UNIVERSITY OF CALIFORNIA,
IRVINE

On Semi-Parametric Regression for Time-to-Event Analyses in Electronic Health Records Studies

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Statistics

by

Kyle Richard Conniff

Dissertation Committee:
Professor Daniel L. Gillen, Chair
Professor Zhaoxia Yu
Associate Professor Luohua Jiang

2024

TABLE OF CONTENTS

Page

LIST OF FIGURES

LIST OF TABLES

Page

LIST OF ALGORITHMS

ACKNOWLEDGMENTS

I would like to thank my advisor, Daniel L. Gillen, for countless hours of support and infinite patience. I have been extremely lucky to have had his mentorship throughout the program. He supported me and battled for me even before I was his student. I would not have made it through the program without his investment in me.

I would also like to thank my co-advisor, Luohua Jiang. She was instrumental in introducing me to research with indigenous populations and data sovereignty. She helped keep me balanced between application and theory, and she introduced me to the field of indigenous research.

I would further like to thank my final committee member, Zhaoxia Yu. She introduced me to Dr. Jiang after learning about my interest in Indigenous research. She also helped me approach Dr Gillen to be my advisor.

The paper in the appendix is published in Alzheimer's & Dementia: The Journal of the Alzheimer's Association.

This material is based on the work supported by the Eugene Cota-Robles Fellowship at the University of California, Irvine, the National Science Foundation Graduate Research Fellowship under Grant No. DGE-1321846, and the National Institute on Aging Diversity Fellowship under Grant No. R01 AG061189-S1. Any opinions, findings, conclusions, and recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the University of California, the National Science Foundation, or the National Institute of Aging.

VITA

Kyle Richard Conniff

EDUCATION

Doctor of Philosophy in Statistics	**2024**
University of California, Irvine	*Irvine, CA*
Bachelor of Arts in Mathematics	**2016**
St. Norbert College	*De Pere, WI*

RESEARCH EXPERIENCE

Graduate Research Assistant	**2016–2024**
University of California, Irvine	*Irvine, California*

TEACHING EXPERIENCE

Teaching Assistant	**Winter 2024**
University of California, Irvine	*Irvine, CA*
Teaching Assistant	**Summer 2021**
University of California, Irvine	*Irvine, CA*
Teaching Assistant	**Summer 2020**
University of California, Irvine	*Irvine, CA*
Teaching Assistant	**Summer 2018**
University of California, Irvine	*Irvine, CA*

REFEREED JOURNAL PUBLICATIONS

Retention of American Indian and Alaska Native Participants in the National Alzheimer's Coordinating Center Uniform Data Set **2023**
Alzheimer's & Dementia

SOFTWARE

R, SAS

ABSTRACT OF THE DISSERTATION

On Semi-Parametric Regression for Time-to-Event Analyses in Electronic Health Records Studies

By

Kyle Richard Conniff

Doctor of Philosophy in Statistics

University of California, Irvine, 2024

Professor Daniel L. Gillen, Chair

Electronic health records (EHRs) have become a powerful resource for studying health outcomes. In time-to-event settings, EHRs are usually subject to interval censoring on the true time of the event. It is common to consider the outcome to be the time of diagnosis. Standard survival analysis tools for right-censored data, such as the Cox proportional hazards model, are commonly used to estimate covariate associations with the time-to-event in such settings. Patients may, however, have access to multiple health care providers across different systems. If patients seek care from external health systems (a phenomenon we call *system migration*), the diagnosis times within the observed system may be erroneously prolonged. No work has considered the performance of the Cox model under system migration. In this dissertation, we show that system migration related to the outcome of interest results in biased estimates of hazard ratios from the Cox model. We develop an extension to the Cox model that adjusts for system migration by 1) estimating the probability of system migration for each patient and 2) uses multiple imputation to adjust diagnosis times for patients identified as migrating across systems. A vital part of this method involves developing a prediction model from patient-specific system usage patterns for estimating the probability of system migration. To improve prediction assessment, we develop an estimator for time-dependent sensitivity and specificity in the recurrent event setting with unbalanced data across subpopulations. Finally, we consider the choice of time scales for assessing the relative risk of disease diagnosis. We compare the appropriateness of two commonly assumed time scales that define risk sets in the Cox model: the age time scale and the time-on-study time scale. Previous research has suggested that the age time scale, corresponding to birth as a time origin, is most appropriate for epidemiological studies. However, simulation studies have suggested that the time-on-study time scale with covariate adjustment for baseline age is more robust to misspecification of the time scale. We investigate the performance of the Cox model under each time scale under varying degrees of model misspecification and further assess the robustness of each approach when modeling time-varying covariates.

Chapter 1

Introduction

Observational studies oftentimes require assumptions about the structure of the data. End-of-life conditions, such as the time of dementia diagnosis, can be particularly difficult to study, especially in underrepresented groups like American Indian and Alaska Native (AI/AN) populations. Low rates of onset, competing risks (e.g. death), limited access to care/travel/in-home help, and possibly long times until diagnosis are only a few of the many barriers to designing a robust prospective observational study of the time-to-dementia diagnosis [1]. Further, to ensure research is equitable across populations, additional care for culturally appropriate study logistics, recruiting, and retention are components of the overall study design [2, 3]. For example, there are numerous reasons, both historical and recent, for the underrepresentation of AI/AN research participants [4, 5, 3, 6, 7, 8]. To study health related outcomes (including dementia) in AI/AN communities, the Indian Health Services (IHS), Tribes, and researchers at the Centers for American Indian and Alaska Native Health have come together and tapped into a vast wealth of data: electronic health records.

In the last decade, electronic health records (EHRs) have become a popular method of analyzing "real-world" health data on individuals [9, 10]. EHRs offer a glimpse into the lives of patients from a medical perspective [9, 10, 11]. In a world where patients all use the same health care system, EHR data would include complete medical histories of all diagnoses and health care related concerns for the nearly the entire population. (Note that there may still be sampling bias due to individuals who do not seek out health care in this setting.) In such a setting, it is possible to study relationships between many health outcomes with a large enough population. In the current health care system of the United States of America, patients may have many options for seeking health care services. When researchers use EHR data, it is commonly assumed (oftentimes implicitly) that the EHR data contains the full health care picture of the patients. When considering end-of-life conditions such as dementia, however, patients may have access to multiple health care options. For example, the vast majority of people in the US qualify for Medicare coverage at 65 [12].

The IHS EHR database is a massive resources for studying the unique health strengths and challenges of AI/AN communities. The IHS is an agency within the Department of Human and Health Services with the goal of providing the high level of health care to AI/AN peoples across the US [13]. Qualifying individuals (generally, reservation-dwelling Tribal members of one of the US's 573 federally recognized Tribes) have no-cost access to IHS or Tribal health clinics [14]. The health records of patients using IHS or Tribal Health services are stored in the National Data Warehouse by IHS, and a subset of that data is available for research upon approval [15]. However, many AI/AN individuals at-risk for dementia also have access to other health coverage options, such as Medicare, Medicaid, or Private health coverage [16]. Further, the availability of services and the accessibility of clinics is not uniform across Tribes and can sometimes necessitate patients seeking care from non-IHS facilities [17]. This may lead to some care for AI/AN patients taking place at non-IHS and non-Tribal health facilities, resulting in missing or delayed information on health diagnoses for some AI/AN individuals in the available data.

In this dissertation, we assess current practices and develop novel statistical methods to best aid the study of dementia in AI/AN communities. Time-to-dementia (either onset or diagnosis) will be our focus throughout the dissertation. Where applicable, the population of interest will be in AI/AN communities. The statistical methodology developed considers reducing bias in associative modeling as well as predictive model selection. The objective of my research is to allow for valid inference when using Cox's proportional hazards (PH) model [18] for observational time-to-event analyses when true model specification is not straightforward and does not fit typical assumptions.

We begin this dissertation in Chapter 2 with a background on semi-parametric modeling of time-to-event data. We then consider multiple imputation as a method for handling missing data followed by its assumptions and asymptotic arguments. Next, we discuss the evaluation of predicting event status in right-censored survival data through the different settings of time-dependent receiver operating characteristic curves. We end this chapter with the importance of time scales in survival analysis and provide a background on a few papers that have driven the conversation around choosing a time scale. This chapter provides the necessary background for the methods developed as part of this dissertation.

In Chapter 3, we consider the impact of a novel type of missing data that may be unknown to the researcher when analyzing EHR data. When patients seek health care from external health systems, we may be missing important diagnoses, including the time of the diagnosis of interest. We call this phenomenon *system migration*. System migration may result in observed diagnosis times that are delayed relative to a patient's true diagnosis time. If unaccounted for, the delayed diagnosis times may result in biased estimates of the hazard ratios in the traditional Cox model. We propose an extension to the Cox model to account for system migration through a two step process. The first step estimates the probability that a patient migrated out-of-system prior to their diagnosis time. The second step uses multiple imputation to adjust risk sets and map diagnosis times to more natural observation times for a given patient identified as potentially migrating out-of-system. Our simulation studies explore how system migration may impact estimates of the hazard ratios in the traditional Cox model, and how the proposed model can reduce the bias by adjusting for potential system migration. Both the Cox model and the proposed method are then applied to simulated IHS data to obtain an example of how system migration may influence our estimates and conclusions in a pseudo-real-world example if it is ignored.

The proposed method developed in Chapter 3 relies upon correctly identifying patients who are migrating out-of-system. This prediction step involves developing a prediction model for the time-to-return-to-clinic of patients. This is a recurrent event setting with a binary outcome. To our knowledge, no one has considered model selection via area under the receiver operating characteristic curve for recurrent events. In Chapter 4, we discuss methods for assessing predictive model selection in the setting of recurrent event, right-censored time-to-event analyses. Time-dependent estimators of the area under the receiver operating characteristic curve have been developed for settings with a single event. However, these methods will treat recurrent events from a given subject as independent. We develop an extension to these models that allow for intrasubject correlation across events. This method is applied to assess the retention of AI/AN research participants from the National Alzheimer's Coordinating Center Uniform Data Set. Differences in the selected models highlight the potential impact of not accounting for intrasubject correlation between events.

