17 % of the sample who took the assessment obtained either a perfect score of 30 or a near perfect score of 29 on the MMSE. Additionally, raw MMSE scores have the undesirable metrological property of inconsistent sensitivity to changes at different points on the scale. MMSE is often included in analyses with underlying normality assumptions despite these clear distributional violations. The normalizing transformation developed by Philipps et al. (2014) transforms the raw MMSE scale ([0, 30]) to [0, 100]. The transformation was achieved by fitting a latent process mixed model where the outcome of interest was latent cognitive level; the model was validated in external samples. The transformation is available in the R package NormPsy (Proust-Lima and Philipps 2018).

Appendix B Handling Missing Data

Missing data is a common challenge in longitudinal studies and if ignored can lead to bias and loss of precision in analyses. Missing data are categorized by the mechanisms assumed to have caused them. Data that are missing independently of all observed and unobserved variables are said to be missing completely at random (MCAR); data that are missing due to observed variables only are said to be missing at random (MAR); and data that are missing due to unobserved variables are said to be missing not at random (MNAR) (Rubin 1987).

For all missing data mechanisms, ignoring missing data by performing a complete-case analysis will lead to precision loss. When data are MCAR, complete-case analyses will be imprecise but unbiased. When data are MAR, the amount of bias in a complete-case analysis depends on several factors including how strongly the missing data mechanism is related to the outcome of interest; when data are MNAR, however, results will be generally be biased (Shaw et al. 2022). The challenge in practice is that we cannot test our assumption of the missing data mechanism, and it is likely the case that the missing data arise from a combination of mechanisms. Thus, it is generally regarded best practice to address the missing data in some way.

In this dissertation, I was concerned with missing data leading to selected subsamples of HRS, HCAP, and ADAMS studies that would bias results, especially when missing data occurred in cognitive variables which are highly correlated with cognitive impairment status. In the ADAMS study, assuming cognitive variables are missing due to cognitive impairment status amounts to an MAR missing data assumption since impairment status is observed for each participant. In HCAP and HRS, this would be an MNAR assumption since we do not have clinical impairment status adjudication for each participant.

Imputing missing data is a strategy for creating a dataset that could be analyzed as if the data were completely observed. There are several data imputation strategies available, and I chose different strategies for each dataset based on the Bayesian latent class mixture modeling framework and the information available in observed data. Details on the specific imputation methods used in this dissertation follow.

B.1: Multiple imputation using fully conditional specification with predictive mean matching

I used multiple imputation with chained equations (MICE) to impute missing data in the ADAMS study. Contrasted with single imputation, which imputes one value for missing data and usually leads to artificially precise estimates (Okpara et al. 2022), MI creates many imputed datasets in which data are analyzed. The "chained equations" portion refers to the definition of separate conditional for each variable to be imputed. This method is also referred to in the literature as "fully conditional specification." The MICE algorithm cycles through all variables to be imputed several times to produce imputations. MI results are combined accounting for within-imputation and between-imputation variance using "Rubin's Rules" (Rubin 1987). Let $\theta_m,\ m = 1, \dots, M$ be an estimate of a quantity of interest from one of M imputed datasets. Then, the following formulas are used to calculate components of a pooled estimate across M imputed datasets:

Pooled mean: $\bar{\theta} = \frac{1}{M}(\sum_{m=1}^{M} \theta_m)$

Within-imputation variance: $V_W = \frac{1}{M}\sum_{m=1}^{M} SE_i^2$

Between-imputation variance: $V_B = \frac{\sum_{m=1}^{M}(\theta_i - \bar{\theta})^2}{M-1}$

Pooled SE: $SE_{pooled} = \sqrt{V_W + V_B + \frac{1}{M}V_B}$

Pooled 95% interval estimates are constructed by taking $\bar{\theta} \pm 1.96\sqrt{SE_{pooled}}$

Since MI methods are widely available across software and have been made increasingly accessible in recent years, MI is generally preferred over single imputation methods. The `mice` package in R provides several MI options that vary predominantly in statistical models used to estimate relationships between variables and are often classified by data type (numeric, binary, ordered, unordered, or any) and structure (e.g., longitudinal) (van Buuren 2019). I chose predictive mean matching (PMM), which is a semi-parametric method that is robust to model misspecification and produces imputations within range of the observed data (L. Tang et al. 2005; Marshall, Altman, and Holder 2010; Kleinke 2017; Shaw et al. 2022). PMM calculates a predicted value of the variable to be imputed for observed and missing participants using linear regression. Missing values are imputed by randomly choosing an observed value from a pool of "nearest neighbor" donors based on proximity between predicted values for observed participants and predicted value of the missing data point and user-specified pool size. For dissertation analyses, I used the `mice` package default of 10 donors.

B.2: Single imputation with stratified hotdeck

I used single imputation methods for imputing neuropsychological data in HCAP and the superpopulation to simplify analyses in the simulation study. As an attempt to impute values based on observations from as similar a unit as possible, I implemented a stratified hotdeck procedure with strata defined by variables that seemed highly correlated with the variable to be imputed.

Hotdeck imputation is a more general case of PMM discussed above. Where PMM uses predicted values from linear regression to define pools of similar donors, general hotdeck procedures can use other metrics to define "nearest neighbors." The stratified hotdeck procedure I implemented in the study matched participants based on cross-classification of important characteristics for the imputation. Continuous variables were binned for the matching procedure. There is no consensus on the best way to implement hotdeck, so I used general principles from MI and PMM. The quality of hotdeck imputations increases with matching precision; however, overly refined matches can result in identical imputed values for several missing observations which undermines the imputation process. It is also possible to refine categories to such an extent that donor pools are empty because the observation missing values what the only one with a specific combination of characteristics. Thus, in this analysis, I required donor pools of at least 15 observations.

See Andridge and Little (2010) for a broad review of hotdeck imputation techniques and implementations.

Appendix C Sampling Distribution Derivations

What follows is a detailed derivation of the synthetic data-generating model described in **Section 3.3.2:**. Let $G_i, i = 1, \dots, 364$ denote the ADAMS adjudicated impairment class (group) for each participant in the ADAMS training sample,

$$G_i = \begin{cases} 1 & \text{if participant } i \text{ is Unimpaired} \\ 2 & \text{if participant } i \text{ has Other impairment} \\ 3 & \text{if participant } i \text{ has MCI} \\ 4 & \text{if participant } i \text{ has Dementia} \end{cases} \quad \text{(C.1)}$$

The mixture distribution we wish to sample from is

$$f(X|\theta_1, \dots\ \theta_4) = \sum_{G=1}^{4} \lambda_G f_G(X), \quad \text{(C.2)}$$

where X represents both categorical and continuous covariates to be modeled (**Table 3.1**), θ_G is the set of parameters for the model in each latent class, and λ_G are the mixture probabilities for the densities f_G. Following the notation of (Schafer 1997), let W_1 and W_2 be the categorical variables race/ethnicity (White, Black, Hispanic) and stroke history (ever/never), and let $Z_1, Z_2, \dots Z_{10}$ be the continuous variables. Then $X = (W, Z)$ is an $n \times 12$ matrix of observed data. Let $C = \{c_d : d = 1, 2, \dots 6\}$ be the vector of observed counts for each contingency cell determined by cross classification of the two categorical variables race/ethnicity and stroke history within each impairment group G and let U be an $n \times 6$ matrix with rows u_i^T, where u_i is a 6-vector with a 1 in position d if observation i falls into cell d and 0s in all other position. All the information about W is contained in C, U, or $U^T U$ = diag(C), and thus we can rewrite the mixture distribution in **Equation (C.2)** as

$$f(X|\theta_1, \dots\ \theta_4) = \sum_{G=1}^{4} \lambda_G f_G(W, Z|\theta_G) = \sum_{G=1}^{4} \lambda_G f_G(C, Z|\theta_G)$$

$$= \sum_{G=1}^{4} \lambda_G f_G(Z|C, \theta_G) f_G(C|\theta_G), \tag{C.3}$$

where $f_G(Z|C, \theta_G)$ is matrix normally distributed, $f_G(Z|C, \theta_G) \sim MN(A\mathrm{B_G},\ V_G,\ \Sigma_\mathrm{G})$, and $f_G(C|\theta_G)$ is multinomially distributed, $f_G(C|\theta_G) \sim M(n_G, \pi_G)$. Following the guidance of (Schafer 1997) for embedding the multi-part model in a Bayesian analysis, conjugate priors for $f_G(Z\,|C, \theta_G) f_G(C|\theta_G)$ were chosen to be independent of each other so that the posteriors would remain independent as well. Matrix normal and inverse Wishart priors were used for $\mathrm{B_G}|\Sigma_\mathrm{G}$ and Σ_G, respectively ($\mathrm{B_G}|\Sigma_\mathrm{G} \sim\ MN(\mathrm{B}_0,\ V_{0_G},\ \Sigma_\mathrm{G}/\kappa_0)$ and $\Sigma_G \sim W_{\nu_0}^{-1}(\Lambda_{0_G}^{-1})$). A Dirichlet prior ($D(\alpha_G)$) was used for $f_G(C|\theta)$. Simulated values of the parameters α_G, V_{0_G}, Σ_G, and $\Lambda_{0_\mathrm{G}}^{-1}$ were based on bootstrapped samples of the ADAMS data as if the posteriors were equal to the prior, and κ_0, ν_0 were hyperparameters in the model.

The full posterior distribution of the model parameters is

$f(\theta_1, \ldots\ \theta_4|\ X) \propto$ likelihood $\times$ prior

$$= f(X\ |\lambda_1, \ldots\ \lambda_4, \pi_1, \ldots\ \pi_4, \mathrm{B}_1, \ldots\ \mathrm{B}_4, \Sigma_1, \ldots \Sigma_4)\ \times$$

$$f(\lambda_1, \ldots\ \lambda_4, \pi_1, \ldots\ \pi_4, \mathrm{B}_1, \ldots\ \mathrm{B}_4, \Sigma_1, \ldots \Sigma_4)$$

$$= f(X\ |\lambda_1, \ldots\ \lambda_4, \pi_1, \ldots\ \pi_4, \mathrm{B}_1, \ldots\ \mathrm{B}_4, \Sigma_1, \ldots \Sigma_4)\ \times$$

$$f(\mathrm{B}_1, \ldots\ \mathrm{B}_4|\Sigma_1, \ldots \Sigma_4, \pi_1, \ldots\ \pi_4, \lambda_1, \ldots\ \lambda_4) f(\Sigma_1, \ldots \Sigma_4|\pi_1, \ldots\ \pi_4, \lambda_1, \ldots\ \lambda_4)\ \times$$

$$f(\pi_1, \ldots\ \pi_4|\lambda_1, \ldots\ \lambda_4) f(\lambda_1, \ldots\ \lambda_4),$$

Given the predicted mixture probabilities, $\lambda_1, \ldots\ \lambda_4$, modeling takes place independently within each subgroup which implies the following posterior distribution:

$$f(\theta_1, \ldots\ \theta_4\ |\ X) \propto \prod_{G=1}^{4} [f_G(X\ |\lambda_G, \pi_G, \mathrm{B_G}, \Sigma_\mathrm{G})\ \times f_G(\mathrm{B_G}|\Sigma_\mathrm{G}, \pi_G, \lambda_G) f_G(\Sigma_\mathrm{G}|\pi_G, \lambda_G) f_G(\pi_G|\lambda_G)]\ \times$$

$$f(\lambda_1, \ldots\ \lambda_4)$$

After accounting for additional independence assumptions in the data-generating model (**Figure 3.3**) and taking advantage of representing $f_G(X)$ as $f_G(Z\,|C)f_G(C)$, the posterior distribution can be simplified into the following pieces within each latent group:

$$f(\theta_1, \dots\ \theta_4|\ X) \propto \prod_{G=1}^{4}[f_G(Z\,|C, \lambda_G, \pi_G, \mathrm{B_G}, \Sigma_\mathrm{G}) f_G(C|\lambda_G, \pi_G, \mathrm{B_G}, \Sigma_\mathrm{G})$$

$$\times f_G(\mathrm{B}_G|\Sigma_\mathrm{G}, \pi_G, \lambda_G) f_G(\Sigma_\mathrm{G}|\pi_G, \lambda_G) f_G(\pi_G|\lambda_G)] f(\lambda_1, \dots\ \lambda_4)$$

$$= \prod_{G=1}^{4}[$$

$$f(Z\,|C, \mathrm{B_G}, \Sigma_\mathrm{G}) \times f(\mathrm{B}_G|\Sigma_G) f(\Sigma_G|\lambda_G) \times \quad \text{(C.4)}$$

$$f(C|\pi_G, \lambda_G) f(\pi_G|\lambda_G)] \times \quad \text{(C.5)}$$

$$f(\lambda_1, \dots, \lambda_4) \quad \text{(C.6)}$$

Posterior sampling distributions will be derived for each of the equations above. Derivations for **(C.4)** and **(C.5)** will be conditional on belonging to impairment group G, which is determined by λ_G. For notational convenience, the index G and parameter λ_G will be removed in their derivations.

For **(C.6)**, the distribution of mixture probabilities is approximated by the multi-part model defined by **Equations (3.2)-(3.4)**.

For **(C.5)**, $f(C|\pi) \sim M(n, \pi)$, where n (the size of the impairment group) is determined by the predicted mixture probability λ_G. The conjugate Dirichlet prior was used, $f(\pi) \sim D(\alpha)$ thus, the posterior distribution is

$$f(\pi|C, \alpha) \sim D(\alpha + C).$$

It takes a little more work to derive the exact sampling distribution for **(C.4)**. The following conjugate priors were used:

$$f(\mathrm{B}|\Sigma) \sim MN_{r\times q}(\mathrm{B}_0, V_0, \Sigma/\kappa_0)$$

$$f(\Sigma) \sim W_{v_0}^{-1}(\Lambda_0^{-1}),$$

where $r = 4$ is the number of effects included in the ANOVA design matrix A for the restricted general location model (see **Section 2.3.3.2:**) and $q = 10$ is the number of continuous covariates in the model. With a matrix normal likelihood, we expect a matrix normal inverse Wishart posterior. What follows are key steps for deriving the posterior distribution:

Writing out posterior distribution,

$$f(\mathrm{B}, \Sigma|Z) \propto |\Sigma|^{-\frac{n}{2}} \exp\left\{-\frac{1}{2}\mathrm{tr}\Sigma^{-1}(Z - UA\mathrm{B})^T(Z - UA\mathrm{B})\right\} \times$$

$$|\Sigma|^{-\frac{v_0+q+1}{2}} \exp\left\{-\frac{1}{2}tr(\Lambda_0\Sigma^{-1})\right\} \times$$

$$|\Sigma|^{-\frac{r}{2}} \exp\{-\frac{\kappa_0}{2}\mathrm{tr}[\Sigma^{-1}(\mathrm{B} - \mathrm{B}_0)^{\mathrm{T}}\mathrm{V}_0^{-1}(\mathrm{B} - \mathrm{B}_0)]\}$$

grouping the parameters,

$$= |\Sigma|^{-\frac{n}{2}}|\Sigma|^{-\frac{v_0+q+1}{2}}|\Sigma|^{-\frac{r}{2}} \exp\left\{-\frac{1}{2}\mathrm{tr}(\Lambda_0\Sigma^{-1})\right\} \times$$

$$\exp\left\{-\frac{1}{2}tr[\Sigma^{-1}(Z - UA\mathrm{B})^T(Z - UA\mathrm{B})]\right\} \times$$

$$\exp\{-\frac{\kappa_0}{2}\mathrm{tr}[\Sigma^{-1}(\mathrm{B} - \mathrm{B}_0)^{\mathrm{T}}\mathrm{V}_0^{-1}(\mathrm{B} - \mathrm{B}_0)]\}$$

expanding all parenthesis with B terms,

$$= |\Sigma|^{-\frac{n}{2}}|\Sigma|^{-\frac{v_0+q+1}{2}}|\Sigma|^{-\frac{r}{2}} \exp\left\{-\frac{1}{2}\mathrm{tr}(\Lambda_0\Sigma^{-1})\right\} \times$$

$$\exp\left\{-\frac{1}{2}tr[\Sigma^{-1}(Z^TZ - 2\mathrm{B}^TA^TU^TZ + \mathrm{B}^TA^TU^TUA\mathrm{B})]\right\} \times$$

$$\exp\{-\frac{\kappa_0}{2}\mathrm{tr}[\Sigma^{-1}(\mathrm{B}^{\mathrm{T}}\mathrm{V}_0^{-1}\mathrm{B} - 2\mathrm{B}^{\mathrm{T}}\mathrm{V}_0^{-1}\mathrm{B}_0 + \mathrm{B}_0^{\mathrm{T}}\mathrm{V}_0^{-1}\mathrm{B}_0)]\}$$

Separating the B terms from the B_0 terms and factoring out operators,

$$= |\Sigma|^{-\frac{n}{2}}|\Sigma|^{-\frac{\nu_0+q+1}{2}}|\Sigma|^{-\frac{r}{2}}\exp\{-\frac{1}{2}\mathrm{tr}[(\Lambda_0 + Z^TZ + \kappa_0 \mathrm{B}_0^T V_0^{-1}\mathrm{B}_0)\Sigma^{-1}]]\} \times$$

$$\exp\{-\frac{1}{2}\mathrm{tr}[\Sigma^{-1}\{\mathrm{B}^{\mathrm{T}}(\mathrm{A}^{\mathrm{T}}\mathrm{U}^{\mathrm{T}}\mathrm{UA} + \kappa_0 \mathrm{V}_0^{-1})\mathrm{B} - 2\mathrm{B}^{\mathrm{T}}(A^TU^TZ - \kappa_0 \mathrm{V}_0^{-1}\mathrm{B}_0)\}]\} \quad \text{(C.7)}$$

We need to complete the square for the matrix normal distribution.

Let $M = (A^TU^TUA + \kappa_0 V_0^{-1})^{-1}$ and $m = (A^TU^TZ - \kappa_0 V_0^{-1}\mathrm{B}_0)$, then the term needed for completing the square is $c = -m^TMm$. Plugging this in to (**C.7**),

$$f(\mathrm{B}, \Sigma|Z) \propto |\Sigma|^{-\frac{n}{2}}|\Sigma|^{-\frac{\nu_0+q+1}{2}}|\Sigma|^{-\frac{r}{2}}\exp\left\{\frac{1}{2}\mathrm{tr}[(\Lambda_0 + Z^TZ + \kappa_0 \mathrm{B}_0 V_0^{-1}\mathrm{B}_0)\Sigma^{-1}]\right\} \times$$

$$\exp\{-\frac{1}{2}\mathrm{tr}[\Sigma^{-1}\{(\mathrm{B} - Mm)^TM^{-1}(\mathrm{B} - Mm) - m^TMm\}]\}$$

grouping the added term with the first exponential,

$$f(\mathrm{B}, \Sigma|Z) \propto |\Sigma|^{-\frac{(\nu_0+q+n+1)}{2}}\exp\left\{\frac{1}{2}\mathrm{tr}[(\Lambda_0 + Z^TZ + \kappa_0 \mathrm{B}_0 V_0^{-1}\mathrm{B}_0 - m^TMm)\Sigma^{-1}]\right\} \times$$

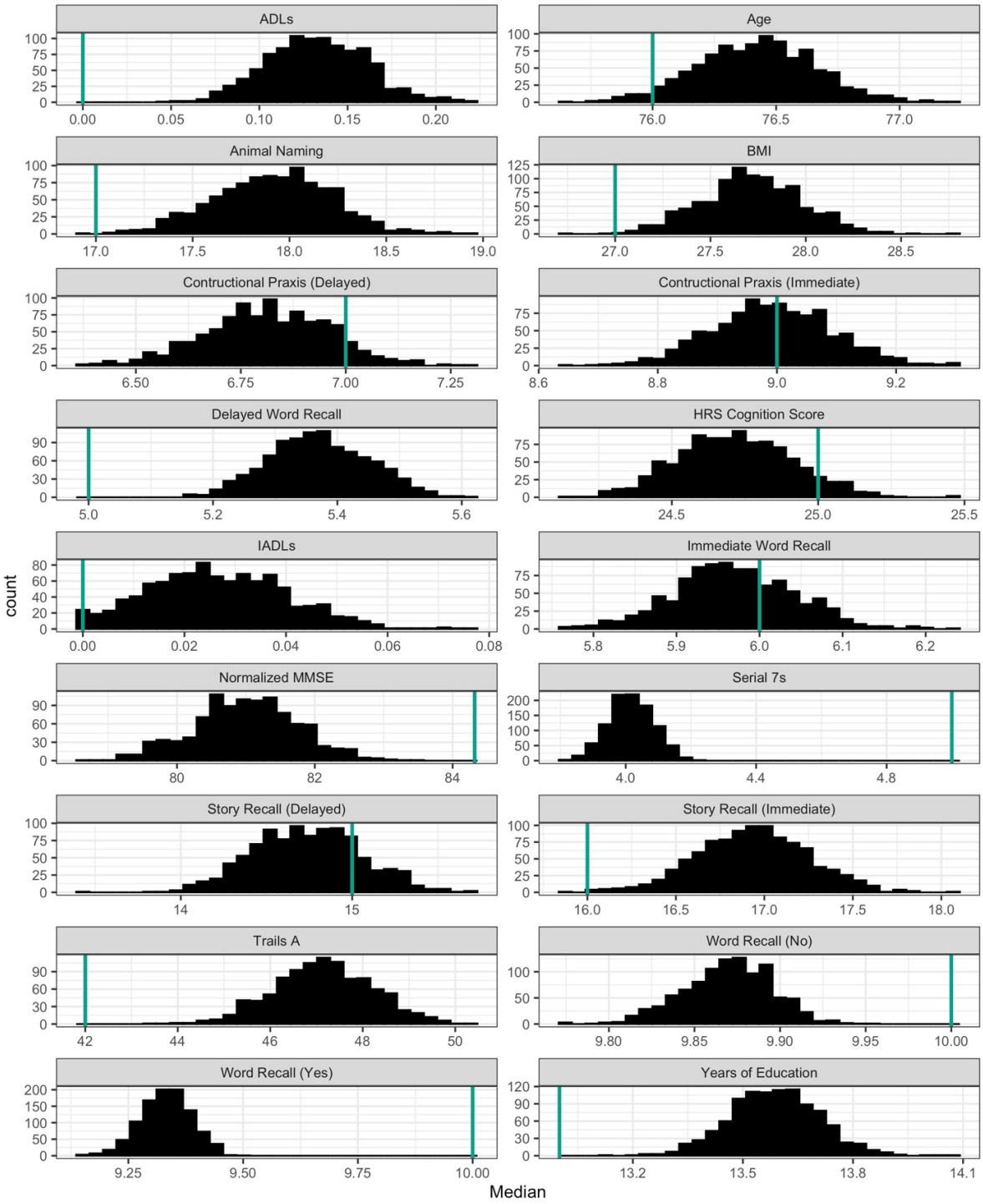
$$|\Sigma|^{-\frac{r}{2}}\exp\{-\frac{1}{2}\mathrm{tr}[\Sigma^{-1}\{(\mathrm{B} - Mm)^TM^{-1}(\mathrm{B} - Mm)\}]\}$$

$$\sim W_{\nu_0+n}^{-1}((\Lambda_0 + Z^TZ + \kappa_0 \mathrm{B}_0^T V_0^{-1}\mathrm{B}_0 - m^TMm)^{-1}) \times MN_{r\times q}(Mm, M, \Sigma)$$

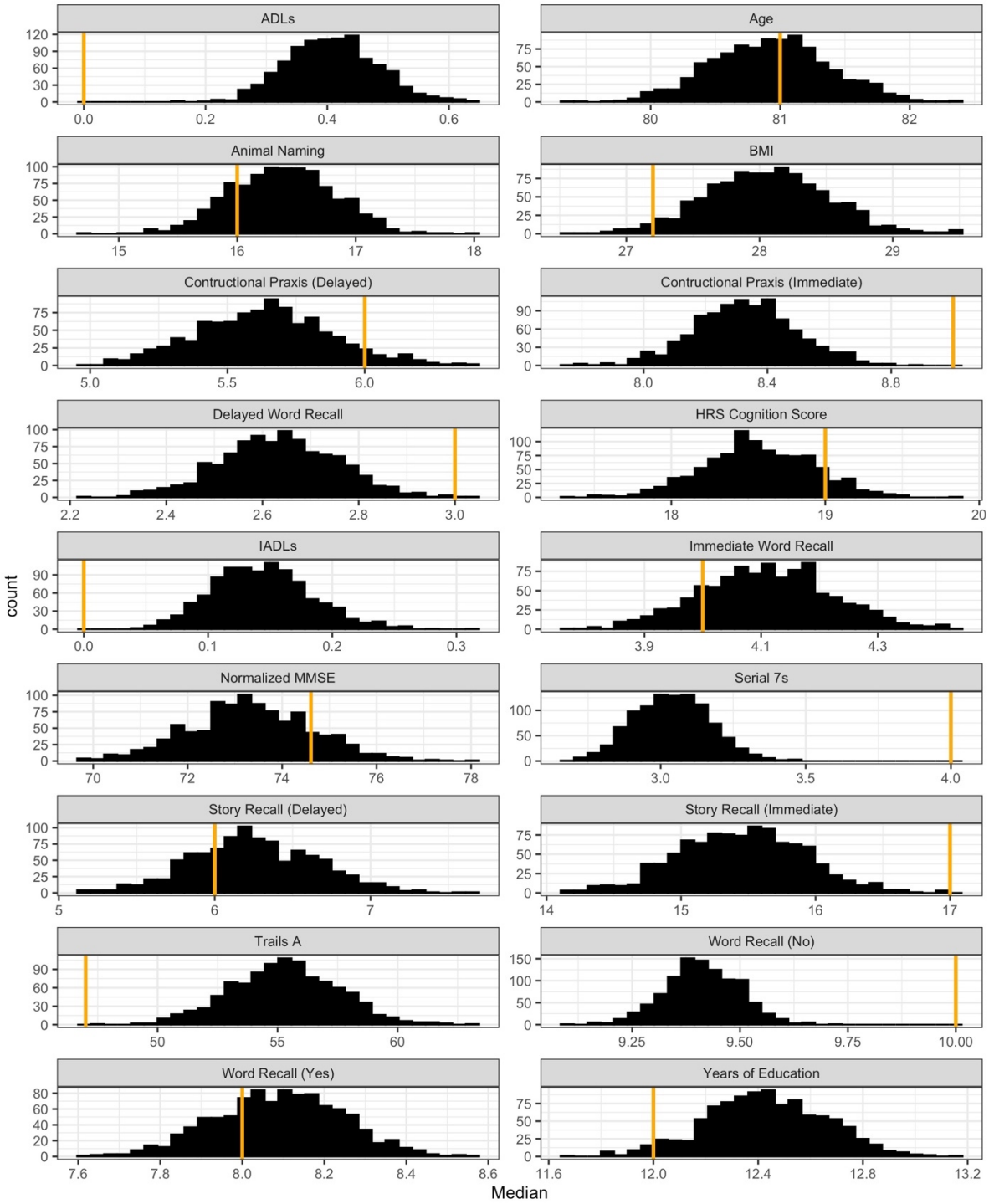
Appendix D Supplementary Figures

D.1: Chapter 4 supplementary figures

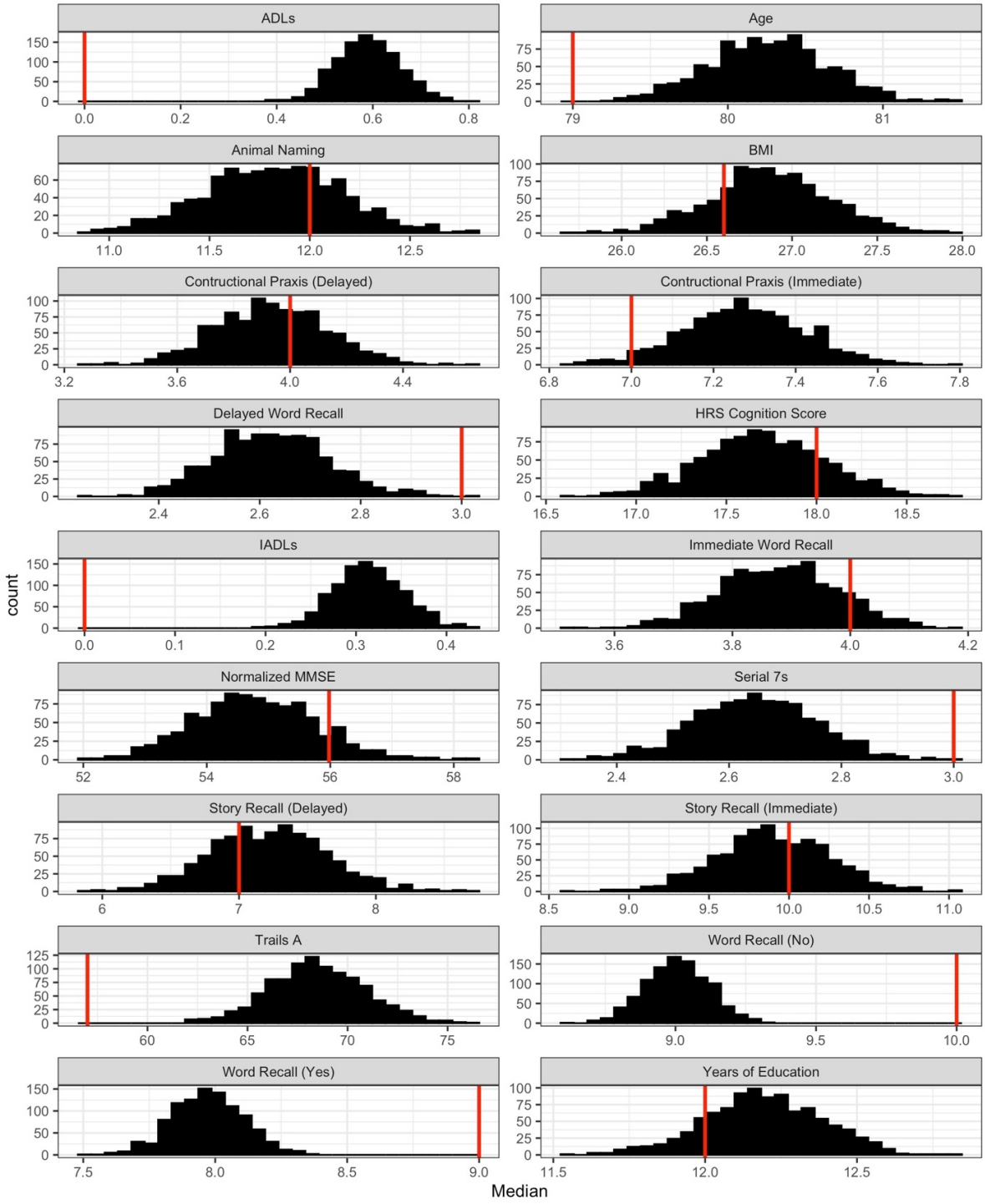
(a) Unimpaired



(b) MCI


ADLs
Age
Animal Naming
BMI
Contructional Praxis (Delayed)
Contructional Praxis (Immediate)
Delayed Word Recall
HRS Cognition Score
IADLs
Immediate Word Recall
Normalized MMSE
Serial 7s
Story Recall (Delayed)
Story Recall (Immediate)
Trails A
Word Recall (No)
Word Recall (Yes)
Years of Education
count
Median

(c) Dementia


ADLs
Age
Animal Naming
BMI
Contructional Praxis (Delayed)
Contructional Praxis (Immediate)
Delayed Word Recall
HRS Cognition Score
IADLs
Immediate Word Recall
Normalized MMSE
Serial 7s
Story Recall (Delayed)
Story Recall (Immediate)
Trails A
Word Recall (No)
Word Recall (Yes)
Years of Education
count
Median