Chapters 3 and 4 develop a method that accounts for system migration and a method for improving the development of prediction models needed for identifying system migration, respectively. We next consider the fundamental aspect of choosing which time scale is best for estimating the relationships between risk factors and dementia diagnosis. In Chapter 5, we compare the two most common time scales considered when using the Cox model. There has been some debate on the choice of time scale and which is best to use for observational and epidemiological studies. Previous simulation studies have considered one of the time scales correct and used simulation to compare the results of using the correct versus incorrect specification of the time scale. We consider a setting where neither time scale is quite correct to assess the bias between models fit under each time scale.

We provide a summary of our contributions and results in Chapter 6. We end with a discussion on future research directions for more robust inference in the setting of missing data and recurrent event prediction.

Chapter 2

Background

2.1 Survival Analysis

In many medical settings, the outcome of interest is the time until some event occurs, say dementia onset. When considering the time until dementia onset, subjects in the population may be excluded from qualifying for the study due to the timing of their event relative to the timing of the conduct of the study, this is called *truncation*. *Left truncation* occurs when subjects are excluded from the study sample because they were not event-free for a sufficient amount of time [19]. For example, if the sample inclusion criteria requires potential participants to be dementia free and 50 years or older, then individuals who do not survive to 50 and individuals who develop dementia younger than 50 will be left truncated. *Right truncation* occurs when subjects are excluded from the sample because they have not experienced the event of interest by a certain age [19]. Returning to the dementia example for left truncation, if the inclusion criteria also requires that individuals have died or been diagnosed with dementia by age 90, then all individuals who survive dementia-free to age 90 will be excluded from the sample and are right truncated. In fact, this study design would suffer from both left and right truncation.

2.1.1 Censoring

The setting where subjects are included in a time-to-event study, but where their time of event is unobserved is called *censoring.* This is a type of missing data where only partial information on the subject's time-to-event is known, but the key value of interest is unknown [20]. Survival analysis methods attempt to handle censoring through careful consideration of what we know about censored subjects without assuming knowledge of their true event times in the estimation procedures. These methods oftentimes assume that the censoring time and the true event time are independent for all subjects conditional on adjustment for covariates collected on the subjects. This assumption, typically termed *non-informative censoring assumption*, says that knowledge of a subject's censoring time does not provide any further information about the subject's true event time conditional on other observed covariates [19]. The assumption of non-informative censoring is similar to the missing at random assumption in general missing data problems, which will be discussed further in the missing data section. If this assumption is broken, many survival methods may produce biased, inconsistent, and unreliable results [19]. Here, we will focus on describing the common types of censoring and we will give brief descriptions of common ways to handle each type of censoring. A more in-depth discussion on censored data and methods for handling censored data can be found in Klein and Moeschberger [19].

The most common type of censoring encountered in practice is *right censoring.* Right censoring is the case where subjects in the study are known to be event-free for some amount of time, but the time of their event is unknown [19]. For example, we observed patient X from age 60 to age 75 and they were dementia-free for the entirety of that time. Then we know that they did not have dementia by age 75, but we do not know if they developed dementia at age 76, or at age 80, or perhaps would never develop dementia during their lifetime. We would consider patient X to be right censored at age 75. Right censoring may exist for several reasons. If the study follow-up or observation period ends before the subject experiences the event of interest, we call this *administrative right censoring.* If a subject dies or experiences another event before the event of interest, we call this *competing risks.* For example, if a subject dies before a dementia diagnosis, their dementia diagnosis time is unknown. Thus, we censor this subject at their time of death. Right censoring can be further categorized by when a subject discontinues the study or is not able to be contacted for further follow-up (lost to follow-up), we call this *random right censoring* [19]. Under administrative right censoring, the non-informative censoring assumption typically holds [19]. With competing risks and random right censoring, however, careful consideration is required to determine that the censoring time is independent of the event time. In the dementia example, patients experiencing memory loss, difficulty communicating, and a reduced ability to plan may not be able to make follow-up appointments for the study and may be lost to follow-up. In this case, these patients are dropping out due to symptoms of dementia and their censoring time may be related to their true time of dementia onset. Right censoring is generally handled by using the information on patients up until the time they are censored. Since we know the patient has not experienced an event for some amount of time, we can take advantage of that and compare their covariate values to people who do experience the event during that period. After the subject has been censored, we no longer use their data for comparisons because we do not know if they are still event-free.

Another type of censoring that is common in medical records data is *left censoring.* Left censoring occurs when the subject has already experienced the event of interest prior to entering the study. In these settings, it is known that the subject has experienced the event of interest, but it is not known when the event occurred, beyond that it has already occurred. For example, in EHR data, patients who have been diagnosed with dementia prior to the date of their first medical record are left-censored. The data may include an indicator of their previous diagnosis, but the time of that diagnosis is unknown. In electronic health records data, left censoring is commonly handled via a washout period, or a time frame after each patient's initial visit where they are observed for the diagnosis of interest [21]. If they have a diagnosis within that timeframe, then they are removed from the study as already having had the event of interest. If they do not have the event of interest, then they are considered event-free and are included in the study from the point at which their washout period ends [21].

A third type of censoring, and one we will focus on in this dissertation, is *interval censoring.* Interval censoring occurs when we know that the event of interest occurred between two times, but we do not know the exact event time. For example, the exact time of dementia onset is rarely known. Instead, we know that patients do not have dementia at one clinic visit and are diagnosed with dementia at the next clinic visit. Usually, the onset of dementia happened between those two visits, and so the true time of dementia onset is interval censored. Right censoring and left censoring can be described in terms of interval censoring. To see these relationships, consider the age of dementia onset. For right censoring, the true age of dementia onset is known to be between the age at which the subject is right censored and infinity. For left censoring, the true age of dementia onset is known to be between birth and the age at which the individual enters into the study. It is important to note that events in EHR data are almost always interval censored. For example, we see a patient without an event at one visit. At the next visit, the patient is diagnosed with the event of interest. Then the time that the event occurred is between those last two points of observation. One way that this is typically handled is to define the primary outcome for the study as the time until diagnosis as the event of interest (as opposed to the time of disease onset).

It is possible for one, two, or all three of these types of censoring to occur. In fact, EHR databases oftentimes include all three [22]. As such, it is important for a researcher to define inclusion and exclusion criteria as well as their outcomes and predictors of interest carefully to limit the impact of these types of censoring.

2.1.2 Statistical Functions of interest in time-to-event data

Survival analysis is focused on maintaining important statistical properties while drawing inference on population parameters in time-to-event settings where censoring is present [19]. Traditional statistical functionals of interest, i.e. the probability density function (PDF) and the cumulative distribution function (CDF), may be of interest in time-to-event settings. We additionally introduce functionals that may be more interpertable. The *survival distribution*, defined as one minus the CDF, answers the probability an event has not been experienced by a given time [19]. We are also typically interested in the rate at which an individual with some characteristics will experience the event in the next instant, given that they have not experienced the event to this point, which is the *hazard function*.

Regression methods in survival analysis tend to focus on estimating the hazard function, or the rate at which an event occurs at a given time conditional on surviving to that time. The hazard function is a more natural quantity to estimate in the presence of censoring because it conditions on a subject having not experienced the event of interest by a given time. This conditioning allows for censored subjects to contribute to the estimation process by providing information to the model up until their censoring time without requiring knowledge of the event time from those subjects [19].

Each of the previously mentioned quantities can be calculated from knowing any one of them. Below, the functionals, plus the cumulative hazard function, are given as well as their relationships to the PDF.

1) Probability Density Function (PDF):

$$f(t) = \lim_{\Delta t \to 0^+} \frac{1}{\Delta t} \Pr(t \leq T < t + \Delta t)$$

2) Cumulative Distribution Function (CDF):

$$F(t) = \Pr(T \leq t) = \int_0^t f(u)du$$

3) Survival Distribution:

$$S(t) = \Pr(T > t) = 1 - \Pr(T \leq t) = 1 - F(t) = 1 - \int_0^t f(u)du$$

4) Hazard Function:

$$\lambda(t) = \lim_{\Delta t \to 0^+} \frac{1}{\Delta t} \Pr(t \leq T < t + \Delta t \mid T \geq t) = \frac{f(t)}{S(t)}$$

5) Cumulative Hazard Function:

$$\Lambda(t) = \int_0^t \lambda(u)du = -\log(S(t)) = -\log\left(1 - \int_0^t f(u)du\right).$$

2.1.3 Cox Model

The most commonly used regression model in survival analysis is the *Cox Proportional Hazards model* (Cox model) [18]. The Cox model considers the hazard function as a multiplicative function of the covariates and adjustment variables. Let t be time, $z_1, \ldots, z_p$ be covariates, and $\beta_1, \ldots, \beta_p$ be parameters values we are interested in estimating, then the Cox model parameterizes the hazard function as:

$$\lambda(t \mid z_1, \ldots, z_p) \ = \lambda_0(t)\ e^{\beta_1 z_1 + \ldots + \beta_p z_p}$$

where $\lambda_0\left(t\right)$ is the hazard rate when all the covariates are zero, or the *baseline hazard.* In this setting, e^{β_j} can be interpreted as the relative difference in the hazard function comparing two subpopulations that differ by one unit in z_j, holding all other covariates constant. Note that this interpretation, and β_j in general, does not depend on t. The "proportional hazards" part of the Cox model is a major assumption which specifies that the effect of a covariate on the hazard is constant over time, which means the effect of z_j on the hazard function is not a function of time [19].

The Cox model accounts for censoring by defining $\delta_i = I\left(T_i \leq C_i\right)$ as the event indicator, where $I\left(\cdot\right)$ is the indicator function returning 1 if the input statement is true and 0 if the input statement is false, T_i is the true event time for subject i, and C_i is the true censoring time for subject i. We also define z_i as the covariate values for the i^{th} subject, and $X_i = \min\left(T_i, C_i\right)$ as the observed time for subject i, where $x_i = T_i$ indicates that $\delta_i = 1$.

The Cox model is powerful in observational studies because it allows for adjustment of potential confounding factors while not assuming a fully parametric model [19]. In fact, the baseline hazard in the model cancels out in the estimation procedure, so only the effects of the covariates are estimated [19]. The Cox model is an example of a semi-parametric model because the baseline hazard is non-parametric and may be infinite dimensional while the functional form of the covariates is explicitly specified [19].

This estimation procedure for the Cox model involves maximizing the partial likelihood. The partial likelihood comes from the conditional probability that a subject with a given covariate value experiences the event of interest at a particular time, given that some subject event occurs at that time [23]. For subject i experiencing an event at time t_i, the partial likelihood is:

$$L_i = \Pr(\text{subject with } z_i \text{ experiences the event at } \ t_i \mid \ \text{an event occurs at } t_i)$$

$$= \frac{\Pr(\text{subject with covariate vector } z_i \text{ experiences the event at } t_i)}{\Pr(\text{some subject in the risk set experiences an event at } t_i)}$$

$$= \frac{\lambda_i(t_i)\Delta t}{\sum_{k\in R(t_i)} \lambda_k(t_i)\Delta t}$$

$$= \frac{\lambda_i(t_i)}{\sum_{k\in R(t_i)} \lambda_k(t_i)},$$

where λ_i is the hazard function for subject i, and $R\,(t_i)$ is the risk set at time t_i. Note that Δt does not depend on the event time, which is why it cancels out in the last line. To get the i^{th} subject's contribution to the partial likelihood for the Cox model, we can substitute in the Cox model's equation for $\lambda(t)$:

$$L_i = \frac{\lambda_0(t_i)\ e^{z_i^T\beta}}{\sum_{k\in R(t_i)} \lambda_0(t_i)e^{z_k^T\beta}}$$

$$= \frac{e^{z_i^T\beta}}{\sum_{k\in R(t_i)} e^{z_k^T\beta}}.$$

Thus, assuming independent event times, the partial likelihood for all D event times is:

$$L = \prod_{i=1}^{D} L_i = \prod_{i=1}^{D} \frac{e^{z_i^T\beta}}{\sum_{k\in R(t_i)} \ \ e^{z_k^T\beta}}.$$

Similar to ordinary least squares and generalized linear models, we can estimate parameter values via setting the partial score function equal to zero and solving. This requires differentiating the log-partial likelihood:

$$\log(L) = \sum_{i=1}^{D} \log\left(\frac{e^{z_i^T\beta}}{\sum_{k\in R(t_i)} e^{z_k^T\beta}}\right)$$

$$= \sum_{i=1}^{D}\left(z_i^T\beta - \log\left(\sum_{k\in R(t_i)} e^{z_k^T\beta}\right)\right)$$

$$\Rightarrow \frac{\partial \log(L)}{\partial\beta} = U(\beta) = \sum_{i=1}^{D}\left(z_i - \sum_{k\in R(t_i)} \frac{z_k e^{z_k^T\beta}}{\sum_{l\in R(t_i)} e^{z_l^T\beta}}\right).$$

The combination of (X_i, δ_i, z_i) allows the Cox model to consider the covariate values for subject i from their time of enrollment in the study until the time of their event or being censored, after which time they no longer contribute to the estimation procedure [19]. In other words, at a given event time, t, the Cox model compares covariate values of the person who experienced the event at time t to the covariate values of everyone still at-risk for experiencing the event. The group of subjects still at-risk for the event are referred to as the *risk set*. The risk set consists of all subjects such that $x_i \geq t$ [19].

For purposes of asymptotics in later chapters, the score function of the partial likelihood can be rewritten in counting process notation. Let $Y_j\left(t\right) = I\left(X_j \geq t\right)$ and $N_i\left(t\right) = I\left(X_i \leq t \cap \delta_i = 1\right)$, then

$$U(\beta) = \sum_{i=1}^{n} \int_{t=0}^{\infty} Z_i - \frac{n^{-1}\sum_{j=1}^{n} Z_j Y_j\left(t\right) e^{Z_j\beta}}{n^{-1}\sum_{j=1}^{n} Y_j\left(t\right) e^{Z_j\beta}} dN_i\left(t\right) = 0 \tag{2.1}$$

where we let the covariate values be random. There is no closed form solution to this equation, thus, in practice, the Newton-Raphson algorithm is used to solve $U(\hat{\beta}) = 0$ to find the parameter estimates [19]. Counting process notation allows us to appeal to Rebolledo's Martingale Central Limit Theorem and state that $\hat{\beta} \sim N_p(\beta_0, \mathcal{I}^{-1}(\beta_0))$, where β_0 denotes the vector of true value of β and $\mathcal{I}(\beta_0) = -\mathrm{E}\left(\frac{\partial U(\beta)}{\partial \beta} \mid_{\beta=\beta_0}\right)$ is the partial information matrix [24]. In practice, it is common to replace $\mathcal{I}\left(\beta_0\right)$ with the observed information matrix $I\left(\beta_0\right) = \frac{\partial U(\beta)}{\partial \beta} \mid_{\beta=\beta_0}$ [24].

2.2 Missing Data

Data collection is never a perfect process. Sometimes collected data is incomplete and missing values. Formally, "missing data are unobserved values that would be meaningful if observed" [20]. For example, in EHR data, some patients may be missing demographic information such as race or income because they did not answer those questions on the form. In this example, since patients have a race and income, but chose not to answer those questions, the data is clearly missing. Another type of missing data may occur with abnormally long gaps between visits. Some patients may seek care from a different health care system, so that a diagnosis is made, but it is unobserved in our data set. This characterizes missing data that may contain important information for the patient's overall health picture, however, we may not even know the data is missing. Two common approaches for dealing with missing data are 1) remove the subjects with the missing data and perform a complete case analysis, or 2) fill in the missing values [20]. To understand the appropriateness of these methods, it is necessary to first understand the patterns and mechanisms behind missing data. We will start with a discussion of these, then we will discuss a few common methods for handling missing data. We will end with a discussion of multiple imputation. More information on all the following topics can be found in Little and Rubin (2019) [20].

2.2.1 Types of Missing Data

Little and Rubin (2019) make the distinction between missingness patterns and missingness mechanisms. *Missingness patterns* describe how the missing values look, or which values are missing and which values are observed. Many of these patterns are named after how they data look in a matrix with blank spaces for the missing values. *Missingness mechanisms* describe why values are missing, or the relationships between the missingness and the variables related to the question at hand. Here we will give a brief overview of missingness patterns. Then we will discuss the mechanisms that lead to missing data.

Missingness Patterns

To describe the patterns of missing data, consider Y, the $n \times (p+1)$ matrix consisting of individuals on the rows $(i = 1 \ldots, n)$ on the rows, and the outcome $(j = 1)$ and covariates $(j = 2, \ldots, p+1)$ on the columns. Then (y_{ij}) is the observed value for the subject i in the j^{th} column. We define the pattern of missing data through rearranging the rows and columns of the missingness indicator matrix, $M = (m_{ij}) = I\,(y_{ij} \text{ is missing})$. Thus, $m_{ij} = 1$ if the y_{ij} value is unobserved, and $m_{ij} = 0$ if the y_{ij} value is observed. The five main patterns of missing data we will mention here are univariate non-response, multivariate non-response, monotone missingness, file matching missingness, and general missingness.

Univariate non-response corresponds to the setting where only a single variable is missing, or there are missing values in only one column of M. For example, this may occur when medical records are collected across several healthcare facilities, and all facilities collect the same information, except for one facility that does not collect information on, say, income. Then subjects from that facility will be missing income information, while all the other columns are collected completely.

Multivariate non-response occurs when multiple columns of M are missing values for the same individuals. In the example above, perhaps one facility does not collect information on income or race. Then two columns will have missing values, but only for patients from the one facility. Note that multivariate non-response can involve multiple sets of this pattern. Say, in the above example, another facility does not collect age and sex information. Then the missingness pattern would be multivariate non-response with two patterns.

Monotone missingness occurs when values for a subject are observed until one is missed, then all values after that are also missing. Longitudinal data is a common setting where monotone missingness can occur. Some patients may be observed for all follow-up, some patients may be observed for most follow-up, and some patients are observed for only a little follow-up. This *attrition* or *loss to follow-up* results in a setting where no further information can be collected on these patients.

File matching missingness occurs when two variables are never observed jointly. For example, when combining medical records from multiple facilities, say one facility is a neurological center and another is an oncology center. Both centers may collect demographic information about their patients, but the neurological center will have additional information on neurological symptoms that the oncology center may not collect. Similarly, the oncology center will have information that the neurological center may not collect. This results in columns of M that are the opposite of each other.

General missingness refers to missing data that does not exhibit any of these patterns. Missing values may be dispersed across different variables for different patients seemingly without meaning. This can be the most difficult form of missingness to handle depending on the amount of missing data and the relationships between the missing values and the observed variables.

Missingness Mechanisms

Here we consider the relationships between the missing values and the variables collected in the data. We will assume that individuals are independent and identically distributed, (i.e. that the rows of (y_i, m_i) are *i.i.d.* over i). We can characterize the missingness mechanism by the conditional distribution of m_i given y_i, say $f_{M|Y}(m_i \mid y_i, \theta)$ where θ are unknown parameters. There are three missingness mechanisms we will classify: missing completely at random (MACR), missing at random (MAR), and missing not at random (MNAR).