(d) Other

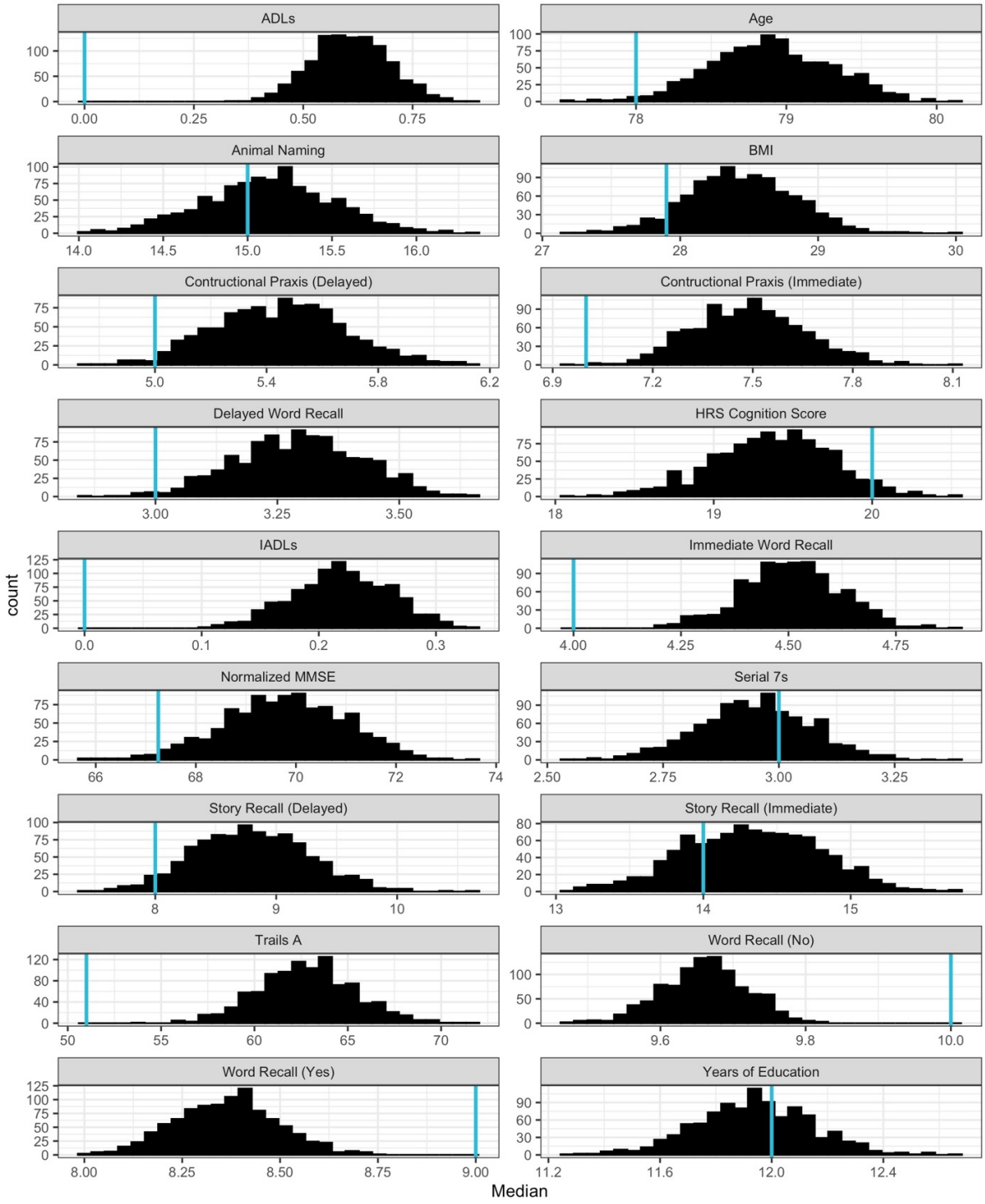
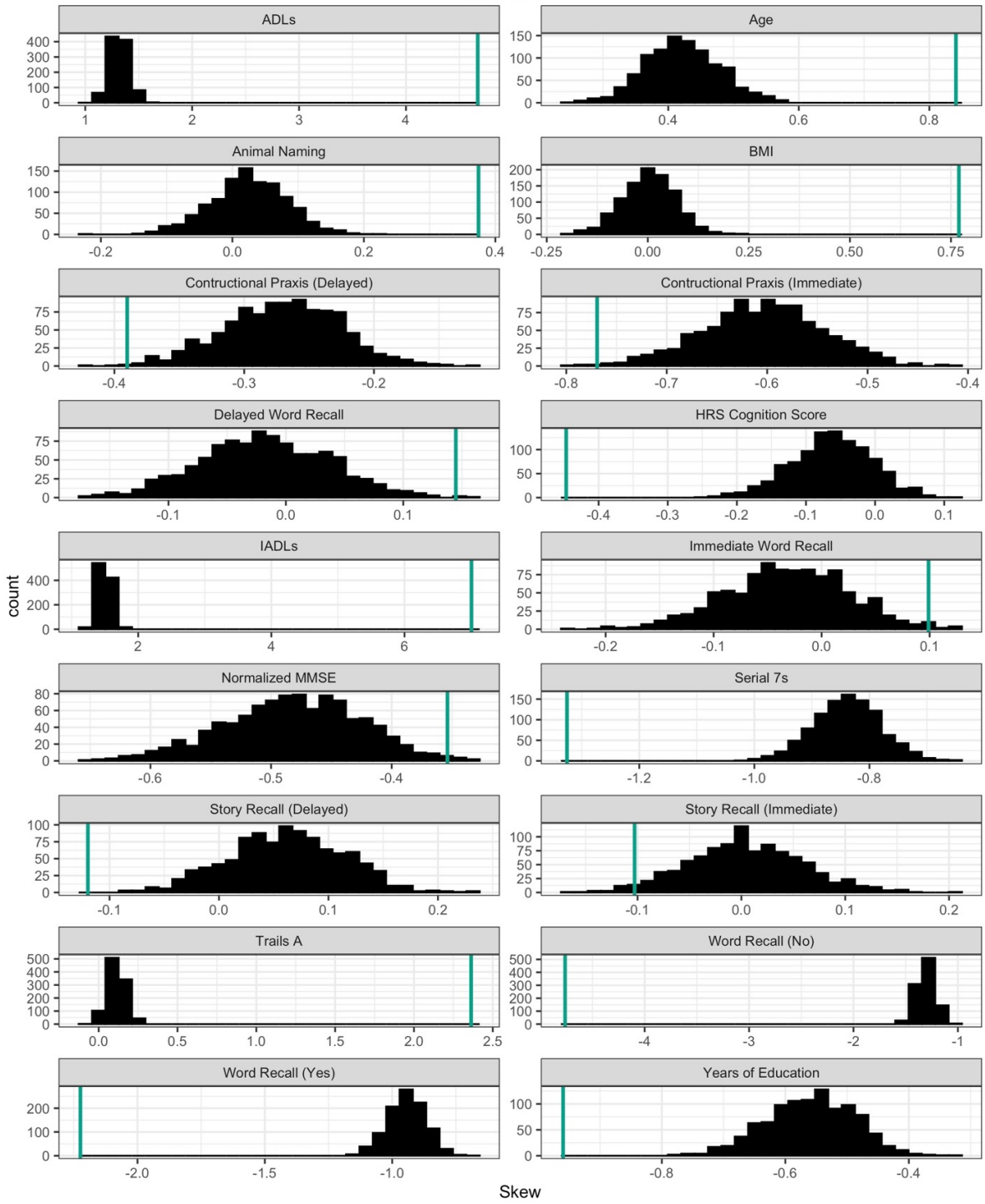
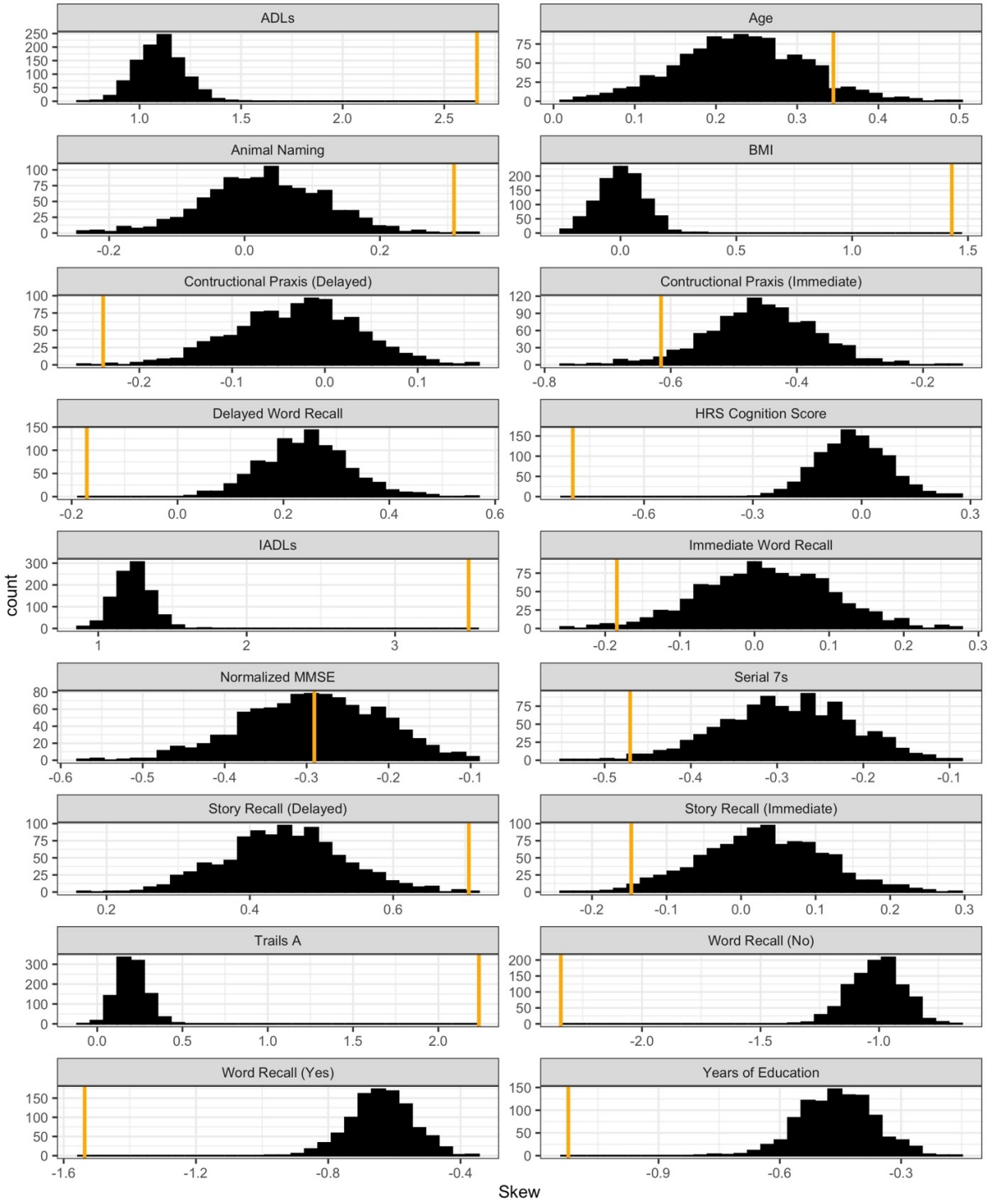


Figure D.1: Posterior predictive distributions of all continuous variable medians for the simulation scenario with $n_{HRS} = 8000$ and 50% HCAP sampling proportion presented in Chapter 4. Distributions are based on 1000 synthetic HCAP datasets by cognitive impairment group: (a) Unimpaired, (b) MCI, (c) Dementia, (d) Other. Colored vertical lines in each panel denote true medians.

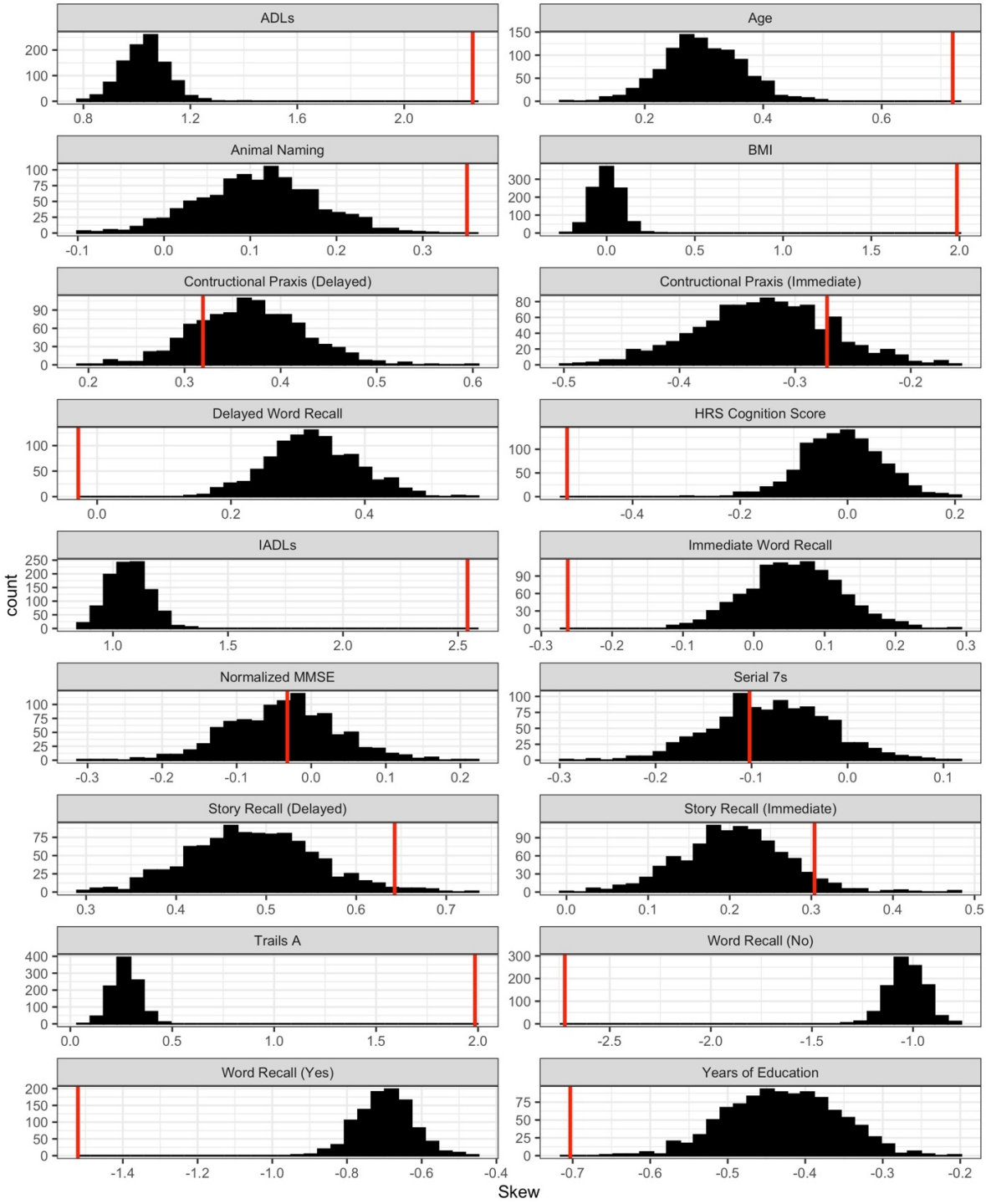
(a) Unimpaired



(b) MCI


ADLs
Age
Animal Naming
BMI
Contructional Praxis (Delayed)
Contructional Praxis (Immediate)
Delayed Word Recall
HRS Cognition Score
IADLs
Immediate Word Recall
Normalized MMSE
Serial 7s
Story Recall (Delayed)
Story Recall (Immediate)
Trails A
Word Recall (No)
Word Recall (Yes)
Years of Education
count
Skew

(c) Dementia


ADLs
Age
Animal Naming
BMI
Contructional Praxis (Delayed)
Contructional Praxis (Immediate)
Delayed Word Recall
HRS Cognition Score
IADLs
Immediate Word Recall
Normalized MMSE
Serial 7s
Story Recall (Delayed)
Story Recall (Immediate)
Trails A
Word Recall (No)
Word Recall (Yes)
Years of Education
count
Skew

(d) Other

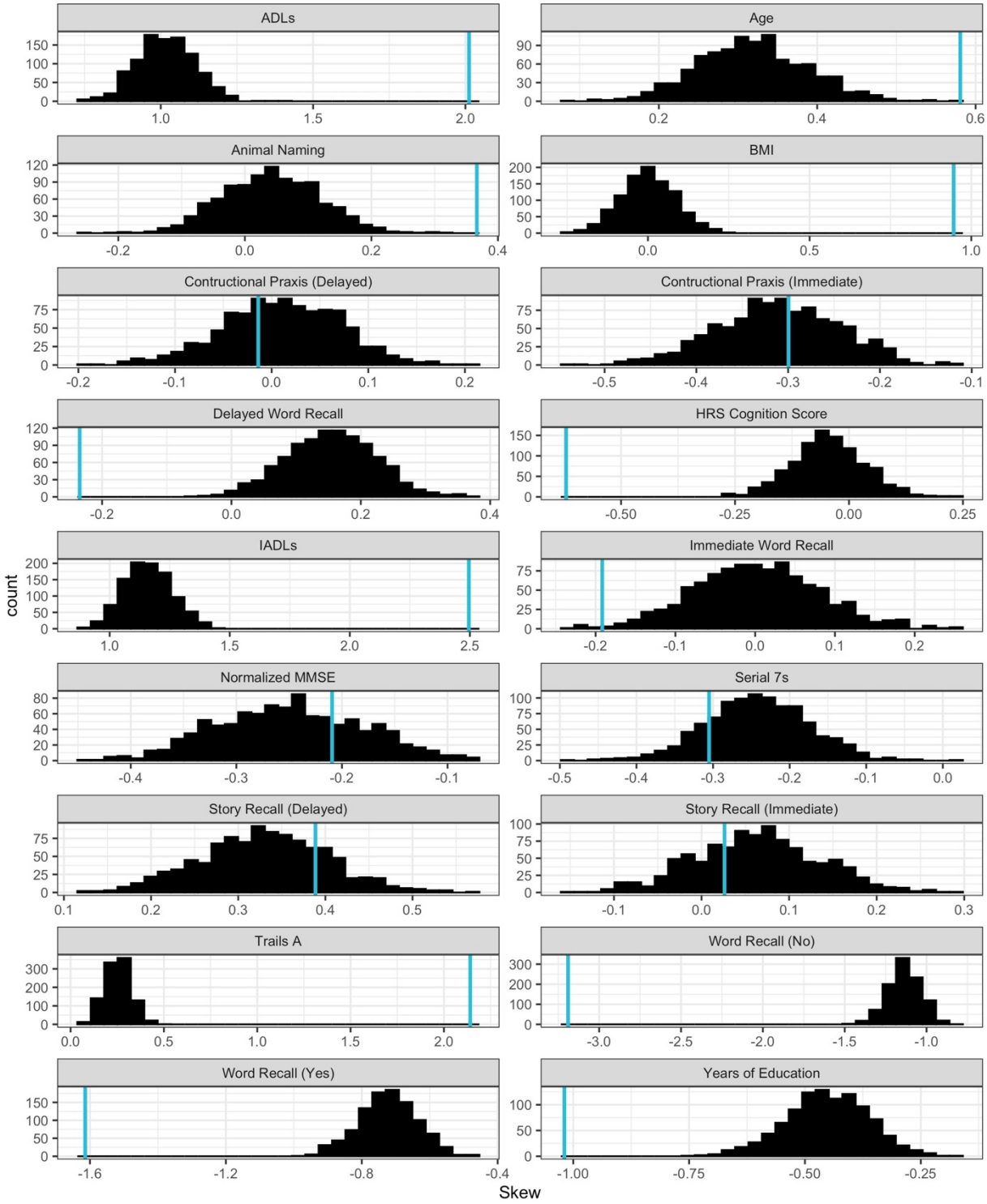
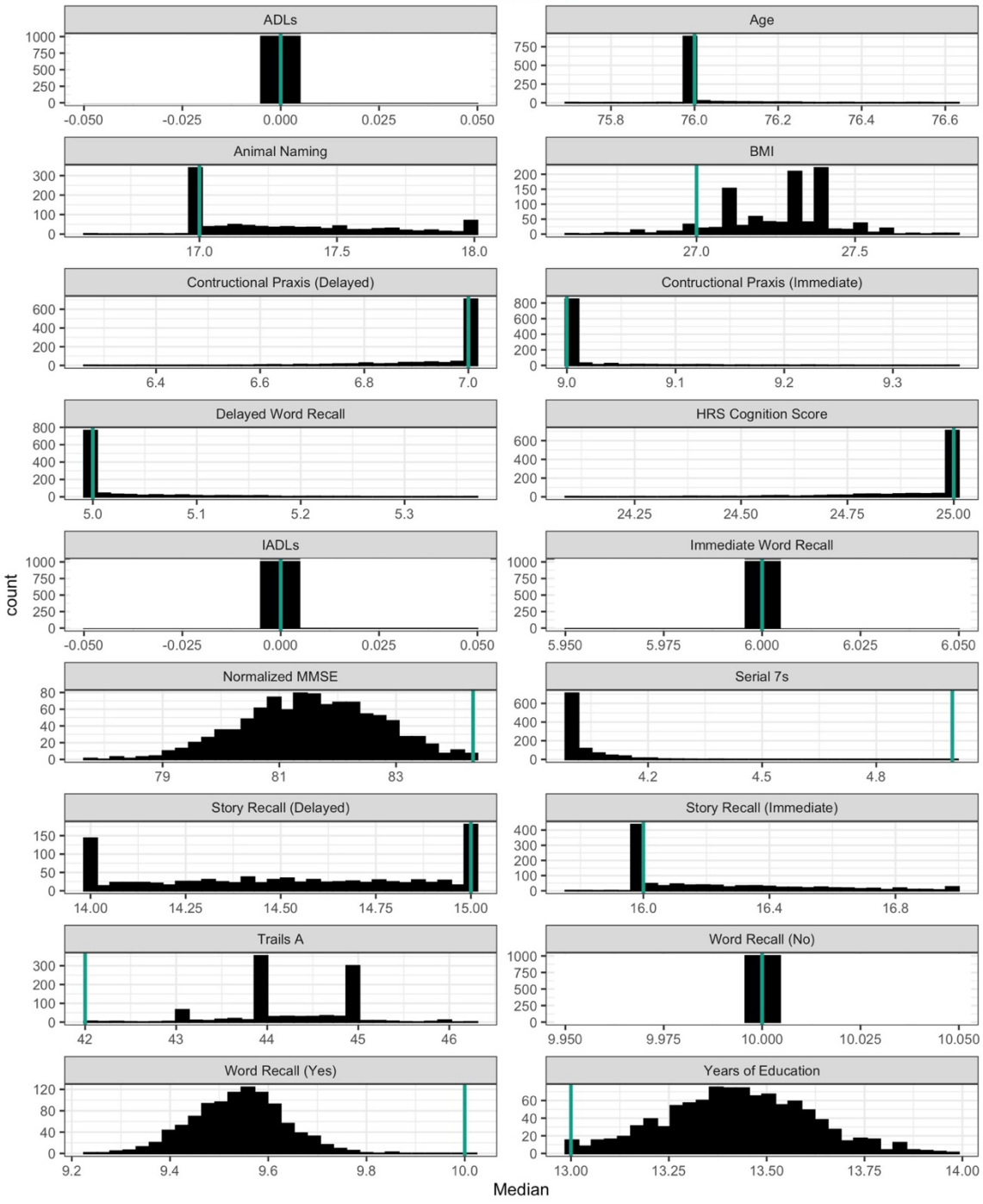


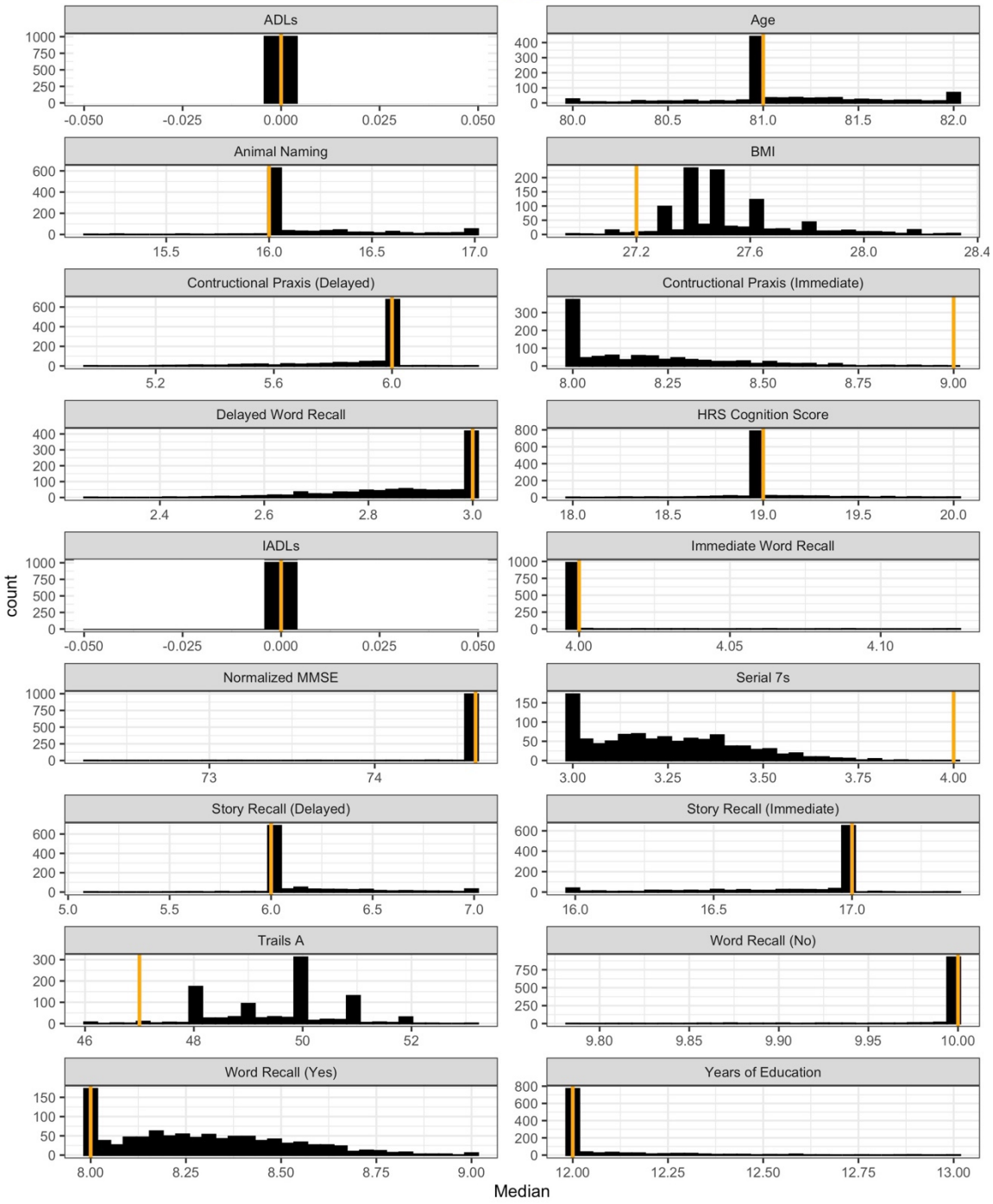
Figure D.2: Posterior predictive distributions of skew for all continuous variables for the simulation scenario with $n_{HRS} = 8000$ and 50% HCAP sampling proportion presented in Chapter 4. Distributions are based on 1000 synthetic HCAP datasets by cognitive impairment group: (a) Unimpaired, (b) MCI, (c) Dementia, (d) Other. Colored vertical lines in each panel denote true skew.

D.2: Chapter 5 supplementary figures

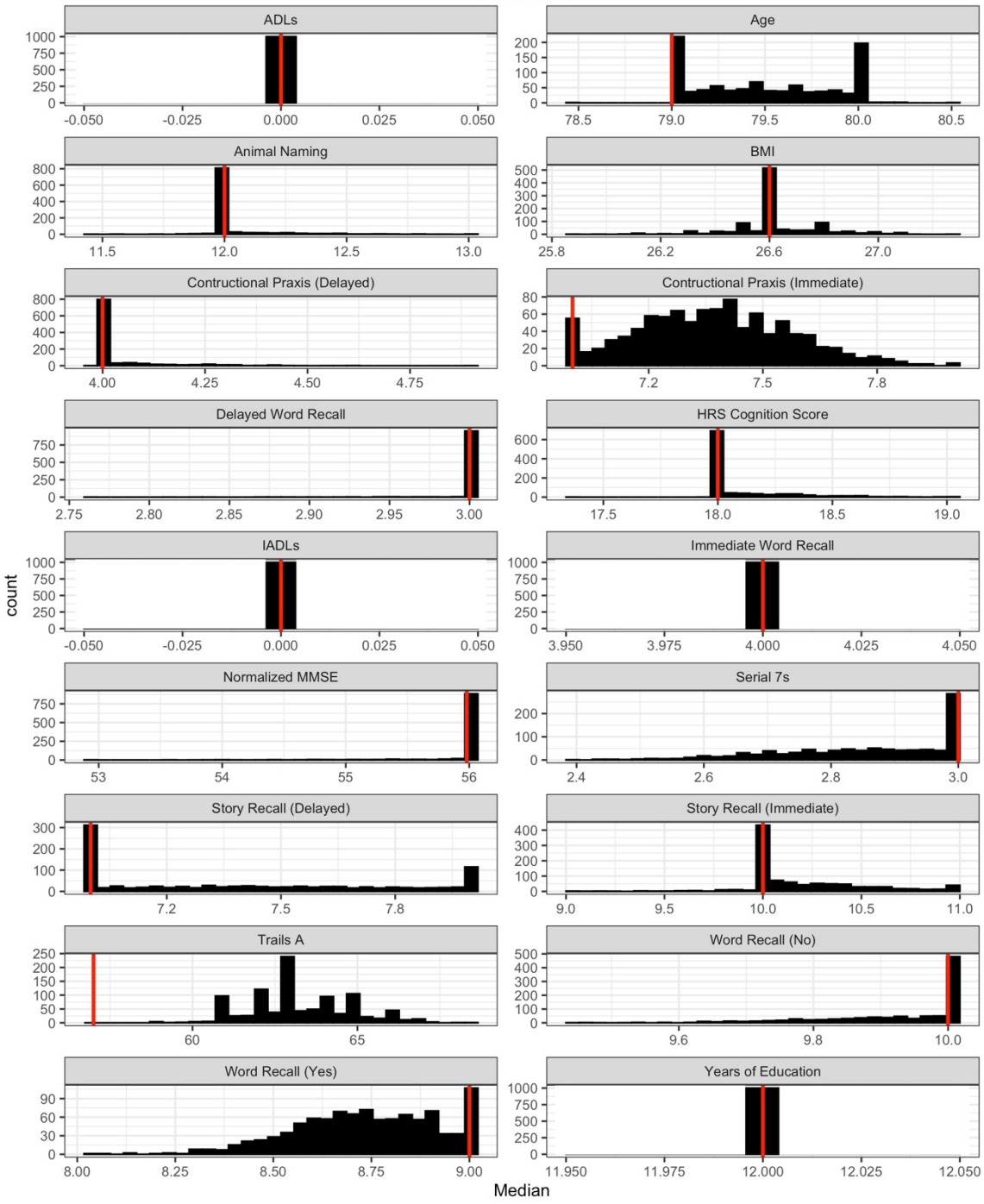
(a) Unimpaired



(b) MCI


ADLs
Age
Animal Naming
BMI
Contructional Praxis (Delayed)
Contructional Praxis (Immediate)
Delayed Word Recall
HRS Cognition Score
IADLs
Immediate Word Recall
Normalized MMSE
Serial 7s
Story Recall (Delayed)
Story Recall (Immediate)
Trails A
Word Recall (No)
Word Recall (Yes)
Years of Education
count
Median

(c) Dementia


ADLs
Age
Animal Naming
BMI
Contructional Praxis (Delayed)
Contructional Praxis (Immediate)
Delayed Word Recall
HRS Cognition Score
IADLs
Immediate Word Recall
Normalized MMSE
Serial 7s
Story Recall (Delayed)
Story Recall (Immediate)
Trails A
Word Recall (No)
Word Recall (Yes)
Years of Education
count
Median

(d) Other

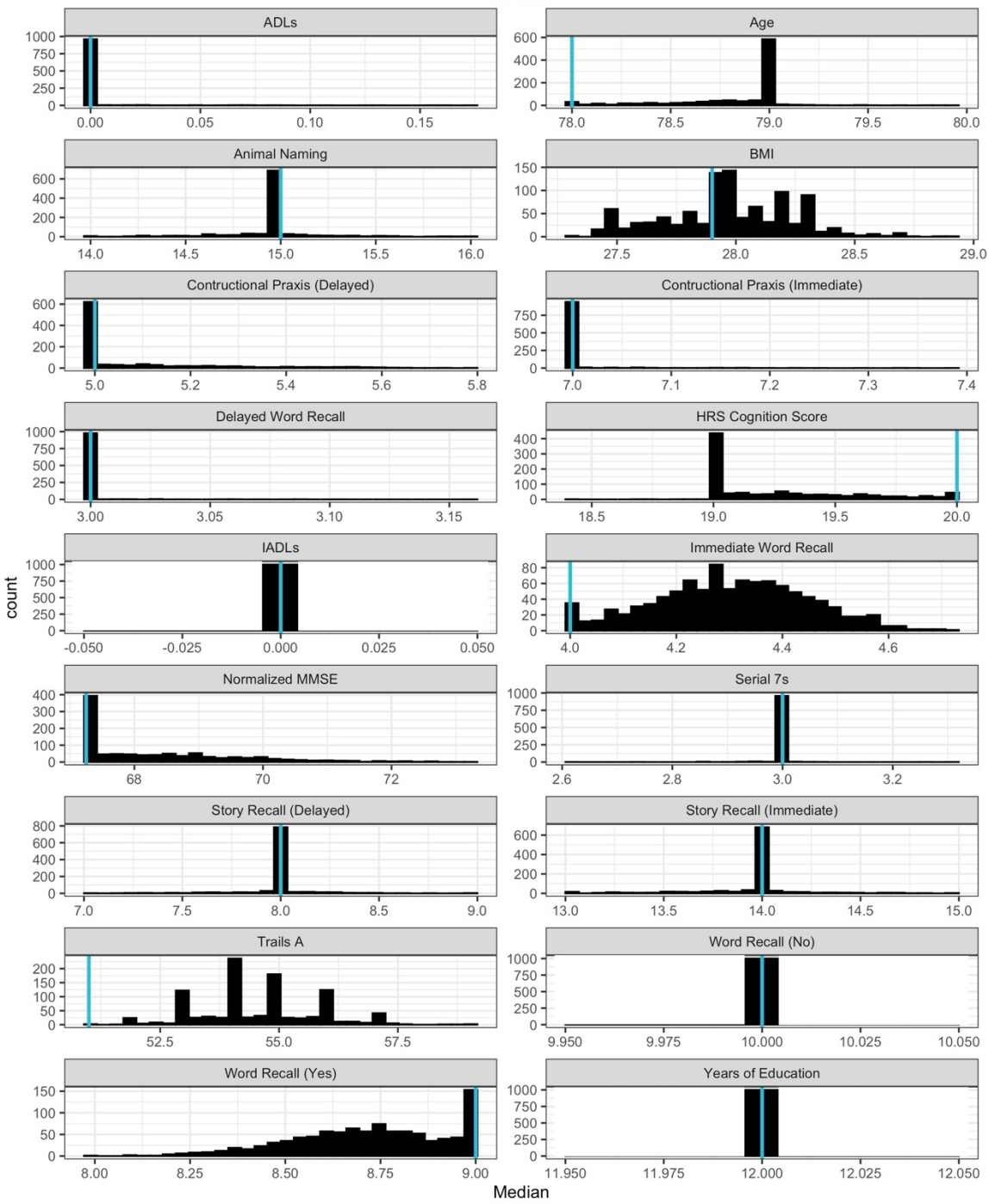
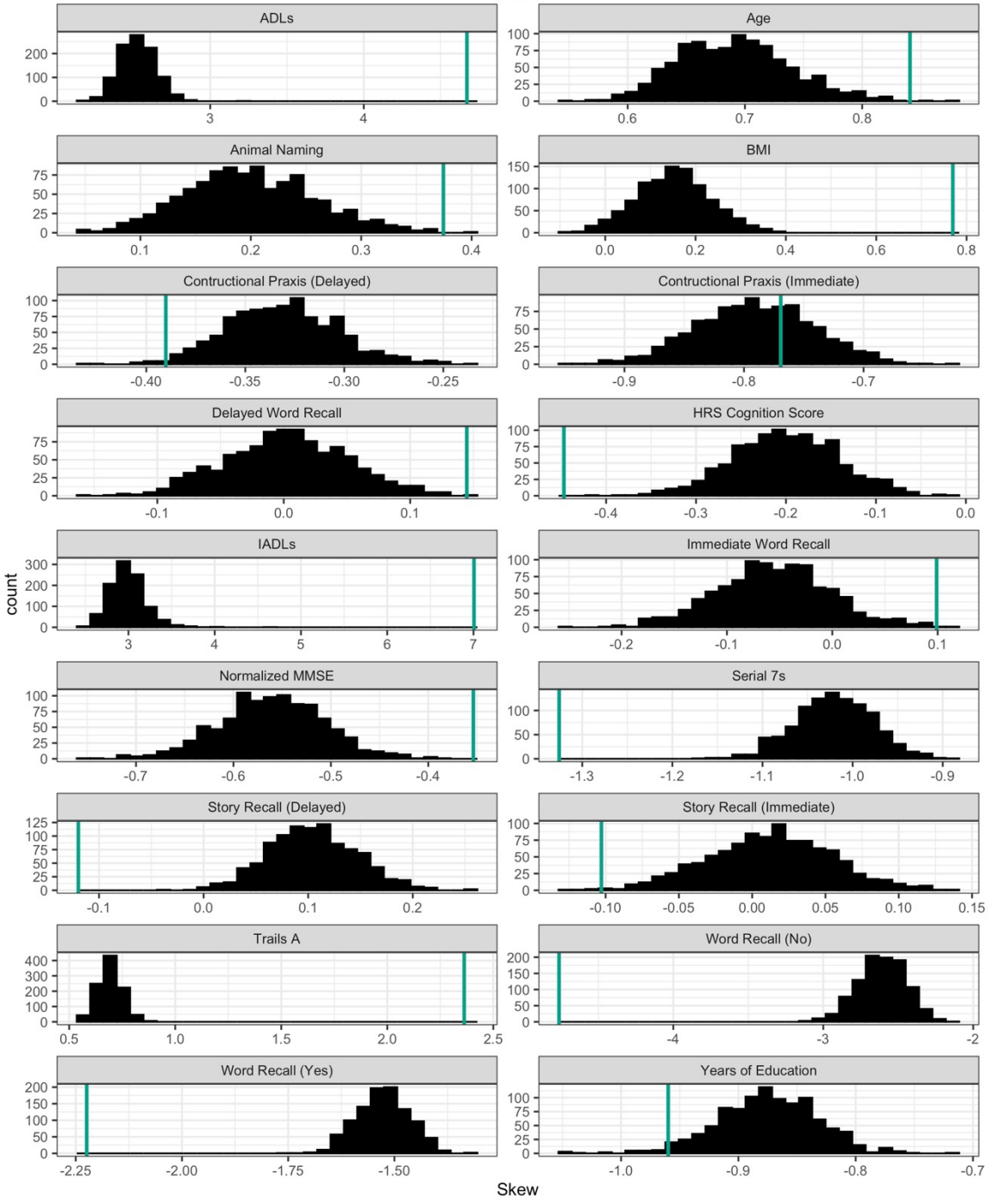
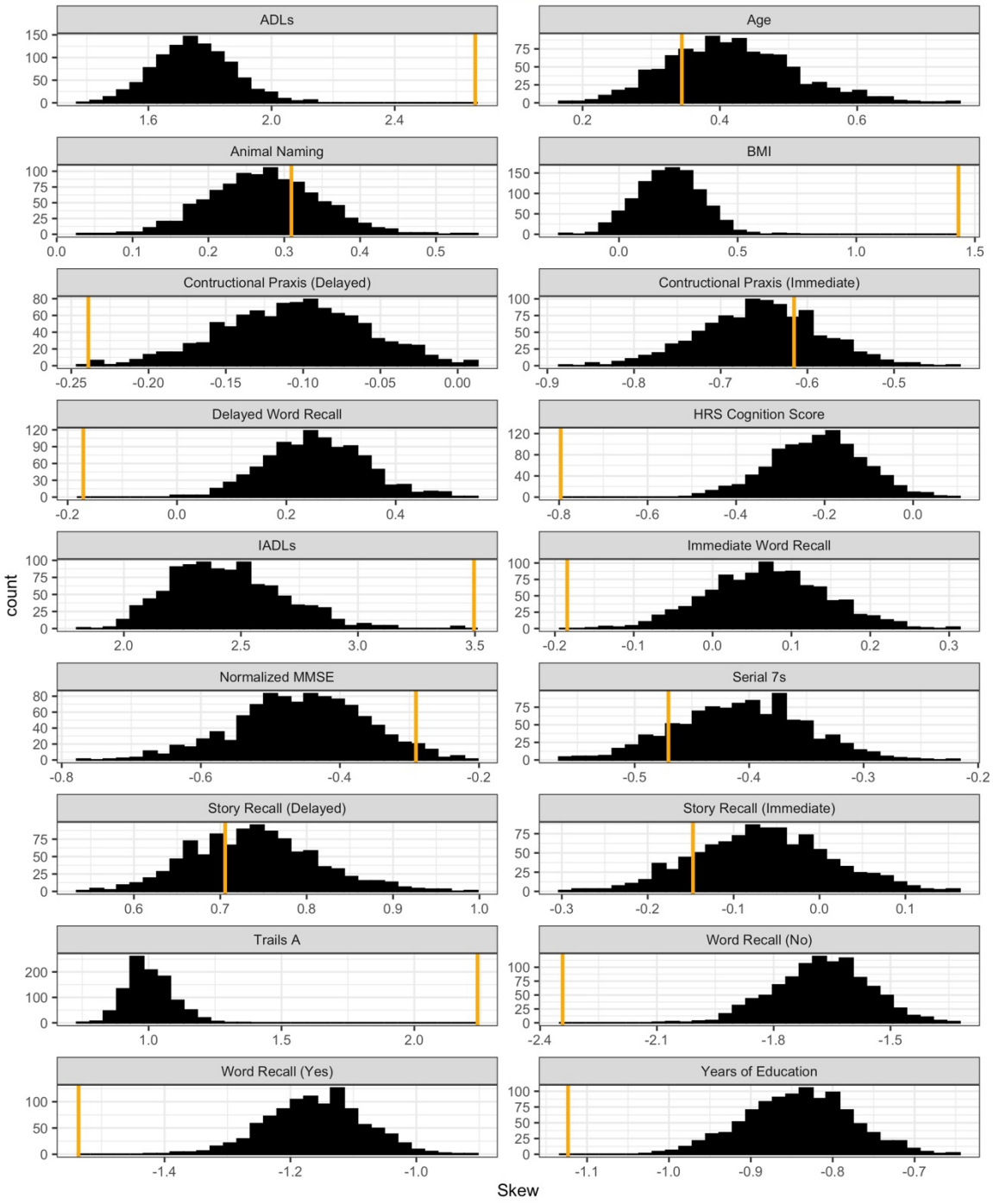