MCAR missingness is defined as missing values that are unrelated to data, collected or uncollected. That is, the missing values are MCAR if, for each subject i and distinct values for y_i, y_i^* in the sample space of Y,

$$f_{M|Y}(m_i \mid y_i, \theta) = f_{M|Y}(m_i \mid y_i^*, \theta).$$

When MCAR is the missingness mechanism, the data set consisting of the complete cases is a random sub-sample of the sample. Thus, an analysis with that data is valid for the same population as the full sample.

MAR missingness occurs when the missing values are only related to observed variables. Let $y_{(0)i}$ denote the values of y_i that are observed and let $y_{(1)i}$ denote the values of y_i that are missing. That is, the missing values are MAR if, for subject i and distinct missing values $\left(y_{(1)i}, y_{(1)i}^*\right)$ in the sample space of $y_{(1)i}$,

$$f_{M|Y}\left(m_i \mid y_{(0)i}, y_{(1)i}, \theta\right) = f_{M|Y}\left(m_i \mid y_{(0)i}^*, y_{(1)i}^*, \theta\right).$$

When MAR is the missingness mechanism for the response variable, Y, and we have a set of fully observed predictors, X, then the missingness of Y cannot be dependent on Y. In this setting, M and Y are conditionally independent given X. Thus,

$$f_{Y|M}\left(y_i \mid x_i, m_i = 1, \theta, \beta\right) = f_{Y|M}\left(y_i \mid x_i, m_i = 0, \theta, \beta\right),$$

where β and θ are unknown parameters. Using this equation, we can reasonably predict the values for $Y \mid m_i = 1$. Thus, we can build a "complete" data set by replacing the $Y_i \mid m_i = 1$ with $\hat{Y}_i \mid m_i = 1$.

MNAR missingness occurs if the missing values depend on what the missing values are. That is, the distribution of m_i depends on the missing components of y_i,

$$f_{M|Y}\left(m_i \mid y_{(0)i}, y_{(1)i}, \theta\right)$$

which does not simplify further and where θ represents unknown parameters. When the missingness mechanism for the outcome, Y, is MNAR the distribution of missing values for y_i depends on those missing values,

$$f_{Y|M}\left(y_i \mid x_i, m_i = 1, \theta, \beta\right),$$

where β and θ represent unknown parameters. Similar to the distribution of $f_{M|Y}$, the distribution $f_{Y|M}$ does not simplify further.

If MNAR is the true missingness mechanism, then additional assumptions are generally required. It is necessary to supplement results with sensitivity analyses including the worst possible cases as well as other potential cases. More information can be found in Little and Rubin (2019) [20].

In terms of survival analysis, right censored data is an example of the MAR missingness mechanism under the non-informative censoring assumption. The missing event time is because the true event time occurs after the data collection ends. The non-informative censoring assumption, however, means that the event time and the censoring time are independent conditional on observed covariates. Thus, missing event times have the same distribution of the observed event times. Accounting for the time up until censoring we avoid biasing the results in time-to-event data [20].

2.2.2 Methods for Handling Missing Data

Here we will discuss common methods for handling missing data. It is necessary to carefully consider the mechanism that describes the missing data and to perform sensitivity analyses to assess how assumptions about the missingness impacts the results. In this section, we will consider complete-case, single imputation, and multiple imputation analyses. More information on these topics can be found in Little and Rubin (2019) [20].

Complete-Case Analysis

Complete-case analyses handle missing data by removing subjects with missing data and performing the analysis on the remaining data as though it is the complete sample. This approach can be advantageous because it is simple and allows for using standard statistical approaches without modification. The disadvantage of this method is that information is discarded, which results in reduced precision and potential bias. If MCAR is the missingness mechanism, then a complete-case analysis provides valid but inefficient results to the full sample populations. If the missingness mechanism is MAR or MNAR, then the analysis may be biased and invalid. In practice, when calculating univariate statistics, it is wasteful to discard known values for subjects who are missing a value from a different variable. An *available-case analysis* uses all available data to calculate univariate statistics to summarize the sample. In this setting, all available values for a given variable contribute to the summary statistics, even if some of those subjects may be removed for a larger analysis or regression.

While this limits the amount of information that is wasted, it is important to characterize which variables are missing information, stratified by key variables in the analysis (e.g. the response variable or predictors of interest), so that it is clear if subjects with missing values are inherently different from subjects without missing values.

Single Imputation

The complete-case and available-case analyses do not consider the relationships between variables as a method for filling in missing data. For example, if a missing value Y_i is highly correlated with an observed covariate X_i, then we could use X_i to predict the value of Y_i and include the imputed value in the analysis. Single imputation refers to the setting where missing values, $y_{(1)i}$ are imputed once each as draws from some distribution, $f(y_{(1)} \mid x, y_{(0)}, \theta)$, where $y_{(1)}$ are the missing outcome values, θ are unknown parameters, x are observed covariates, and $y_{(0)}$ are observed outcome values from the full sample. The analysis is then performed on the data set which includes these imputed values. The key component of imputation methods involve how the imputed values are determined. Ideally, $f(y_{(1)} \mid x, y_{(0)}, \theta)$ is the predictive distribution of $y_{(1)}$. Since this distribution is unknown, it is necessary to estimate the predictive distribution based on the observed data. Common methods of estimating the predictive distribution of the missing values include: mean imputation, regression imputation, hot deck imputation, stochastic regression imputation [25], and composite methods [26, 27]. More information on each of these methods can be found in Little and Rubin [20].

Single imputation can be attractive as a method because it can appear as though the data is complete, and statistical methods can be applied to the data per usual. However, only imputing a value once does not account for the uncertainty in that predicted value, which generally results in underestimated standard errors. Further, if the imputation method did not draw values from the correct distribution, then results may be biased. In general, imputation methods should be conditional on observed variables because it reduces bias, improves precision, and preserves associations between variables. Imputation methods should also be multivariate, where applicable, and the imputed values should be draws from a predictive distribution [20].

2.2.3 Multiple Imputation

One way to account for the uncertainty of predicted values is to impute each missing value multiple times, known as *multiple imputation*. The method of multiple imputation was developed from a Bayesian perspective with imputed values being draws from a posterior predictive distribution of the missing values. These missing values should be imputed by independent draws from the *posterior predictive* distribution of the missing values, $Y_{(1)}$. Posterior predictive means we condition on the observed values, $Y_{(0)}$, when predicting the missing values. The method involves creating multiple data sets, where each data set has unique draws (or imputations) from the predictive distribution of the missing values. We consider a standard Frequentist approach where our predictive distribution is developed from a prediction model, rather than a posterior predictive distribution. Estimates and standard errors from each of the imputation data sets are combined with estimates of the uncertainty in the imputation process. The uncertainty from the imputation process can be estimated using the variation in the estimates from each imputation data set. Multiple imputation shares all the advantages of single imputation, while making up for the disadvantages by accounting for the uncertainty around the imputed values. The only disadvantage of multiple imputation compared to single imputation is that multiple imputation requires more steps for the analysis. That is, multiple data sets must be created, the model is run multiple times, and the results from those models are combined.

To perform multiple imputation, we will draw D imputations for each missing value, $Y_{(1)}$, from a prediction model, $f(y_{(1)} \mid x, y(0), \hat{\gamma})$, where x are observed covariates that are predictive of the observed outcome values $y(0)$, and $\hat{\gamma}$ are estimated parameters for the predictive distribution of Y. Within each of the $d = \ 1, \ldots, D$ imputation data sets, estimate θ and $\text{var}(\hat{\theta})$ with $\hat{\theta}_d$ and $\hat{\text{V}}(\hat{\theta}_d)$. Then we combine the estimates such that:

$$\bar{\theta}_D = \frac{1}{D} \sum_{d=1}^{D} \hat{\theta}_d$$

and

$$\hat{V}(\bar{\theta}_D) = \frac{1}{D} \sum_{d=1}^{D} \hat{V}(\hat{\theta}_d) + \frac{D+1}{D} * \frac{1}{D-1} \sum_{d=1}^{D} (\hat{\theta}_d - \bar{\theta}_D)^2.$$

Averaging over the imputation data sets improves efficiency over estimation in the single imputation setting. The variance term can be decomposed as weighted average of the within-imputation variance, $V(\hat{\theta}_d)$ and the between-imputation variance $\frac{1}{D-1} \sum_{d=1}^{D} (\hat{\theta}_d - \bar{\theta}_D)^2$, where $\frac{D+1}{D}$ serves as a finite imputation adjustment.

Inference and interval estimation from the above procedure typically comes from the asymptotic distribution. $\bar{\theta}_D$ is normal with the variance $V(\bar{\theta}_D)$ if the following conditions hold: 1) the imputation method is proper, 2) the complete-data inference method is valid, and 3) the sample is large enough to appeal to asymptotic arguments. In this setting, a proper imputation method must satisfy that 1) the multiple imputation method inferences follow the distribution of the complete-data estimates, 2) the within-imputation set variance component is centered at the variance of the complete-data statistics, and 3) between-imputation variance is controlled. More information can be found in [28].

2.3 Receiver Operating Characteristic Curves

Oftentimes in survival analysis, the instantaneous rate of failure is modeled, but the interest truly lies in predicting if the event will occur. Thus, the underlying goal is building a model or marker adept at classification for a binary outcome. *Receiver operating characteristic* (ROC) curves are a common method of comparing the performance of classifiers and markers across a wide-range of possible classification thresholds. In this section, we will discuss ROC curves in the binary setting first. Then we will discuss time-dependent ROC curves appropriate in survival analysis.