Figure D.3: Posterior predictive distributions of all continuous variable medians for the simulation scenario with $n_{HRS} = 8000$, 50% HCAP sampling proportion, and 50% race-stratified SRS calibration presented in Chapter 5. Distributions are based on 1000 synthetic HCAP datasets by cognitive impairment group: (a) Unimpaired, (b) MCI, (c) Dementia, (d) Other. Colored vertical lines in each panel denote true medians.

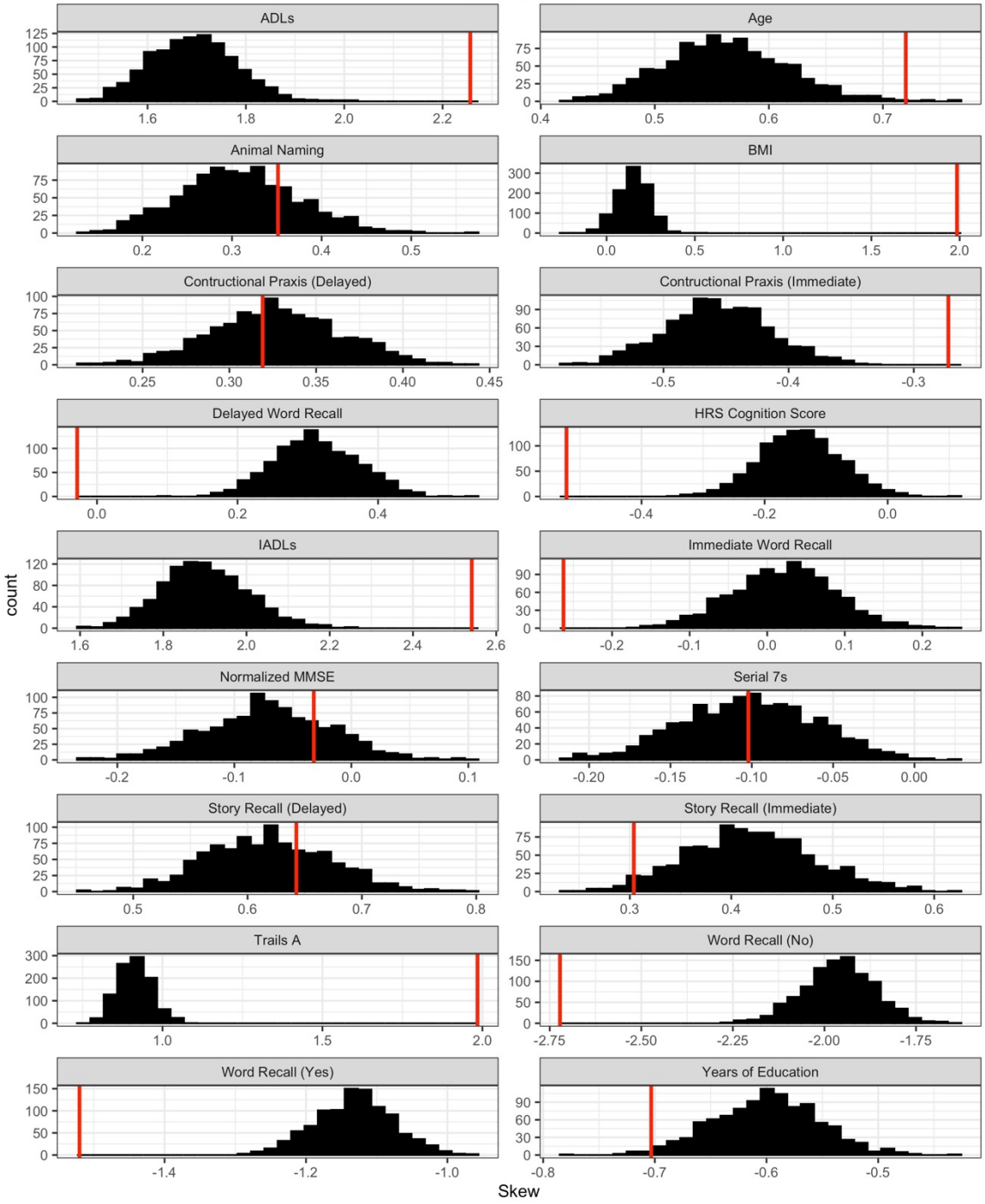
(a) Unimpaired



(b) MCI


ADLs
Age
Animal Naming
BMI
Contructional Praxis (Delayed)
Contructional Praxis (Immediate)
Delayed Word Recall
HRS Cognition Score
IADLs
Immediate Word Recall
Normalized MMSE
Serial 7s
Story Recall (Delayed)
Story Recall (Immediate)
Trails A
Word Recall (No)
Word Recall (Yes)
Years of Education
count
Skew

(c) Dementia


ADLs
Age
Animal Naming
BMI
Contructional Praxis (Delayed)
Contructional Praxis (Immediate)
Delayed Word Recall
HRS Cognition Score
IADLs
Immediate Word Recall
Normalized MMSE
Serial 7s
Story Recall (Delayed)
Story Recall (Immediate)
Trails A
Word Recall (No)
Word Recall (Yes)
Years of Education
count
Skew

(d) Other

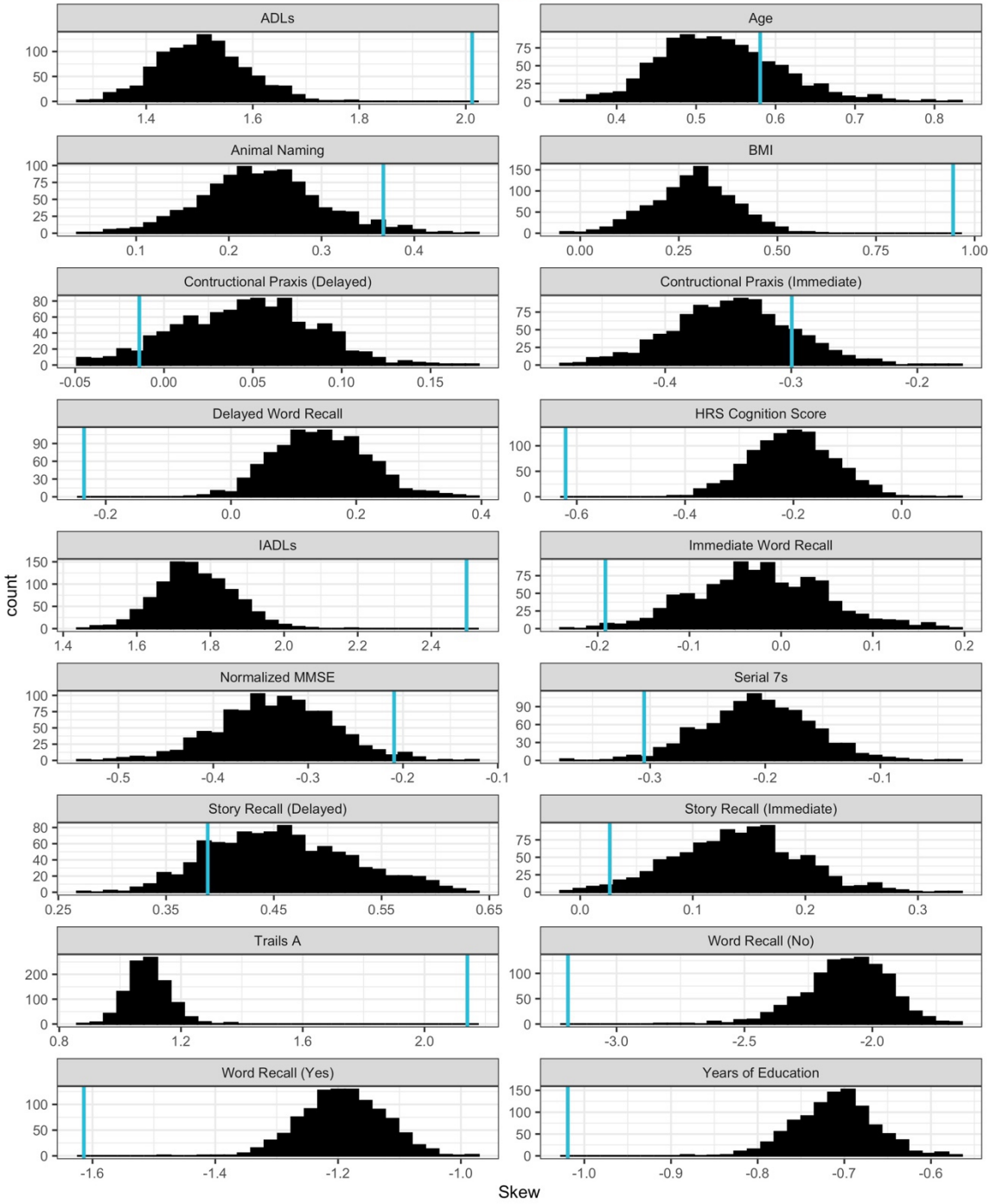


Figure D.4: Posterior predictive distributions of skew for all continuous variables for the simulation scenario with $n_{HRS} = 8000$, 50% HCAP sampling proportion, and 50% race-stratified SRS calibration presented in Chapter 5. Distributions are based on 1000 synthetic HCAP datasets by cognitive impairment group: (a) Unimpaired, (b) MCI, (c) Dementia, (d) Other. Colored vertical lines in each panel denote true skew.

D.3: Chapter 6 supplementary figures

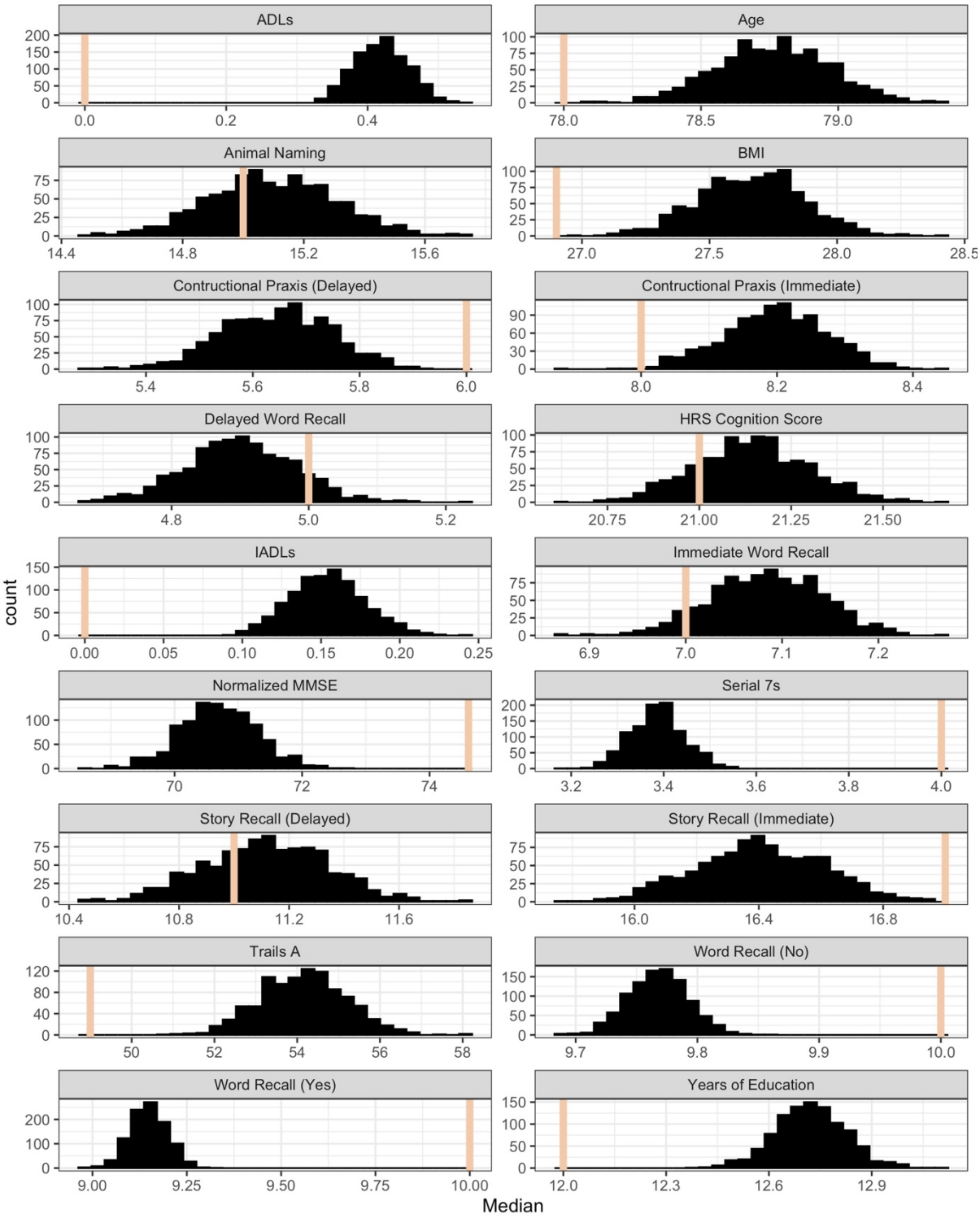


Figure D.5: Posterior predictive distributions of medians for all continuous variables for HCAP 2016 analysis presented in Chapter 6. Distributions are based on 1000 synthetic HCAP 2016 datasets. Colored vertical lines in each panel denote observed medians in the HCAP 2016 data.