2.3.1 Sensitivity, Specificity, and Area Under the Curve

In the setting of a fixed binary outcome, ROC curves portray the ability of a classifier to distinguish event status by plotting sensitivity against (1 – specificity) of a classifier over the range of possible thresholds. Let's say our event of interest disease status and our classifier is a test for presence of the disease in a subject. *Sensitivity* is the probability that the test result is positive for the disease given that the subject has the disease. *Specificity* is the probability that the test is negative for the disease given that the subject does not have the disease. In practice, it is common to summarize an ROC curve by the area under the ROC curve (AUC), which estimates the probability that a randomly sampled subject with the disease has a higher risk score for the disease than a randomly sampled subject who is disease-free. An AUC equal to 0.5 suggests that the classifier is no better at identifying disease than random chance. The closer AUC is to 1, the better the classifier is at separating subjects with the disease from subjects without the disease. ROC curves are a powerful tool because they allow for comparison of different markers and models across all possible thresholds. These comparisons are possible as long as the markers and models are identifying the same outcome, regardless of how they come to the identification of that outcome [29].

2.3.2 Time-Dependent ROC Curves

Since disease status is rarely fixed across time, Heagerty et al (2000) proposed time-dependent ROC curves [30]. These curves estimate sensitivity, specificity, and AUC at a meaningful (i.e. clinically relevant) time point. Estimating these quantities in a time-dependent setting requires attention to the definitions of which subjects are allowed to be cases and controls. For example, if interest is in studying the performance of the classifier at the median event time (say, $t = 7$), then should our controls (i.e. non-diseased subjects) be everyone who has not developed the disease before $t = 7$? Or should a subject with a disease time of $t = 5$ contribute to the specificity for $t < 5$, and contribute to the cases at and/or after $t = 5$? This distinction led to the a few different specifications for time-dependent ROC methods depending on how they deal with the changing of disease status. For estimating specificity, controls are defined as either *static*, meaning that everyone who is disease-free until some time t^* is considered a control, or *dynamic*, meaning that subjects contribute as controls up until their event time. Similarly, for estimating sensitivity, cases are defined as either *incident*, meaning that a subject is only defined as a case at the time of their event, or *cumulative*, meaning that a case is all subjects who experienced the event up until (and including) time t. From these four settings, we can define *dynamic specificity* as the probability that subject i's ($i = 1, \dots, n$) risk score is smaller than some threshold given that their event time, T_i, is greater than some time t. *Static specificity* is defined as the probability that subject i's risk score is smaller than some threshold, c, given that their event time is greater than some fixed time t^*, where t^* is usually taken be some time equal to or greater than the time of interest in the study. These expressions can be written as: $\text{Spec}_D(c,\ t) = Pr(M_i \leq c \mid T_i > t)$ and $\text{Spec}_{\bar{D}}(c, t^*) = Pr(M_i \leq c \mid T_i > t^*)$, where $\bar{D}$ is used to denote "static" to avoid confusion. Similarly, *cumulative sensitivity* is defined as the probability that subject i's risk score, M_i, is greater than some threshold value, c, given that the subject's event time, T_i, is between baseline and some specified time t.

Incident sensitivity is defined as the probability that subject i's risk score is greater than some threshold given that their time of event is exactly time t. These can be expressed by: $\text{Sens}_C(t, c) = Pr(M_i > c \mid 0 < T_i \leq t)$ and $\text{Sens}_I(t, c) = P(M_i > c \mid T_i = t)$. These ideas are summarized by identifying the methods for estimating time-dependent sensitivity and specificity as belonging to one of three categories: cumulative sensitivity and dynamic specificity (C/D), incident sensitivity and static specificity ($I/\bar{D}$), or incident sensitivity and dynamic specificity (I/D). Several estimators have been proposed for each, but we will focus on the incident/dynamic estimator of Heagerty and Zheng (2005) [31] and the cumulative/dynamic estimator of Chambless and Diao (2006) [32].

Heagerty and Zheng (2005) [31] developed an I/D estimator to allow for subjects to contribute to the specificity calculation up until the time of their event. At their time of event, they contribute to the calculation of the sensitivity. After their event time, subjects are removed from the risk set and contribute no further to either estimator. This mirrors how a subject contributes information to the partial likelihood of the Cox model. The sensitivity in their estimator measures the expected proportion of subjects with a risk score greater than the threshold given that the individuals fail at time t. The specificity in their estimator measures the expected proportion of subjects with a risk score smaller than the threshold given that they have not experienced an event prior to time t. This definition effectively splits the risk set at time t into subjects who have the event at time t and subjects who are event-free past time t. Besides the natural relationship with how information is contributed to the partial likelihood, the I/D framework also allows for time-varying risk scores, time-specific summary measures, and time-average summary measures. When the interest is in assessing the accuracy of a Cox model under proportional hazards, Heagerty and Zheng proposed estimating Sens_I with

$$\widehat{\text{Sens}}_{HZ}(t, c) = \sum_k I(M_k > c) \frac{Y_k(t) e^{M_k \hat{\gamma}}}{\sum_j Y_j(t) e^{M_j \hat{\gamma}}}$$

where M_k is the risk score generated as the linear predictors from a Cox model and $Y_k(t) = I(T_k > t)$ is the at-risk indicator for subject k. They propose estimating Spec_D with

$$\widehat{\text{Spec}}_{HZ}(t, c) = \sum_k I(M_k \leq c) \frac{Y_k(t+)}{\sum_j Y_j(t+)}$$

where $Y_k(t+) = \lim_{\delta \to 0} Y_k(t + |\delta|)$. More information can be found in [31].

Alternatively, Chambless and Diao (2006) [32] developed sensitivity and specificity estimators in the C/D framework. They similarly considered the setting where the linear predictors estimated via a Cox model serve as the risk score. Heagerty and Zheng (2005) [31] note that the C/D framework is appropriate when the scientific goal is to develop a risk score that distinguishes between subjects who survive event-free past t' and those who experience the event prior t', where t' is a singular time, or a collection of times $t_1\prime, \ t_2\prime, \ \ldots, \ t_q\prime$, of scientific interest. This framework classifies every subject as either a case or a control for every fixed time t, so that subjects contribute to the estimation of the specificity prior to their event time and contribute to the estimation of the sensitivity at all $t \geq T_i$, where T_i is their event time. Chambless and Diao (2006) [32] denote $\text{Sens}_{CD}(t, c) = \frac{E(1-S(t|M))I(c<M)}{E(1-S(t|Z))}$ and $\text{Spec}_{CD}(t, c) = \frac{E(S(t|M)I(c<M))}{E(S(t|M))}$, where $E(\cdot)$ is the expectation, M are the linear predictors from a Cox model, and $S(t \mid M)$ is the survival function. The values can be estimated by replacing the expectations with means over all subjects and $S(t \mid M)$ with estimated survival functions from a Cox model, $\hat{S}(t \mid Z) = exp\left\{-\hat{\Lambda}_0(t)exp(Z\hat{\beta})\right\}$, where $\hat{\Lambda}_0(t)$ is the Breslow estimator for the baseline hazard [33] and $\hat{\beta}$ is estimated from the Cox model.

2.4 Time Scales

Survival analysis models consider the time until an event occurs. The time component defines the ordering of subject event times as well as the risk sets at each event time. We typically define time from some origin, called the *time origin*, which is the point at which subjects are initially considered at-risk for the event. The time origin defines how time is measured in the model, which is called the *time scale* [34]. For the Cox model, there are two common time scales: age and time-on-study [34]. In the age time scale, the time origin is birth and the study is measuring the risk of experiencing the event at a particular age [35]. When fitting a Cox model under the age time scale, no covariate adjustment for age is included [35]. In the time-on-study time scale, the time origin is the first visit, or examination, or study enrollment. This time scale measures the risk of experiencing the event from the time observation begins. When fitting a Cox model with this time scale, baseline age or age-at-entry to the study is included as an adjustment covariate [34]. In general, the Cox model uses the time scale to order the events and determine risk sets [34]. The time of the events is not accounted for in the estimation of covariates [34]. Figure 2.1 displays how the age, left truncated age, and time-on-study time scales may order events differently. The age time scale will order events from youngest to oldest in terms of age. The left truncated age time scale also orders events by age from youngest to oldest, but it only compares people who were enrolled in the study at similar ages. In this time scale, the risk set for individual i who experiences the event at time, a_i, consists only of individuals, j, such that $a_{0j} \leq a_i \leq a_j$, where a_{0j} is the age-at-entry to the study for individual j. The time-on-study time scale orders events in relation to how long the subjects have been enrolled in the study. In this section, we consider previous work that has suggested when the models may estimate the same quantities. Then we discuss the performance of the models in previous simulation studies.

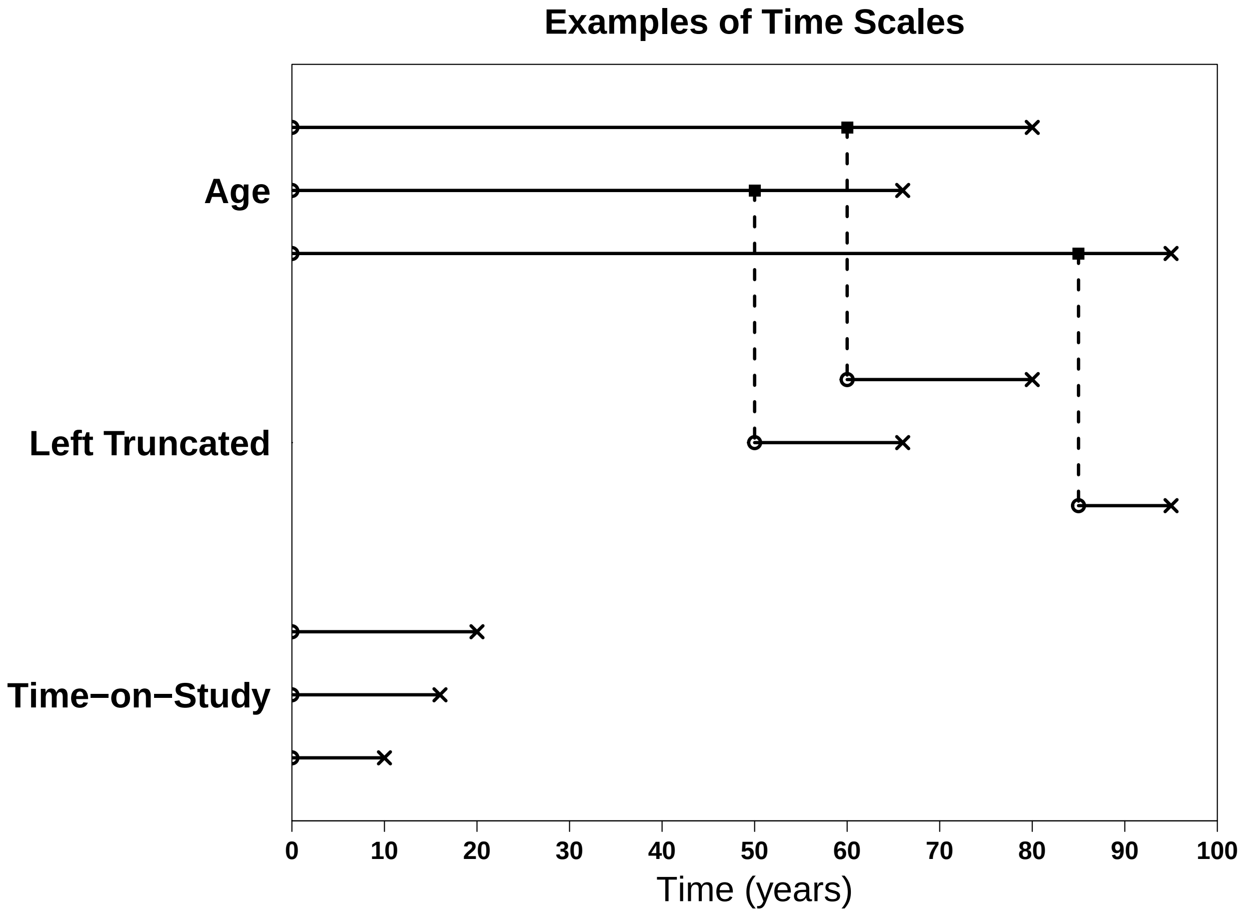


Figure 2.1: This figure presents an example of the three most common time scale specifications. The squares represent the age at which the individuals entered the study. They X's represent when the individuals experienced the event of interest.

2.4.1 Models and Likelihoods

Korn et al. (1997) previously hypothesized that there are two settings where the age time scale and the time-on-study time scale would be equivalent [36]. Under the age time scale, let a be the observed age of the event or censoring and x be the vector of covariate values across subjects ($i = 1, \ldots, n$). The Cox model for the age time scale is:

$$\lambda_A(a \mid x) = \lambda_{0A}(a)e^{\beta x}, \tag{2.2}$$

where $\lambda_{0A}(a)$ is the baseline hazard for experiencing the event at a particular age and β is the coefficient for x. Similarly, let t be the observed time of event or censoring for a subject, then the Cox model with time-on-study as the time scale is:

$$\lambda_T(t \mid x, a_0) = \lambda_{0T}(t)e^{\beta x + \gamma a_0}, \tag{2.3}$$

where β is the coefficient for x, γ is the coefficient for a_0, and $\lambda_{0T}(t)$ is the baseline hazard.

For these two models, Korn et al. (1997) [36] argued that inference on the covariate x would be identical if 1) $\lambda_{0A}(a) \approx ce^{\psi a}$, for some ψ, and 2) if the covariate of interest and baseline age are independent. Simulation studies of these results have had mixed results [34, 35, 37].

Chalise et al. (2012) [34] showed that Korn's first condition, while mathematically true, is not necessarily going to result in equivalent estimates because the baseline hazard does not contribute to the estimation process of the coefficients. They further note that the time scale is important in the Cox model because it specifies the order of the observed events and determines the risk sets at each event time. In general, the risk sets will be different between the age and time-on-study time scales. To observe how the estimation process between these two models differ, let's consider the partial likelihoods for each of the Cox models. The partial likelihood for equation 2.2 is

$$PL_A(\beta) = \prod_{j=1}^{n} \left(\frac{e^{\beta_A x_j}}{\sum_{i \in R_{jA}} e^{\beta x_j}} \right)^{\delta_j}. \tag{2.4}$$

The partial likelihood for the equation 2.3 is

$$PL_T(\beta) = \prod_{j=1}^{n} \left(\frac{e^{\beta_T x_j + \gamma a_{0j}}}{\sum_{i \in R_{jT}} e^{\beta x_j + \gamma a_{0j}}} \right)^{\delta_j}. \tag{2.5}$$

To estimate β_A and β_T, we maximize these partial likelihoods. These coefficient estimates will be the same when the 2.4 is equal to 2.5. This occurs when a_0 cancels out of the numerator and denominator in equation [34]. That is, the partial likelihoods will be the same when a_0 is constant. Chalise et al. (2012) [34] shows this result in simulation.

2.4.2 Time Scale Performance in Simulations

There have been three groups to assess the assumptions of Korn [36] and the performance of the models under the two time scales through simulation. There are two models that are considered in each of the simulation studies, with additional variations considered in a few of the papers. We will focus on the results for the two models in each paper as well as the results for an additional model considered by two of the groups. Consider the models:

M1) age time scale model

$$\lambda_A(a \mid x) = \lambda_{0A}(a) e^{\beta x}$$

M2) age time scale model with left truncation adjustment

$$\lambda_A(a \mid x, a_0) = \lambda_{0A}(a \mid a_0) e^{\beta x}.$$

M3) time-on-study time scale model with covariate adjustment for baseline age,

$$\lambda_T(t \mid x, a_0) = \lambda_{0T}(t)e^{\beta x + \gamma a_0}$$

Thiebaut and Benichou (2004) [35] conducted the first large scale simulation study to assess Korn's hypotheses. They consider two main goals in their simulations: 1) to investigate Korn's condition that independence between the covariate of interest and baseline age would result in similar estimates across the time scales, and 2) to assess the amount of bias for different degrees of association between covariates and baseline age. They only assess our models M1 and M3 under the assumption that age was the correct time scale. They conclude that independence between baseline age and the covariate of interest is not a condition for equality between the models, as their results show bias in the time-on-study models with adjustment for baseline age. However, the simulation results in their Table 1 suggest that bias is in third decimal place and only exists for hazard ratios of 5, 10, and 50, which are large hazard ratios to observe in practice. As for their second goal, they similarly note that bias occurs when the covariate of interest and baseline age are moderately and highly correlated, as well as for a time-varying covariate. Again, they present results for a hazard ratio of 5 and "significant" biases ranging from 0.008 to 0.051 on M3. They conclude with a general recommendation for the use of the age time scale for the analysis of epidemiological cohort studies.

Pencina et al. (2007) [37] similarly consider the choice of time scale and Korn et al's conditions for equivalence between parameter estimates. In addition to the two models studied by Thiebaut and Benichou (2004) [37] (M1 and M3), Pencina et al. consider the age time scale model with adjustment for baseline age in the baseline hazard (M2 above). They investigate the bias between the models for varying correlations between baseline age and the covariate of interest under both the age time scale being correct and the time-on-study time scale being correct. For all levels of correlation and both time scales, their simulations show that M3 has smaller empirical bias than do the age time scale models (M1 and M2). They conclude that M2 is generally close to M3, while M1 consistently estimates different quantities.

Pencina et al. also assesses Korn's condition of independence between baseline age and the covariate of interest under both time scales. Contrary to Thiebaut and Benichou's [35] conclusion, Pencina et al's simulation studies show equal bias in all three models when there was no correlation between baseline age and the covariate of interest. They state that while these models may estimate the same quantities in this setting, it is unlikely for baseline age to be unrelated to covariates. They conclude that it is best to consider which time scale is most appropriate for the data, and that adjusting baseline age is necessary, regardless of the time scale that best fits the data.

Chalise et al. [34, 38] is the most recent group to consider the above models. They reference the Korn, Thiebaut, and Pencina papers, as well as the contradictions among them as the motivation for their work. Their first paper [34] derives the partial likelihoods of M1 and M3 to examine when the models estimate different quantities. They argue why the assertions from Korn et al. may be correct mathematically, but do not hold in practice. Their simulations in the paper focus on how M1, M2, and M3 will be the same when the variation in baseline age is 0. They then investigate the performance of the three models as the variation of baseline age increases. When time-on-study is the correct time scale, they conclude that adjustment for baseline age in the time-on-study model allows M3 to account for the variation in baseline age and produce more robust results than models M1 and M2. When age is the correct time scale, they further conclude that M3 estimates are reasonably close to the estimates from M1 and M2.

In their second paper, Chalise et al. [38] assess the predictive ability of models M1, M2, and M3 on the two time scales according to their concordance indices. Across all settings considered and regardless of time scale, M3 has the best predictive ability. They further note that this holds true with censoring rates from ranging from 10% to 90%. They conclude that M3 has the best predictive power in their simulations. Between their two papers, they conclude that M3 appears to be the most robust of the models with respect to misspecification of the time scale.