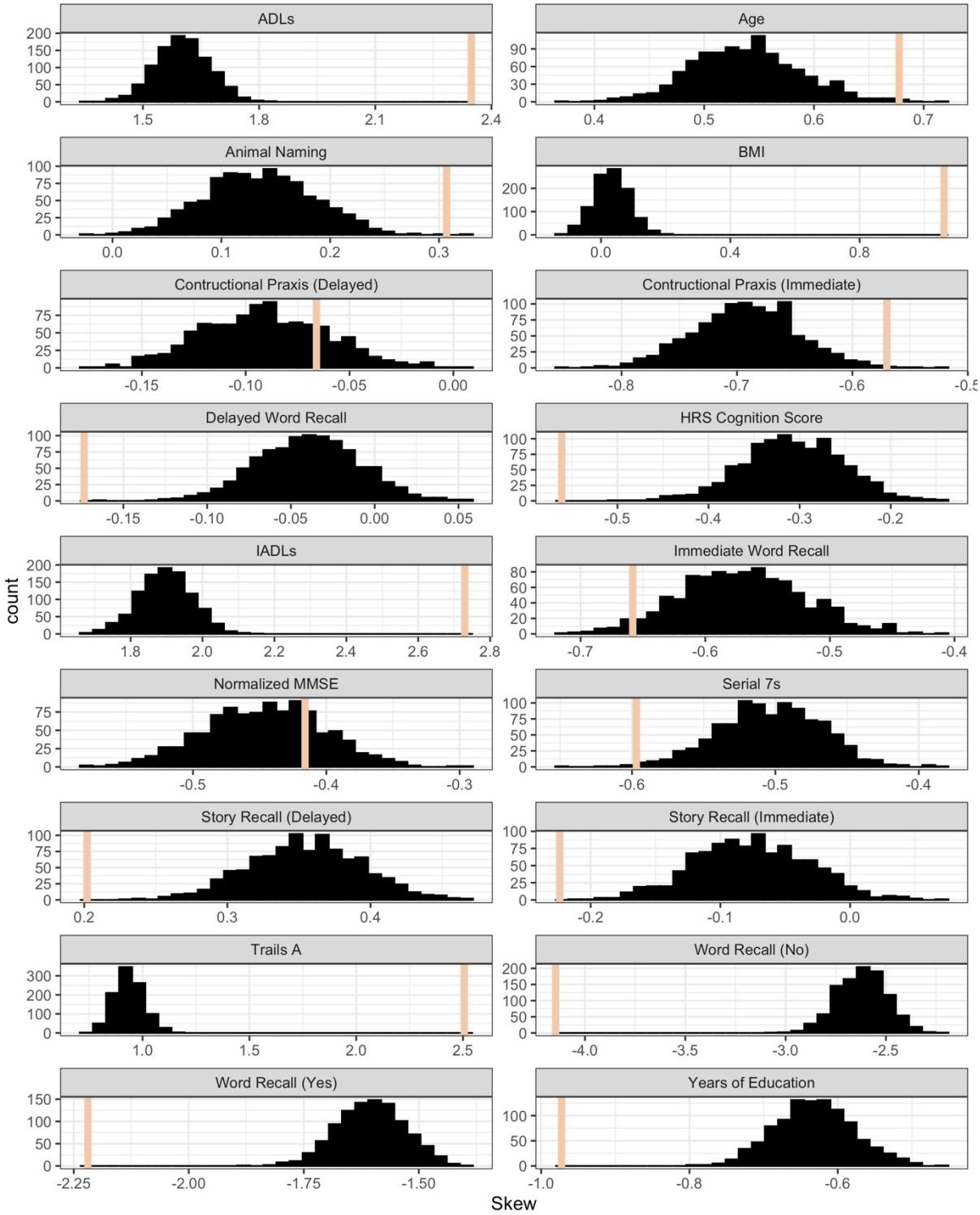


Figure D.6: Posterior predictive distributions of skewness for all continuous variables for HCAP 2016 analysis presented in Chapter 6. Distributions are based on 1000 synthetic HCAP 2016 datasets. Colored vertical lines in each panel denote observed skewness in the HCAP 2016 data.

Appendix E Supplementary Tables

E.1: Chapter 4 supplementary tables

Variable	ADAMS Baseline (2002) N = 826	HCAP 70+ Baseline (2016) N = 2,298	HRS 70+ (2016) N = 6,313	Super-population N = 1,000,000
Age, Mean (SD)	81.3 (7.0)	78.8 (5.9)	78.8 (6.0)	78.8 (6.0)
Female, n (%)	476 (57.6%)	1,379 (60.0%)	3,765 (59.6%)	596,718 (59.7%)
Race/Ethnicity, n (%)				
White	590 (71.4%)	1,767 (76.9%)	4,837 (76.6%)	765,663 (76.6%)
Black	153 (18.5%)	325 (14.1%)	895 (14.2%)	142,182 (14.2%)
Hispanic	83 (10.0%)	206 (9.0%)	581 (9.2%)	92,155 (9.2%)
Years of Education, Mean (SD)	10.0 (4.4)	12.6 (3.1)	12.6 (3.2)	12.6 (3.2)
Employment status, n (%)				
Working	58 (7.0%)	190 (8.3%)	548 (8.7%)	86,592 (8.7%)
Not working	61 (7.4%)	305 (13.3%)	809 (12.8%)	128,287 (12.8%)
Retired	707 (85.6%)	1,803 (78.5%)	4,956 (78.5%)	785,121 (78.5%)
Married/Partnered, n (%)	327 (39.6%)	1,236 (53.8%)	3,356 (53.2%)	531,414 (53.1%)
BMI, Mean (SD)	25.9 (5.4)	27.8 (5.6)	27.7 (5.6)	27.7 (5.6)
History of stroke, n (%)	151 (18.3%)	307 (13.4%)	796 (12.6%)	125,963 (12.6%)
History of diabetes, n (%)	158 (19.1%)	681 (29.6%)	1,858 (29.4%)	294,720 (29.5%)
History of heart disease, n (%)	267 (32.3%)	849 (36.9%)	2,306 (36.5%)	365,608 (36.6%)
History of hypertension, n (%)	447 (54.1%)	1,676 (72.9%)	4,579 (72.5%)	726,009 (72.6%)
Current smoker, n (%)	60 (7.3%)	143 (6.2%)	407 (6.4%)	64,479 (6.4%)
Alcohol consumption, n (%)				
No drinking	681 (82.4%)	1,543 (67.1%)	4,247 (67.3%)	673,197 (67.3%)
Moderate drinking	108 (13.1%)	567 (24.7%)	1,579 (25.0%)	249,798 (25.0%)
Heavy drinking	37 (4.5%)	188 (8.2%)	487 (7.7%)	77,005 (7.7%)
ADLs, Mean (SD)	0.9 (1.4)	0.4 (1.0)	0.4 (1.0)	0.4 (1.0)
IADLs, Mean (SD)	0.6 (1.0)	0.2 (0.5)	0.2 (0.5)	0.2 (0.5)
Immediate word recall, Mean (SD)	5.5 (2.4)	7.0 (1.8)	4.8 (1.7)	4.8 (1.7)
Delayed word recall, Mean (SD)	3.6 (2.7)	4.9 (2.6)	3.8 (2.0)	3.8 (2.0)
Serial 7s, Mean (SD)	2.2 (2.0)	3.3 (1.7)	3.3 (1.7)	3.3 (1.7)
Item naming (cactus): correct, n (%)	629 (76.2%)	2,119 (92.2%)	5,781 (91.6%)	915,717 (91.6%)
Item naming (scissor): correct, n (%)	794 (96.1%)	2,272 (98.9%)	6,213 (98.4%)	984,307 (98.4%)
President naming: correct, n (%)	674 (81.6%)	2,180 (94.9%)	6,047 (95.8%)	957,739 (95.8%)

Backwards count (20): correct, n (%)	601 (72.8%)	2,099 (91.3%)	5,779 (91.5%)	915,086 (91.5%)
HRS total cognition, Mean (SD)	15.1 (6.4)	20.8 (5.3)	20.8 (5.3)	20.8 (5.3)
Subjective cognitive status, n (%)				
Same as 2 years ago	518 (62.7%)	1,584 (68.9%)	4,423 (70.1%)	701,102 (70.1%)
Better than 2 years ago	59 (7.1%)	46 (2.0%)	129 (2.0%)	20,412 (2.0%)
Worse than 2 years ago	249 (30.1%)	668 (29.1%)	1,761 (27.9%)	278,486 (27.8%)
Total MMSE (normalized), Mean (SD)	51.0 (25.1)	70.5 (19.1)		70.4 (19.2)
Animal naming, Mean (SD)	11.5 (5.6)	15.5 (6.3)		15.5 (6.2)
Word recall (yes), Mean (SD)	8.1 (2.3)	8.9 (1.8)		8.6 (1.8)
Word recall (no), Mean (SD)	8.9 (2.1)	9.6 (1.2)		9.5 (1.2)
Immediate story recall, Mean (SD)	13.3 (9.0)	16.4 (6.4)		14.2 (6.3)
Delayed story recall, Mean (SD)	9.3 (8.5)	11.6 (7.4)		10.0 (6.7)
Immediate constructional praxis, Mean (SD)	8.3 (2.1)	8.1 (2.3)		8.1 (2.3)
Delayed constructional praxis, Mean (SD)	2.8 (1.1)	5.6 (3.2)		5.5 (3.2)
Trails A, Mean (SD)	86.0 (62.4)	58.1 (34.5)		57.6 (33.3)
Impairment group, n (%)				
Unimpaired	307 (37.2%)			373,112 (37.3%)
MCI	98 (11.9%)			163,587 (16.4%)
Dementia	273 (33.1%)			259,032 (25.9%)
Other	148 (17.9%)			204,269 (20.4%)

Table E.1: Sample characteristics for waves of multiply-imputed ADAMS, hotdeck imputed HCAP, and HRS relevant to the simulation study and the superpopulation generated for the simulation study. Cells that are grayed out indicated measures that are not available for the dataset.

Variable	Unimpaired γ_1	Other γ_2	MCI γ_3	Dementia γ_4
Age	X	X	X	X
Black	X	X	X	X
Hispanic	X	X	X	X
Female	X	X	X	X
Years of Education	X	X	X	X
Not working	X	X	X	X
Retired	X	X	X	X
Married/partnered		X	X	X
Normalized MMSE	X	X	X	X
Immediate word recall	X	X		X
Delayed word recall	X		X	X
Serial 7s	X	X	X	X
Animal naming	X		X	X
Word recall (yes)	X	X	X	X
Word recall (no)	X	X	X	X
Immediate story recall	X	X	X	X
Delayed story recall	X	X	X	X
Backwards count (20)	X		X	X
Immediate constructional praxis	X	X	X	X
Delayed constructional praxis	X	X	X	X
Trails A	X	X	X	X
HRS total cognition	X	X	X	X
Item naming (cactus)	X	X	X	X
Item naming (scissor)	X	X	X	X
President naming	X	X	X	X
Subjective cognition: better	X	X	X	X
Subjective cognition: worse	X	X	X	X
ADLs	X	X	X	X
IADLs	X	X	X	X
BMI	X		X	X
History of stroke	X	X	X	X
History of diabetes	X	X	X	X
History of heart disease	X	X	X	X
History of hypertension	X	X	X	X
Current smoker	X	X	X	X
Moderate drinking	X	X	X	X
Heavy drinking	X	X	X	X

Table E.2: Variables selected using LASSO regression for inclusion in each cognitive impairment class model described by Equations **(4.1) - (4.5)**. An X denotes a variable selected for inclusion. Grayed-out cells indicate variables that were not selected.

Variable	Dataset	Range	Bins
Age	HRS	70-107	[70, 85), 85+
Years of education	HRS	0-17	Less than HS, HS, some college +
Immediate word recall	HRS	0-10	[0, 6), [6, 8), [8, 10]
Delayed word recall	HRS	0-10	[0, 5), [5, 7), [7, 10]
Serial 7s	HRS	0-5	[0, 5), 5
HRS total cognition	HRS	0-35	Quintiles*
MMSE (normalized)	HCAP	0-100	Quartiles
Word list recall (yes)	HCAP	0-10	[0, 9), 9+
Word list recall (no)	HCAP	0-10	[0, 10), 10
Animal naming	HCAP	0-43	Quintiles
Story recall (immediate)	HCAP	0-35	Quartiles
Story recall (delayed)	HCAP	0-35	Quintiles
Constructional praxis (immediate)	HCAP	0-11	[0, 8), [8, 11), 11
Constructional praxis (delayed)	HCAP	0-11	[0, 5), [5, 7), [7, 9), 9+
Trails A	HCAP	0-300	Quintiles

*HRS total cognition was categorized by quintiles for HRS hotdeck imputation but was categorized by quartiles for generating values in the superpopulation.

Table E.3: Binning for continuous and ordered categorical variables for hotdeck imputation in HRS and HCAP. If a variable is available in HRS, it is also available in HCAP.

Variable to be imputed	Variables used for matching
Subjective cognition: better	Sex/gender; race/ethnicity; age; education; HRS total cognition
Subjective cognition: worse	Sex/gender; race/ethnicity; age; education; HRS total cognition
Immediate word recall	Sex/gender; race/ethnicity; age; education; HRS total cognition
Delayed word recall	Sex/gender; race/ethnicity; age; education; HRS total cognition; immediate word recall
Serial 7s	Sex/gender; race/ethnicity; age; education; HRS total cognition
Backwards count (20)	Sex/gender; race/ethnicity; age; education; HRS total cognition; serial 7s
Item naming: scissor	Sex/gender; race/ethnicity; age; education; HRS total cognition
Item naming: cactus	Sex/gender; race/ethnicity; age; education; HRS total cognition; item naming: scissor
President naming	Sex/gender; race/ethnicity; age; education; HRS total cognition
MMSE	Sex/gender; race/ethnicity; age; education; HRS total cognition
Word recall (yes)	MMSE; immediate word recall; delayed word recall
Word recall (no)	MMSE; immediate word recall; delayed word recall word recall (yes)
Immediate story recall	MMSE; immediate word recall; delayed word recall word recall (yes); word recall (no)
Delayed story recall	MMSE; word recall (yes); word recall (no)
Animal naming	MMSE
Immediate constructional praxis	MMSE
Delayed constructional praxis	MMSE; immediate constructional praxis
Trails A	MMSE; serial 7s; backwards count (20)

Table E.4: Variables used for matching in HCAP 2016 hotdeck imputation.

Variable to be imputed	Variables used for matching
MMSE	Race/ethnicity; education
Word recall (yes)	MMSE; delayed word recall
Word recall (no)	MMSE; delayed word recall; word recall (yes)
Immediate story recall	MMSE; immediate word recall
Delayed story recall	MMSE; delayed word recall; word recall (yes); word recall (no)
Animal naming	MMSE
Immediate constructional praxis	MMSE
Delayed constructional praxis	MMSE; immediate constructional praxis
Trails A	MMSE; serial 7s

Table E.5: Variables used for matching in superpopulation hotdeck imputation.

References

Albert, Marilyn S., Steven T. DeKosky, Dennis Dickson, Bruno Dubois, Howard H. Feldman, Nick C. Fox, Anthony Gamst, et al. 2011. "The Diagnosis of Mild Cognitive Impairment Due to Alzheimer's Disease: Recommendations from the National Institute on Aging-Alzheimer's Association Workgroups on Diagnostic Guidelines for Alzheimer's Disease." *Alzheimer's & Dementia : The Journal of the Alzheimer's Association* 7 (3): 279. *https://doi.org/10.1016/J.JALZ.2011.03.008*.

Alpérovitch, A., P. Amouyel, J. F. Dartigues, P. Ducimetière, B. Mazoyer, K. Ritchie, C. Tzourio, et al. 2003. "Vascular Factors and Risk of Dementia: Design of the Three-City Study and Baseline Characteristics of the Study Population." *Neuroepidemiology* 22 (6): 316–25. *https://doi.org/10.1159/000072920*.

Alzheimer's Association. 2021. "2021 Alzheimer's Disease Facts and Figures." *Alzheimers and Dementia* 17 (3): 1–108.

———. 2022. "2022 Alzheimer's Disease Facts and Figures." *Alzheimers Dementia* 18: 1–122.

Andridge, Rebecca R., and Roderick J.A. Little. 2010. "A Review of Hot Deck Imputation for Survey Non-Response." *International Statistical Review* 78 (1): 64. *https://doi.org/10.1111/J.1751-5823.2010.00103.X*.

Arevalo-Rodriguez, Ingrid, Nadja Smailagic, Marta Roquéi Figuls, Agustín Ciapponi, Erick Sanchez-Perez, Antri Giannakou, Olga L. Pedraza, Xavier Bonfill Cosp, and Sarah Cullum. 2015. "Mini-Mental State Examination (MMSE) for the Detection of Alzheimer's Disease and Other Dementias in People with Mild Cognitive Impairment (MCI)." *The Cochrane Database of Systematic Reviews* 2015 (3). *https://doi.org/10.1002/14651858.CD010783.PUB2*.

Babulal, Ganesh M., Yakeel T. Quiroz, Benedict C. Albensi, Eider Arenaza-Urquijo, Arlene J. Astell, Claudio Babiloni, Alex Bahar-Fuchs, et al. 2019. "Perspectives on Ethnic and Racial Disparities in Alzheimer's Disease and Related Dementias: Update and Areas of Immediate Need." *Alzheimer's and Dementia* 15 (2): 292–312. *https://doi.org/10.1016/j.jalz.2018.09.009*.

Bazeley, Pat. 2011. "Integrative Analysis Strategies for Mixed Data Sources:" *American Behavioral Scientist* 56 (6): 814–28. *https://doi.org/10.1177/0002764211426330*.

Belin, Thomas R., and Donald B. Rubin. 1995. "The Analysis of Repeated-measures Data on Schizophrenic Reaction Times Using Mixture Models." *Statistics in Medicine* 14 (8): 747–68. https://doi.org/10.1002/sim.4780140805.

Bennett, D. A., R. S. Wilson, J. A. Schneider, D. A. Evans, L. A. Beckett, N. T. Aggarwal, L. L. Barnes, J. H. Fox, and J. Bach. 2002. "Natural History of Mild Cognitive Impairment in Older Persons." *Neurology* 59 (2): 198–205. *https://doi.org/10.1212/WNL.59.2.198*.

Bennett, David A., Julie A. Schneider, Aron S. Buchman, Lisa L. Barnes, Patricia A. Boyle, and Robert S. Wilson. 2012. "Overview and Findings from the Rush Memory and Aging Project." *Current Alzheimer Research* 9 (6): 663. /pmc/articles/PMC3439198/.

Bienias, Julia L., Laurel A. Beckett, David A. Bennett, Robert S. Wilson, and Denis A. Evans. 2003. "Design of the Chicago Health and Aging Project (CHAP)." *Journal of Alzheimer's Disease* 5 (5): 349–55. *https://doi.org/10.3233/JAD-2003-5501*.

Bodner, Todd E. 2008. "What Improves with Increased Missing Data Imputations?" *Structural Equation Modeling* 15 (4): 651–75. *https://doi.org/10.1080/10705510802339072*.

Borenstein, Michael, Larry v Hedges, Julian P T Higgins, and Hannah R Rothstein. 2010. "A Basic Introduction to Fixed-Effect and Random-Effects Models for Meta-Analysis." *Research Synthesis Methods* 1 (August): 97–111. *https://doi.org/10.1002/jrsm.12*.

Boulanger, Bruno, and Bradley P. Carlin. 2021. "How and Why Bayesian Statistics Are Revolutionizing Pharmaceutical Decision Making." *Clinical Researcher*, 2021.

Brandt, J, M Spencer, and M Folstein. 1988. "The Telephone Interview for Cognitive Status ." *Neuropsychiatry, Neuropsychology and Behavioral Neurology* 1 (2): 111–17. *https://jhu.pure.elsevier.com/en/publications/the-telephone-interview-for-cognitive-status-3*.

Buckley, Rachel F, Elizabeth C Mormino, Rebecca E Amariglio, Michael J Properzi, Jennifer S Rabin, Yen Ying Lim, Kathryn v Papp, et al. 2018. "Sex, Amyloid, and APOEε4 and Risk of Cognitive Decline in Preclinical Alzheimer's Disease: Findings from Three Well-Characterized Cohorts HHS Public Access." *Alzheimers Dement* 14 (9): 1193–1203. *https://doi.org/10.1016/j.jalz.2018.04.010*.

Cao, Yi, Heather Allore, Brent vander Wyk, and Roee Gutman. 2022. "Review and Evaluation of Imputation Methods for Multivariate Longitudinal Data with Mixed-Type Incomplete Variables." *Statistics in Medicine*, 1–33. *https://doi.org/10.1002/SIM.9592*.

Casella, George. 1985. "An Introduction to Empirical Bayes Data Analysis." *The American Statistician* 39 (2): 83–87.

Chin, Alexander L, Selamawit Negash, and Roy Hamilton. 2011. "Diversity and Disparity in Dementia: The Impact of Ethnoracial Differences in Alzheimer's Disease." *Alzheimer Dis Assoc Disord* 25 (3): 187–95. *https://doi.org/10.1097/WAD.0b013e318211c6c9*.

Crimmins, Eileen M., Jung Ki Kim, Kenneth M. Langa, and David R. Weir. 2011. "Assessment of Cognition Using Surveys and Neuropsychological Assessment: The Health and Retirement Study and the Aging, Demographics, and Memory Study." *The Journals of*

Gerontology. Series B, Psychological Sciences and Social Sciences 66 Suppl 1 (Suppl 1): i162. *https://doi.org/10.1093/geronb/gbr048*.

Cunningham, Timothy J., Janet B. Croft, Yong Liu, Hua Lu, Paul I. Eke, and Wayne H. Giles. 2017. "Vital Signs: Racial Disparities in Age-Specific Mortality Among Blacks or African Americans — United States, 1999–2015." *MMWR. Morbidity and Mortality Weekly Report* 66 (17): 444–56. *https://doi.org/10.15585/mmwr.mm6617e1*.

Dahal, Prabin, Kasia Stepniewska, Philippe J. Guerin, Umberto D'Alessandro, Ric N. Price, and Julie A. Simpson. 2019. "Dealing with Indeterminate Outcomes in Antimalarial Drug Efficacy Trials: A Comparison between Complete Case Analysis, Multiple Imputation and Inverse Probability Weighting." *BMC Medical Research Methodology* 19 (1). *https://doi.org/10.1186/S12874-019-0856-Z*.

Demirovic, Jasenka, Ronald Prineas, David Loewenstein, Judy Bean, Ranjan Duara, Steven Sevush, and Jose Szapocznik. 2003. "Prevalence of Dementia in Three Ethnic Groups: The South Florida Program on Aging and Health." *Annals of Epidemiology* 13 (6): 472–78. *https://doi.org/10.1016/S1047-2797(02)00437-4*.

Drechsler, Jörg. 2015. "Multiple Imputation of Multilevel Missing Data—Rigor Versus Simplicity:" *Http://Dx.Doi.Org/10.3102/1076998614563393* 40 (1): 69–95. *https://doi.org/10.3102/1076998614563393*.

Du, Jiacong, Jonathan Boss, Peisong Han, Lauren J Beesley, Stephen A Goutman, Stuart Batterman, Eva L Feldman, and Bhramar Mukherjee. 2022. "Variable Selection with Multiply-Imputed Datasets: Choosing between Stacked and Grouped Methods." *Journal of Computational and Graphical Statistics* JA (JA): 1–35. *https://doi.org/10.1080/10618600.2022.2035739/FORMAT/EPUB*.

Efron, Bradley., and Robert J. Tibshirani. 1994. *An Introduction to the Bootstrap*. 1st ed. Chapman & Hall/CRC.

Elwood, RW. 1991. "The Wechsler Memory Scale-Revised: Psychometric Characteristics and Clinical Application." *Neuropsychology Review* 2 (2): 179–201. *https://doi.org/10.1007/BF01109053*.

Ferretti, Maria Teresa, Julie Martinkova, Ewelina Biskup, Thomas Benke, Gino Gialdini, Zuzana Nedelska, Katrin Rauen, et al. 2020. "Sex and Gender Differences in Alzheimer's Disease: Current Challenges and Implications for Clinical Practice." *European Journal of Neurology*, February. *https://doi.org/10.1111/ene.14174*.

Fillenbaum, Gerda G., Albert Heyman, Marc S. Huber, Max A. Woodbury, Jack Leiss, Kenneth E. Schmader, Arline Bohannon, and Barbara Trapp-Moen. 1998. "The Prevalence and 3-Year Incidence of Dementia in Older Black and White Community Residents." *Journal of Clinical Epidemiology* 51 (7): 587–95. https://doi.org/10.1016/S0895-4356(98)00024-9.

Fishman, Ezra. 2017. “Risk of Developing Dementia at Older Ages in the United States.” *Demography* 54: 1897–1919. *https://doi.org/10.1007/s13524-017-0598-7*.

Folstein, Marshal F., Susan E. Folstein, and Paul R. McHugh. 1975. “Mini-Mental State: A Practical Method for Grading the Cognitive State of Patients for the Clinician.” *Journal of Psychiatric Research* 12: 189–98.

Freedman, Vicki A, and Judith D Kasper. 2019. “Cohort Profile: The National Health and Aging Trends Study (NHATS).” *International Journal of Epidemiology* 48 (4): 1044–45. *https://doi.org/10.1093/ije/dyz109*.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. 2010. “Regularization Paths for Generalized Linear Models via Coordinate Descent.” *Journal of Statistical Software* 33 (1): 1–22.

Gabry, Jonah, Daniel Simpson, Aki Vehtari, Michael Betancourt, and Andrew Gelman. 2019. “Visualization in Bayesian Workflow.” *Journal of the Royal Statistical Society. Series A: Statistics in Society* 182 (2): 389–402. *https://doi.org/10.1111/rssa.12378*.

Gasquoine, Philip G. 2009. “Race-Norming of Neuropsychological Tests.” *Neuropsychology Review* 19: 250–62. *https://doi.org/10.1007/s11065-009-9090-5*.

Gelman, Andrew, John B. Carlin, Hal Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. 2014. *Bayesian Data Analysis*. 3rd ed. CRC Press.

Gelman, Andrew, Aki Vehtari, Daniel Simpson, Charles C Margossian, Bob Carpenter, Yuling Yao, Lauren Kennedy, Jonah Gabry, Paul-Christian Bürkner, and Martin Modrák. 2020. “Bayesian Workflow.” *https://doi.org/10.48550/arXiv.2011.01808*

Gianattasio, Kan Z., Adam Ciarleglio, and Melinda C. Power. 2020. “Development of Algorithmic Dementia Ascertainment for Racial/Ethnic Disparities Research in the US Health and Retirement Study.” *Epidemiology* 31 (1): 126–33. *https://doi.org/10.1097/EDE.0000000000001101*.

Gianattasio, Kan Z., Qiong Wu, M. Maria Glymour, and Melinda C. Power. 2019. “Comparison of Methods for Algorithmic Classification of Dementia Status in the Health and Retirement Study.” *Epidemiology* 30 (2): 291–302. *https://doi.org/10.1097/EDE.0000000000000945*.

Gómez-Olivé, F Xavier, Livia Montana, Ryan G Wagner, Chodziwadziwa W Kabudula, Julia K Rohr, Kathleen Kahn, Till Bärnighausen, et al. 2018. “Cohort Profile: Health and Ageing in Africa: A Longitudinal Study of an INDEPTH Community in South Africa (HAALSI).” *International Journal of Epidemiology* 47 (3): 689–90. *https://doi.org/10.1093/IJE/DYX247*.

Gross, Alden L., Jason J. Hassenstab, Sterling C. Johnson, Lindsay R. Clark, Susan M. Resnick, Melissa Kitner-Triolo, Colin L. Masters, et al. 2017. “A Classification Algorithm for

Predicting Progression from Normal Cognition to Mild Cognitive Impairment across Five Cohorts: The Preclinical AD Consortium." *Alzheimer's and Dementia: Diagnosis, Assessment and Disease Monitoring* 8 (January): 147–55. *https://doi.org/10.1016/j.dadm.2017.05.003*.

Grund, Simon, Oliver Lüdtke, and Alexander Robitzsch. 2017. "Multiple Imputation of Missing Data for Multilevel Models: Simulations and Recommendations." *Https://Doi.Org/10.1177/1094428117703686* 21 (1): 111–49. *https://doi.org/10.1177/1094428117703686*.

Harvey, Philip D. 2019. "Domains of Cognition and Their Assessment." *Https://Doi.Org/10.31887/DCNS.2019.21.3/Pharvey* 21 (3): 227–37. *https://doi.org/10.31887/DCNS.2019.21.3/PHARVEY*.

Hayes-Larson, Eleanor, Crystal Shaw, Sarah F. Ackley, Scott C. Zimmerman, M. Maria Glymour, Rebecca E. Graff, John S. Witte, Lindsay C. Kobayashi, and Elizabeth Rose Mayeda. 2022. "The Role of Dementia Diagnostic Delay in the Inverse Cancer-Dementia Association." *The Journals of Gerontology. Series A, Biological Sciences and Medical Sciences* 77 (6): 1254–60. *https://doi.org/10.1093/GERONA/GLAB341*.

Hebert LE, Scherr PA, Bienias JL, Bennett DA, Evans da. 2003. "Alzheimer Disease in the US Population: Prevalence Estimates Using the 2000 Census." *Archives of Neurology* 60 (8): 1119–22. *https://doi.org/doi:10.1001/archneur.60.8.1119*.

Hedges, Larry v. 1992. "Meta-Analysis." *Journal of Educational and Behavioral Statistics* 17 (4): 279–96. *https://doi.org/10.3102/10769986017004279*.

Heeringa, Steven G, Gwenith G Fisher, Michael Hurd, Kenneth M Langa, Mary Beth Ofstedal, Brenda L Plassman, Willard L Rodgers, and David R Weir. 2009. "Aging, Demographics and Memory Study (ADAMS) Sample Design, Weighting and Analysis for ADAMS."

Herzog, AR, and RB Wallace. 1997. "Measures of Cognitive Functioning in the AHEAD Study." *The Journals of Gerontology. Series B, Psychological Sciences and Social Sciences* 52 Spec No (SPEC. ISS.): 37–48. *https://doi.org/10.1093/GERONB/52B.SPECIAL_ISSUE.37*.

Hron, K., M. Templ, and P. Filzmoser. 2010. "Imputation of Missing Values for Compositional Data Using Classical and Robust Methods." *Computational Statistics & Data Analysis* 54 (12): 3095–3107. *https://doi.org/10.1016/J.CSDA.2009.11.023*.

Huque, Md Hamidul, John B. Carlin, Julie A. Simpson, and Katherine J. Lee. 2018. "A Comparison of Multiple Imputation Methods for Missing Data in Longitudinal Studies." *BMC Medical Research Methodology* 18 (1): 1–16. https://doi.org/10.1186/s12874-018-0615-6.

Hurd, Michael D., Paco Martorell, Adeline Delavande, Kathleen J. Mullen, and Kenneth M. Langa. 2013. "Monetary Costs of Dementia in the United States." *The New England Journal of Medicine* 368 (14): 1326–34. *https://doi.org/10.1056/NEJMSA1204629*.

Jack, Clifford R., Marilyn S. Albert, David S. Knopman, Guy M. McKhann, Reisa A. Sperling, Maria C. Carrillo, Bill Thies, and Creighton H. Phelps. 2011. "Introduction to the Recommendations from the National Institute on Aging-Alzheimer's Association Workgroups on Diagnostic Guidelines for Alzheimer's Disease." *Alzheimer's & Dementia : The Journal of the Alzheimer's Association* 7 (3): 257–62. *https://doi.org/10.1016/J.JALZ.2011.03.004*.

Jorm, Anthony F. 2004. "The Informant Questionnaire on Cognitive Decline in the Elderly (IQCODE): A Review." *International Psychogeriatric Association* 16 (3): 1–19. *https://doi.org/10.1017/S1041610204000390*.

Kasper, Judith D, Vicki A Freedman, and Brenda Spillman. 2013. "Classification of Persons by Dementia Status in the National Health and Aging Trends Study."

Kleinke, Kristian. 2017. "Multiple Imputation Under Violated Distributional Assumptions: A Systematic Evaluation of the Assumed Robustness of Predictive Mean Matching:" *Journal of Educational and Behavioral Statistics* 42 (4): 371–404. *https://doi.org/10.3102/1076998616687084*.

Knopman, David S., Rebecca F. Gottesman, Albert Richey Sharrett, Lisa M. Wruck, Beverly Gwen Windham, Laura Coker, Andrea L.C. C. Schneider, et al. 2016. "Mild Cognitive Impairment and Dementia Prevalence: The Atherosclerosis Risk in Communities Neurocognitive Study." *Alzheimer's and Dementia: Diagnosis, Assessment and Disease Monitoring* 2: 1–11. *https://doi.org/10.1016/j.dadm.2015.12.002*.

Kukull, Walter A., Roger Higdon, James D. Bowen, Wayne C. McCormick, Linda Teri, Gerard D. Schellenberg, Gerald van Belle, Lance Jolley, and Eric B. Larson. 2002. "Dementia and Alzheimer Disease Incidence: A Prospective Cohort Study." *Archives of Neurology* 59 (11): 1737–46. *https://doi.org/10.1001/archneur.59.11.1737*.

Kuller, Lewis H., Oscar L. Lopez, James T. Becker, Yuefang Chang, and Anne B. Newman. 2016. "Risk of Dementia and Death in the Long Term Follow Up of the Pittsburgh CHS Cognition Study." *Alzheimers Dement* 12 (2): 170–83. *https://doi.org/10.1016/j.physbeh.2017.03.040*.

Langa, Kenneth M, Brenda L Plassman, Robert B Wallace, A Regula Herzog, Steven G Heeringa, Mary Beth Ofstedal, James R Burke, et al. 2005. "The Aging, Demographics, and Memory Study: Study Design and Methods." *Neuroepidemiology* 25: 181–91. *https://doi.org/10.1159/000087448*.

Langa, Kenneth M., Lindsay H. Ryan, Ryan McCammon, Richard N. Jones, Jennifer J. Manly, Deborah A. Levine, Amanda Sonnega, Madeline Farron, and David R. Weir. 2020. "The

Health and Retirement Study Harmonized Cognitive Assessment Protocol Project: Study Design and Methods." *Neuroepidemiology* 54 (1): 64–74. *https://doi.org/10.1159/000503004*.

Langa, KM, M Kabeto, and D Weir. 2009. "Report on Race and Cognitive Impairment Using HRS in, 2010 Alzheimer's Disease Facts and Figures."

Lee, Jinkook, and Aparajit B. Dey. 2020. "Introduction to LASI-DAD: The Longitudinal Aging Study in India-Diagnostic Assessment of Dementia." *Journal of the American Geriatrics Society* 68 (S3): S3–4. *https://doi.org/10.1111/JGS.16740*.

Little, Roderick J A, and Mark D Schluchter. 1985. "Maximum Likelihood Estimation for Mixed Continuous and Categorical Data with Missing Values." *Biometrika* 72 (3): 497–512.