In general, we note that Chalise has consistent results across a few papers and that their methods build on partial likelihood based arguments. However, we find that there are holes in the literature. For example, in late-life disease, such as dementia, is either time scale truly correct? The risk of dementia for most people is nearly 0 until some point in mid-life. After that, risk may increase as a function of age. Further, subjects are not enrolled in studies with knowledge of the age at which their risk started to change. Also, none of these studies assess the bias of the models when a time-varying covariate is of interest and the time scale is misspecified. Thiebaut and Benichou consider a time-varying covariate, but from the standpoint of correlation with baseline age. Our work in Chapter 5 assesses the bias of the M1, M2, and M3 when neither time scale is correct under various settings of correlation between the covariate of interest and time, as well as when there is a time-varying covariate.

Chapter 3

Accounting for System Migration Bias in EHR-based Time-to-Event Studies

3.1 Introduction

Electronic health records (EHRs) have grown in popularity for health-related research over the past decade. They offer a low-cost way to access large amounts of longitudinal data with many potential health outcomes [11]. Establishing causal relationships using EHR data is, however, subject to several limitations including: confounding bias, selection bias, informed-presence bias, and misclassification bias. Another potential, yet largely unstudied, bias arises when a patient's medical history is not fully captured by the EHR system, resulting in potentially informative missing data [39]. Hagar et al. (2014) [22] note that missing EHR data could occur for a variety of reasons including a move or change in insurance, no health concerns, an unrelated condition, mortality, a reflection of the patient's attitude toward healthcare, or seeking care from another facility for the outcome-of-interest. This last case is most concerning as it is likely to result in a violation of the commonly employed non-informative censoring assumption in time-to-event analyses. We refer to patients seeking care from another health system (regardless of the reason) as *system migration* (SM). The occurrence of SM has been particularly hypothesized among American Indian and Alaska Native (AI/AN) patients in the Indian Health Services and Tribal health systems (I/T system).

The Indian Health Services (IHS) provides federally-funded healthcare to qualifying AI/AN patients [13]. In order to study health disparities and health service utilization in AI/AN communities, the *IHS Data Delivery Project* (IHS DDP) created an EHR data set of over 600,000 AI/AN patients who are representative of the overall I/T system population in terms of age and sex [16]. The IHS DDP team has recently focused on studying incidence, prevalence, risk factors, and costs associated with dementia in the I/T population [40, 41, 42, 43]. They note it is difficult to identify incident dementia cases with clinical diagnostic codes (e.g. ICD-9 codes) because the codes and previous diagnoses do not carry forward [40]. Even using a single dementia-related ICD-9 code and a five-year follow-up period, they were only able to identify 88% of prevalent dementia cases among I/T patients 65+ with a dementia-related ICD-9 code in the baseline year of the data (2007). In studying time-to-event outcomes, such as time-to-dementia onset, it is vital to accurately capture diagnoses at the earliest time possible. While Jiang et al. (2021) [40] suggests potential left-censoring as the cause of uncaptured events, it has also been noted that patients at-risk for dementia in the I/T health system have more health care options than just I/T systems. For example, 93.5% of AI/AN patients aged 65+ had Medicare coverage in the IHS DDP in fiscal year 2010 [16]. Further, I/T system clinics and hospitals do not all offer the same services, especially for specialty care, like dementia. I/T patients in some areas may have to seek services for certain dementia-related symptoms and conditions from outside the I/T health system [44]. On the data-level, patients who receive care from non-I/T clinics will have unobserved clinic visits that do not appear in the IHS data. This SM results in unknown intermittent missing data at the subject-level. Thus, the I/T system data can be characterized by two different groups of system-usage: 1) patients who exclusively utilize the I/T system and 2) patients who migrate across systems. The result is interval-censoring that occurs heterogeneously across patients, resulting in potentially elongated intervals for patients partaking in SM. Specifically, for patients who receive out-of-system diagnoses, SM results in delayed diagnostic times within the IHS data, assuming they are eventually observed to receive these diagnoses within the I/T system. We hypothesize that SM will impact coefficient estimates in time-to-event analyses if such migration is differential across covariate-defined subpopulations.

Figure 3.1 illustrates two settings in which system migration may arise in I/T system data. In setting I, the patient receives a dementia diagnosis upon returning to the I/T system, resulting in interval-censoring on a clinic visit time. In setting II, the patient migrates out-of-system without experiencing a diagnosis after (or during) the migration. While this setting may be of concern if patients who migrate out-of-system are systematically different from patients who remain in-system, in the current chapter we focus only on the first setting.

Ideally, patients would have standardized lengths of intervals between visits. If a patient migrated out-of-system in this setting, identifying SM would be trivial, as would be imputing an expected clinic visit time. Even under random interval-censoring with short intervals ($<$ 2 years), Law and Brookmeyer (1992) [45] show imputation methods can result in valid estimates in the time-to-event setting. Specifically, they proposed imputing the midpoints of intervals, which assumes that events occur uniformly throughout the intervals. Under SM, this assumption seems inappropriate because patients receiving a diagnosis out-of-system may be more likely stay out-of-system for follow-up and treatment, resulting in a longer duration of SM than patients who are not diagnosed out-of-system.

In more general interval-censoring settings, Pan (2000) [46] proposed using multiple imputation algorithms to create right-censored data. Pan's method imputes “exact” event times for all subjects with a diagnosis.

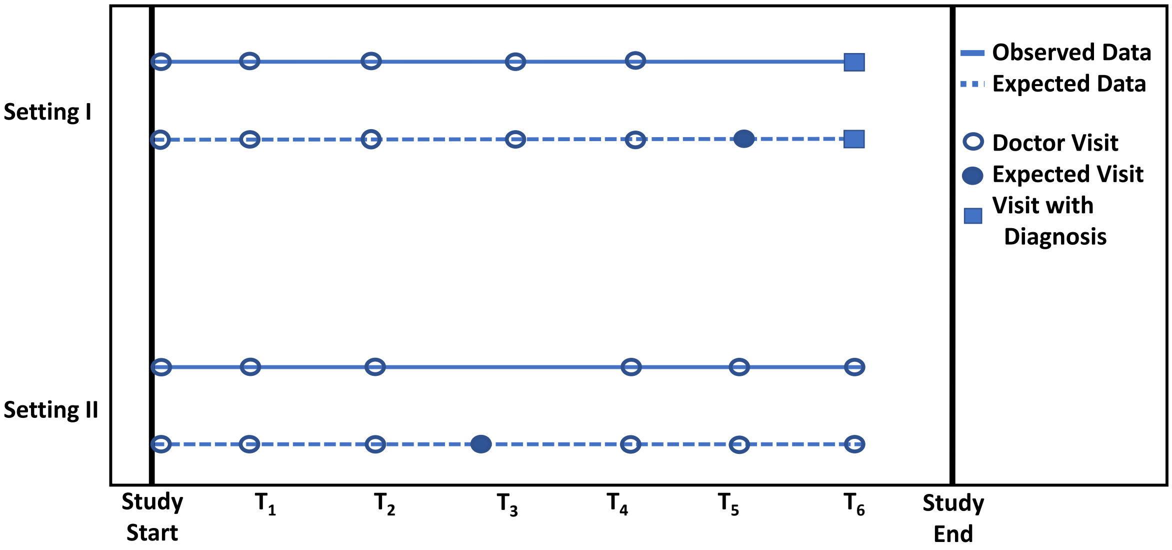


Figure 3.1: Examples of system migration in the IHS data. In setting I, the patient has an extended gap and is diagnosed with dementia on their first visit post-gap. In this setting it is unclear from the observed EHR data alone if the patient was unobserved because they were healthy, or if they were unobserved due to SM. In setting II, the patient has an extended gap between two adjacent visits without a diagnosis. It is again unclear if they were unobserved due to SM or simply not needing health care.

In this chapter, we propose a multiple imputation technique to impute return-to-clinic times similar to Pan (2000) [46], but only for subjects who are predicted to be migrating out-of-system based upon observable characteristics. In section 3.2, we propose the Migration-Adjusted Cox model: a two-step method that 1) estimates a probability of system migration for each patient and 2) uses multiple imputation to set earlier diagnosis times for patients identified as potentially migrating out-of-system. In section 3.3 we explore the performance of the Migration-Adjusted Cox model in simulation studies under two common missingness mechanisms. Section 3.4 contains an application of the Migration-Adjusted Cox model to data simulated based on the IHS EHR data, illustrating potential impacts of failing to adjust for SM in these data. Finally, we end with a discussion of the proposed methods, overall utility and limitations of the approach, and avenues of future research in section 3.5.

3.2 Methods

Due to SM status being unknown for each patient, we propose the Migration Adjusted Cox (MA-Cox) model, a two-step process of adjusting for system migration within the Cox model. In the first step, we predict the propensity for SM for each patient using observed data from their history of clinic visits. Any gap between two subsequent visits that is uncharacteristically long for the given individual conditional on covariates can be identified as a potential period of system migration. In the second step, for patients identified with periods of potential system migration, we impute a return-to-clinic time that is more normative for that individual based upon the system usage history profile of patients with similar characteristics. We provide the details of the two steps in the next two subsections. The section concludes with overall specification of the MA-Cox estimation algorithm. Asymptotic arguments for the proposed estimation procedure are provided where appropriate.

3.2.1 Step One: Estimating the Probability of System Migration

Suppose individual i has K_i observed clinic visits at times $X^V_{i0} < X^V_{i1} < ... < X^V_{iK_i}$ where $X^V_{iK_i} = min(C_i, T^V_i + X^V_{iK_i-1})$; where T^V_i is the the true time between visits, $i = 1, \dots, n$. That is, the final observed visit time for subject i is the minimum of their censoring time and the true time of their next system visit. Let $\delta^V_{ik} = I(t = X^V_{ik})$ be an indicator equal to 1 if the patient has a system visit at time X^V_{ik} and 0 otherwise, $k = 1, \dots, K_i$. In this setting, $\delta^V_{ik} = 1$ for all $k < K_i$. Let Z^V_i denote the covariate vector to be used to model the inter-visit time for patient i.