Liu, C, and DB Rubin. 1998. "Ellipsoidally Symmetric Extensions of the General Location Model for Mixed Categorical and Continuous Data." *Biometrika* 85 (3): 673–88. *https://doi.org/10.1093/biomet/85.3.673*.

Lopez, Oscar L., James T. Becker, Yue-Fang Fang Chang, Robert A. Sweet, Steven T. DeKosky, Michael H. Gach, Owen T. Carmichael, Eric McDade, and Lewis H. Kuller. 2012. "Incidence of Mild Cognitive Impairment in the Pittsburgh Cardiovascular Health Study-Cognition Study." *Neurology* 79 (15): 1599–1606. *https://doi.org/10.1212/WNL.0b013e31826e25f0*.

Lopez, Oscar L., Lewis H. Kuller, Annette Fitzpatrick, Diane Ives, James T. Becker, and Norman Beauchamp. 2003. "Evaluation of Dementia in the Cardiovascular Health Cognition Study." *Neuroepidemiology* 22 (1): 1–12. *https://doi.org/10.1159/000067110*.

Lüdtke, Oliver, Alexander Robitzsch, and Simon Grund. 2017. "Multiple Imputation of Missing Data in Multilevel Designs: A Comparison of Different Strategies." *Psychological Methods* 22 (1): 141–65. *https://doi.org/10.1037/MET0000096*.

Manly, J. J., and R. J. Echemendia. 2007. "Race-Specific Norms: Using the Model of Hypertension to Understand Issues of Race, Culture, and Education in Neuropsychology." *Archives of Clinical Neuropsychology* 22 (3): 319–25. *https://doi.org/10.1016/j.acn.2007.01.006*.

Manly, Jennifer J. 2005. "Advantages and Disadvantages of Separate Norms for African Americans." *Clinical Neuropsychologist* 19 (2): 270–75. *https://doi.org/10.1080/13854040590945346*.

Manly, Jennifer J., Sandra Bell-McGinty, Ming-X. Tang, Nicole Schupf, Yaakov Stern, and Richard Mayeux. 2005. "Implementing Diagnostic Criteria and Estimating Frequency of Mild Cognitive Impairment in an Urban Community." *Archives of Neurology* 62 (11): 1739–46. https://doi.org/10.1001/archneur.62.11.1739.

Manly, Jennifer J, Richard N Jones, Kenneth M Langa, ; Lindsay, H Ryan, Deborah A Levine, Ryan Mccammon, Steven G Heeringa, and David Weir. 2022. "Estimating the Prevalence of Dementia and Mild Cognitive Impairment in the US: The 2016 Health and Retirement Study Harmonized Cognitive Assessment Protocol Project." *JAMA Neurology*, October. *https://doi.org/10.1001/JAMANEUROL.2022.3543*.

Marshall, Andrea, Douglas G. Altman, and Roger L. Holder. 2010. "Comparison of Imputation Methods for Handling Missing Covariate Data When Fitting a Cox Proportional Hazards Model: A Resampling Study." *BMC Medical Research Methodology* 10: 112. *https://doi.org/10.1186/1471-2288-10-112*.

Mayeda, Elizabeth Rose, M. Maria Glymour, Charles P. Quesenberry, and Rachel A. Whitmer. 2016. "Inequalities in Dementia Incidence between Six Racial and Ethnic Groups over 14 Years." *Alzheimer's and Dementia* 12 (3): 216–24. *https://doi.org/10.1016/j.jalz.2015.12.007*.

Mayeux, Richard, Christiane Reitz, Adam M Brickman, Mary N Haan, Jennifer J Manly, M Maria Glymour, Christopher C Weiss, et al. 2011. "Operationalizing Diagnostic Criteria for Alzheimer's Disease and Other Age-Related Cognitive Impairment-Part 1." *Alzheimers Dement* 7 (1): 15–34. *https://doi.org/10.1016/j.jalz.2010.11.005*.

McKhann, Guy M., David S. Knopman, Howard Chertkow, Bradley T. Hyman, Clifford R. Jack, Claudia H. Kawas, William E. Klunk, Walter J. Koroshetz, Jennifer J. Manly, Richard Mayeux, Richard C. Mohs, John C. Morris, Martin N. Rossor, Philip Scheltens, Maria C. Carrillo, Bill Thies, Sandra Weintraub, and Creighton H. Phelps. 2011. "The Diagnosis of Dementia Due to Alzheimer's Disease: Recommendations from the National Institute on Aging-Alzheimer's Association Workgroups on Diagnostic Guidelines for Alzheimer's Disease." *Alzheimer's & Dementia : The Journal of the Alzheimer's Association* 7 (3): 269. *https://doi.org/10.1016/J.JALZ.2011.03.005*.

———. 2011. "The Diagnosis of Dementia Due to Alzheimer's Disease: Recommendations from the National Institute on Aging-Alzheimer's Association Workgroups on Diagnostic Guidelines for Alzheimer's Disease." *Alzheimer's & Dementia* 7 (3): 263–69. *https://doi.org/10.1016/j.jalz.2011.03.005*.

Mclachlan, Geoffrey J, Sharon X Lee, and Suren I Rathnayake. 2019. "Finite Mixture Models." *Annual Review of Statistics and Its Applications* 6: 355–78. *https://doi.org/10.1146/annurev-statistics*.

Mehta, Kala M., and Gwen W. Yeo. 2017. "Systematic Review of Dementia Prevalence and Incidence in United States Race/Ethnic Populations." *Alzheimer's & Dementia* 13 (1): 72–83. *https://doi.org/10.1016/j.jalz.2016.06.2360*.

Mejia-Arango, Silvia, Rene Nevarez, Alejandra Michaels-Obregon, Belem Trejo-Valdivia, Laura Rosario Mendoza-Alvarado, Ana Luisa Sosa-Ortiz, Adrian Martinez-Ruiz, and Rebeca Wong. 2020. "The Mexican Cognitive Aging Ancillary Study (Mex-Cog): Study Design

and Methods." *Archives of Gerontology and Geriatrics* 91 (November): 104210. *https://doi.org/10.1016/J.ARCHGER.2020.104210*.

Morris, JC, A Heyman, RC Mohs, JP Hughes, G van Belle, G Fillenbaum, ED Mellits, and C Clark. 1989. "The Consortium to Establish a Registry for Alzheimer's Disease (CERAD). Part I. Clinical and Neuropsychological Assessment of Alzheimer's Disease." *Neurology* 39 (9): 1159–65. *https://doi.org/10.1212/WNL.39.9.1159*.

Mungas, Dan, Crystal Shaw, Eleanor Hayes-Larson, Charles DeCarli, Sarah Tomaszewski Farias, John Olichney, Hector Hernandez Saucedo, et al. 2021. "Cognitive Impairment in Racially/Ethnically Diverse Older Adults: Accounting for Sources of Diagnostic Bias." *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring* 13 (1): e12265. *https://doi.org/10.1002/DAD2.12265*.

Nasinski, Maciej. 2021. "MiceFast: Fast Imputations Using 'Rcpp' and 'Armadillo.'" *https://CRAN.R-project.org/package=miceFast*.

Okpara, Chinenye, Chidozie Edokwe, George Ioannidis, Alexandra Papaioannou, Jonathan D. Adachi, and Lehana Thabane. 2022. "The Reporting and Handling of Missing Data in Longitudinal Studies of Older Adults Is Suboptimal: A Methodological Survey of Geriatric Journals." *BMC Medical Research Methodology* 22 (1): 122. *https://doi.org/10.1186/S12874-022-01605-W*.

Olkin, I., and R. F. Tate. 1961. "Multivariate Correlation Models with Mixed Discrete and Continuous Variables." *The Annals of Mathematical Statistics* 32 (2): 448–65. *https://doi.org/10.1214/aoms/1177705052*.

Olsen, Maren K., and Joseph L. Schafer. 2001. "A Two-Part Random-Effects Model for Semicontinuous Longitudinal Data." *Journal of the American Statistical Association* 96 (454): 730–45. *https://doi.org/10.1198/016214501753168389*.

Philipps, Viviane, Hélène Amieva, Sandrine Andrieu, Carole Dufouil, Claudine Berr, Jean-François Dartigues, Hélène Jacqmin-Gadda, and Cécile Proust-Lima. 2014. "Normalized Mini-Mental State Examination for Assessing Cognitive Change in Population-Based Brain Aging Studies." *Neuroepidemiology* 43: 15–25. *https://doi.org/10.1159/000365637*.

Plassman, B. L., K. M. Langa, G. G. Fisher, S. G. Heeringa, D. R. Weir, M. B. Ofstedal, J. R. Burke, et al. 2007. "Prevalence of Dementia in the United States: The Aging, Demographics, and Memory Study." *Neuroepidemiology* 29: 125–32. *https://doi.org/10.1159/000109998*.

Prina, A. Matthew, Rosie Mayston, Yu-Tzu Wu, and Martin Prince. 2019. "A Review of the 10/66 Dementia Research Group." *Social Psychiatry and Psychiatric Epidemiology* 54 (1): 10. *https://doi.org/10.1007/S00127-018-1626-7*.

Proust-Lima, Cecile, and Viviane Philipps. 2018. “NormPsy: Normalisation of Psychometric Tests.”

R Core Team. 2020. “R: A Language and Environment for Statistical Computing.” Vienna: R Foundation for Statistical Computing.

Rajan, Kumar B., Jennifer Weuve, Lisa L. Barnes, Elizabeth A. McAninch, Robert S. Wilson, and Denis A. Evans. 2021. “Population Estimate of People with Clinical AD and Mild Cognitive Impairment in the United States (2020–2060).” *Alzheimer’s & Dementia* 17 (12): 1975. *https://doi.org/10.1002/ALZ.12362*.

Reger, Mark A., Robert K. Welsh, G. Stennis Watson, Brenna Cholerton, Laura D. Baker, and Suzanne Craft. 2004. “The Relationship between Neuropsychological Functioning and Driving Ability in Dementia: A Meta-Analysis.” *Neuropsychology* 18 (1): 85–93. *https://doi.org/10.1037/0894-4105.18.1.85*.

Reitan, Ralph M., and Deborah Wolfson. 1992. *Neuropsychological Evaluation of Older Children*. Neuropsychology Press.

Rosen, W. G., R. C. Mohs, and K. L. Davis. 1984. “A New Rating Scale for Alzheimer’s Disease.” *American Journal of Psychiatry* 141 (11): 1356–64. *https://doi.org/10.1176/AJP.141.11.1356*.

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

Rubin, Donald B. 1981. “The Bayesian Bootstrap.” *The Annals of Statistics* 9 (1): 130–34. *https://doi.org/10.1214/aos/1176345338*.

———. 1984. “Bayesianly Justifiable and Relevant Frequency Calculations for the Applied Statistician.” *Annals of Statistics* 12 (4): 1151–72. *https://doi.org/10.1214/AOS/1176346785*.

Rubin, Donald B. 1996. “Multiple Imputation after 18+ Years.” *Journal of the American Statistical Association* 91: 473–89. *https://doi.org/10.1080/01621459.1996.10476908*.

Schafer, J L. 1997. *Analysis of Incomplete Multivariate Data*. Edited by DR Cox, V Isham, N Keiding, N Reid, and H Tong. 1st ed. Chapman & Hall.

Shaw, Crystal, Eleanor Hayes-Larson, M Maria Glymour, Carole Dufouil, Timothy J Hohman, Rachel A Whitmer, Lindsay C Kobayashi, Ron Brookmeyer, and Elizabeth Rose Mayeda. 2021. “Evaluation of Selective Survival and Sex / Gender Differences in Dementia Incidence Using a Simulation Model.” *JAMA Network Open* 4 (3): 1–11. *https://doi.org/10.1001/jamanetworkopen.2021.1001*.

Shaw, Crystal, Yingyan Wu, Scott C Zimmerman, Eleanor Hayes-Larson, Thomas R Belin, Melinda C Power, M Maria Glymour, and Elizabeth Rose Mayeda. 2022. "Comparison of Imputation Strategies for Incomplete Longitudinal Data in Lifecourse Epidemiology." *American Journal of Epidemiology* In Press.

Smith A. 1968. "The Symbol-Digit Modalities Test: A Neuropsychologic Test of Learning and Other Cerebral Disorders. ." In *Learning Disorders*, edited by J Helmuth, 83–91. Seattle: Special Child Publications.

Sonnega, Amanda, Jessica D Faul, Mary Beth Ofstedal, Kenneth M Langa, John WR Phillips, and David R Weir. 2014. "Cohort Profile: The Health and Retirement Study (HRS)." *International Journal of Epidemiology* 43 (2): 576–85. *https://doi.org/10.1093/ije/dyu067*.

Stef van Buuren. 2019. *Flexible Imputation of Missing Data, Second Edition*. *https://stefvanbuuren.name/fimd/sec-rastering.html*.

Steptoe, Andrew, Elizabeth Breeze, James Banks, and James Nazroo. 2013. "Cohort Profile: The English Longitudinal Study of Ageing." *International Journal of Epidemiology* 42 (6): 1648. *https://doi.org/10.1093/IJE/DYS168*.

Tang, Lingqi, Juwon Song, Thomas R. Belin, and Jürgen Unützer. 2005. "A Comparison of Imputation Methods in a Longitudinal Randomized Clinical Trial." *Statistics in Medicine* 24 (14): 2111–28. *https://doi.org/10.1002/SIM.2099*.

Tang, M.-X, P. Cross, H. Andrews, D.M. Jacobs, S. Small, K. Bell, C. Merchant, et al. 2001. "Incidence of AD in African-Americans, Caribbean Hispanics, and Caucasians in Northern Manhattan." *Neurology* 56: 49–56.

Tibshirani, Robert. 1996. "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society. Series B (Methodological)* 58 (1): 267–88. *https://www.jstor.org/stable/2346178*.

Titterington, DM. 1985. *Statistical Analysis of Finite Mixture Distributions*. New York: Wiley.

Trittschuh, Emily H., Paul K. Crane, Eric B. Larson, Brenna Cholerton, Wayne C. Mccormick, Susan M. McCurry, James D. Bowen, Laura D. Baker, and Suzanne Craft. 2011. "Effects of Varying Diagnostic Criteria on Prevalance of Mild Cognitive Impairment in a Community Based Sample." *J Alzheimers Dis* 25 (1): 163–73. *https://doi.org/10.1161/CIRCULATIONAHA.110.956839*.

Tschanz, JoAnn T., Maria C. Norton, Peter P. Zandi, and Constantine G. Lyketsos. 2013. "The Cache County Study on Memory in Aging: Factors Affecting Risk of Alzheimer's Disease and Its Progression after Onset." *International Review of Psychiatry* 25 (6): 673–85. *https://doi.org/10.3109/09540261.2013.849663*.

U.S. Department of Agriculture, and U.S. Department of Health and Human Services. 2020. "Dietary Guidelines for Americans." *https://www.dietaryguidelines.gov/sites/default/files/2020-12/Dietary_Guidelines_for_Americans_2020-2025.pdf*.

Vable, Anusha M., Chloe W. Eng, Elizabeth Rose Mayeda, Sanjay Basu, Jessica R. Marden, Rita Hamad, and M. Maria Glymour. 2018. "Mother's Education and Late-Life Disparities in Memory and Dementia Risk among US Military Veterans and Non-Veterans." *Journal of Epidemiology and Community Health* 72 (12): 1162–67. *https://doi.org/10.1136/jech-2018-210771*.

Walter, Stefan, Jessica R. Marden, Laura D. Kubzansky, Elizabeth R. Mayeda, Paul K. Crane, Shun Chiao Chang, Marilyn Cornelis, David H. Rehkopf, Shubhabrata Mukherjee, and M. Maria Glymour. 2016. "Diabetic Phenotypes and Late-Life Dementia Risk: A Mechanism-Specific Mendelian Randomization Study." *Alzheimer Disease and Associated Disorders* 30 (1): 15–20. *https://doi.org/10.1097/WAD.0000000000000128*.

Weir, David R., Kenneth M. Langa, and Lindsay H. Ryan. 2016. "Harmonized Cognitive Assessment Protocol (HCAP): Study Protocol Summary." *http://hrsonline.isr.umich.edu/index.php?p=shoavail&iyear=ZU*].

White, Ian R., Patrick Royston, and Angela M. Wood. 2011. "Multiple Imputation Using Chained Equations: Issues and Guidance for Practice." *Statistics in Medicine* 30 (4): 377–99. *https://doi.org/10.1002/sim.4067*.

Wilson, R. S., N. T. Aggarwal, L. L. Barnes, C. F. Mendes De Leon, L. E. Hebert, and D. A. Evans. 2010. "Cognitive Decline in Incident Alzheimer Disease in a Community Population." *Neurology* 74 (7): 951–55. *https://doi.org/10.1212/WNL.0b013e3181ec684c*.

Wu, Qiong, Eric J Tchetgen Tchetgen, Theresa L Osypuk, Kellee White, Mahasin Mujahid, and M Maria Glymour. 2013. "Combining Direct and Proxy Assessments to Reduce Attrition Bias in a Longitudinal Study." *Alzheimer Dis Assoc Disord* 27 (3): 207–12. *https://doi.org/10.1097/WAD.0b013e31826cfe90*.

Zhang, Anru, Lawrence D. Brown, and T. Tony Cai. 2019. "Semi-Supervised Inference: General Theory and Estimation of Means." *Annals of Statistics* 47 (5): 2538–66. *https://doi.org/10.1214/18-AOS1756*.

Zhao, Yaohui, Yisong Hu, James P Smith, John Strauss, and Gonghuan Yang. 2014. "Cohort Profile: The China Health and Retirement Longitudinal Study (CHARLS)." *International Journal of Epidemiology* 43 (1): 61–68. https://doi.org/10.1093/IJE/DYS203.