To estimate the probability of SM, we begin by creating a prediction model for time-between-visits of the following form:

$$\hat{\gamma}_i(t \mid Z^V_i) = \gamma_0(t^V - X^V_{ik}) e^{Z^V_i \hat{\theta}}, \tag{3.1}$$

where $\gamma_0(.)$ denotes the baseline hazard function, Z^V_i can be selected via any model selection procedure, and $\hat{\theta}$ is obtained by maximizing the partial likelihood score function. In sections 3.3 and 3.4, we employ forward-backward stepwise model selection with AIC as the covariate selection criterion.

From the model in (3.1), we estimate the individual level return-time distribution, $S^V_i(t^V) = P(t^V > X_{ik} - X_{ik-1})$, with $\hat{S}_i(t^V) = (\hat{S}^V_0(t^V))^{exp\{Z^V_i \hat{\theta}\}}$, where $\hat{S}^V_0(t^V)$ is the baseline return-time function. $\hat{S}^V_0(t^V)$ can be estimated by exponentiating the baseline hazard of the return-time function so that $\hat{S}^V_i(t^V) = exp\{\hat{\gamma}_0(t^V - X^V_{ik}) exp\{Z^V_i \hat{\theta}\}\}$, with $\hat{\gamma}_0(t^V - X^V_{ik})$ the baseline hazard estimated via the Breslow estimator [33]:

$$\hat{\gamma}_0(t^V - X^V_{ik}) = \sum_{j: X^V_{ik} \leq t^V + X^V_{ik-1}} \frac{1}{\sum_{i=1}^{n} \sum_{k=1}^{K_i} exp(Z^V \hat{\theta}) \delta^V_{ik}}. \tag{3.2}$$

Finally, we identify the estimated probability of SM for patient i as $\hat{p}^{SM}_i = \hat{S}^V_i(X^V_{iK} - X^V_{iK-1})$.

3.2.2 Step Two: Map System Migration Diagnostic Times to Present System

Suppose individual i has an observed disease event time, X_i^D, where $X_i^D = min(C_i, T_i^D)$ with C_i denoting the censoring time and T_i^D denoting the disease diagnosis time. Let $\delta_i^D = I(t = T_i^D) = I(X_i^D = T_i^D)$ be an indicator equal to 1 if the patient has the diagnosis of interest, and 0 otherwise. Let Z_i^D be the specified set of covariates associated with the time-to-diagnosis for subject i. To accurately estimate the hazard ratios associated with covariates Z^D, we need to account for individuals who migrate out of system for their diagnosis of the event of interest. Here, we use multiple imputation to map the diagnostic times of potential SM cases back to "normal" system visitation patterns. Within each of the $l = 1, ..., L$ imputation data sets, we will impute diagnosis times for potential cases of SM and then model the hazard of experiencing the event as $\tilde{\lambda}_i^D(t|Z_i^D) = \tilde{\lambda}_0^D(t)e^{Z_i^D\beta^l}$, where $\tilde{\lambda}_i^D$ and $\tilde{\lambda}_0^D$ are the hazard function and baseline hazard function for for the l^{th} imputation data set. Going forward, we will represent diagnostic time variables for the l^{th} data set with a tilde. That is, we define $\tilde{X}_i^D = X_i^{D;l}$ is the observed time for subject i in the l^{th} imputation data set, $\tilde{t}^D = t^{D;l}$ represents the event times in the l^{th}, and $\tilde{Y}_i^D = Y_i^{D,l} = I(\tilde{X}_i^D \geq \tilde{t}^D)$ is the at-risk indicator for subject i in imputation set l.

We identify potential cases of SM as patients who have abnormally long gaps between their visit with a diagnosis and the prior visit using the approach outlined in section 3.2.1. However, not everyone with an abnormally long gap is necessarily seeking services from an external system. To account for this, within each imputation data set, we draw $M_i^l \sim$ Bernoulli($\hat{p}_i^{SM}$)) as an indicator for whether patient i is an "actual" SM case, where $M_i^l = 1$ represents SM case, and $M_i^l = 0$ represents no SM for that patient. Next, among the patients with $M_i^l = 1$ we only impute diagnosis times for patients with diagnoses (i.e. $\delta_i^D = 1$). Note that our diagnosis times are draws from the estimated return-time distribution, $\hat{S}_i^V(t^V \mid Z_i^V)$, added to their penultimate visit time, $X_{iK_i-1}^V$. We sample return times via the inverse probability integral transform. That is, for subject i such that $M_i^l = 1$ and $\delta_i^D = 1$, we sample U_i from Uniform$(0,1)$ and set the imputation return-time, $\tilde{X}_i$, as the return-time from the inverse return-time distribution plus the time of the penultimate observed visit. That is, $X_i^l = (\hat{S}_i^V)^{-1}(U_i \mid Z_i^V) + X_{iK-1}^V$, where $(\hat{S}_i^V)^{-1}(U_i \mid Z_i^V)$ is the return-time from the inverse return time distribution. Now, we only adjust the observed diagnosis time if the imputed time is earlier than the observed time, $\tilde{X}_i^D = min(T_i^D, X_i^l)$. For the rest of the patients in the data set (i.e. for patients with $M_i^l = 0$ or $\delta_i^D = 0$), we do not change their observed times: $\tilde{X}_i^D = X_i^D$.

To estimate $\hat{\beta}$, we fit a Cox PH model with each imputation data set to obtain imputation-specific estimates of β^l. That is, solve:

$$U^l(\beta^l) = \sum_{i=1}^{n} \int_{\tilde{t}^D=0}^{\infty} \left(Z_i^D - \frac{n^{-1}\sum\limits_{q=1}^{n} \tilde{Y}_q^D(\tilde{t}^D) Z_q^D exp\{Z_q^D \beta^l\}}{n^{-1}\sum\limits_{q=1}^{n} \tilde{Y}_q^D(\tilde{t}^D) exp\{Z_q^D \beta^l\}} \right) dN_i^l(\tilde{t}^D) \overset{\triangle}{=} 0, \qquad (3.3)$$

where $N_i^l(\cdot)$ is the counting process for events, and $\tilde{Y}_i^D(\tilde{t}^D) = I(\tilde{X}_i^D \geq \tilde{t}^D)$ is the at-risk indicator for individual i, both for imputation data set l. Using [47] for multiple imputation, $\hat{\beta} = \frac{1}{L}\sum\limits_{l=1}^{L} \hat{\beta}^l$ and $\hat{V}(\hat{\beta}) = \frac{1}{L}\sum\limits_{l=1}^{L} \hat{V}_{SI}(\hat{\beta}^l) + \frac{L+1}{L}\left(\frac{\sum_{l=1}^{L}(\hat{\beta}^l - \hat{\beta})^2}{L-1}\right)$, where $\hat{V}_{SI}(\hat{\beta}^l)$ is the standard variance estimator for a Cox model. It can be demonstrated that $\hat{\beta}$ is consistent for β and asymptotically normal, and that $\hat{V}(\hat{\beta})$ is consistent for the desired quantity. Details in section 3.2.3.

Algorithm 1 Algorithm for the MA-Cox model estimating procedure.

```
Identify subset of covariates, $\mathbf{Z}^V$, which minimizes honest assessment of out-of-sample prediction error of time-to-next visit with the model $\hat{\gamma}_i(t \mid Z^V) = \gamma_0(t^V - X_{ik}^V)e^{Z_i^V\hat{\theta}}$.
Compute $\hat{\gamma}_0(t^V - X_{ik}^V) = \sum_{i:X_{ik}^V \leq t^V + X_{ik-1}^V} \frac{1}{\sum_{i=1}^n \sum_{k=1}^{K_i} exp(Z^V\hat{\theta})\delta_{ik}^V}$ [33] corresponding to $\hat{\gamma}_i(t \mid Z^V)$
Calculate $\hat{S}_i^V(t^V) = (\hat{S}_0^V(t^V))^{e^{Z_i^V\hat{\theta}}} = (e^{\hat{\gamma}_0(t^V - X_{ik}^V)})^{e^{Z_i^V\hat{\theta}}}$
Estimate $p_i^{SM}$ by $\hat{p}_i^{SM} = \hat{S}_i^V(X_{iK}^V - X_{iK-1}^V)$
for each of the $l$ imputation data set do
  Draw $M_i \sim$ Bernoulli($\hat{p}_i^{SM}$) as an indicator for whether subject $i$ was interval-censored ($M_i = 1$), or not ($M_i = 0$);
  if $M_i = 1$ and $\delta_i^D = 1$ then
    Draw $U_i \sim$ Uniform$(0, 1)$;
    Set $X_i^l = (\hat{S}_i^V)^{-1}(U_i \mid Z_i^D) + X_{iK-1}^V$;
    Set $\tilde{X}_i^D = min(T_i^D, X_i^l)$;
  end if
  else $\tilde{X}_i^D = X_i^D$
  To estimate $\hat{\beta}^{(l)}$ on the imputed data set, solve:
      $$U(\beta^l) = \sum_{i=1}^{n} \int\limits_{t^D=0}^{\infty} \left( Z_i^D - \frac{n^{-1}\sum_{q:\tilde{X}_q^D \geq \tilde{X}_i^D} Z_q^D exp\{Z_q^D\beta^l\}}{n^{-1}\sum_{q:\tilde{X}_q^D \geq \tilde{X}_i^D} exp\{Z_q^D\beta^l\}} \right) dN_i^l(\tilde{t}^D) \triangleq 0.$$
    Estimate $\hat{V}_{SI}(\hat{\beta}_l)$ with
        $$\hat{V}_{SI}(\hat{\beta}_l) = n^{-1}\sum_{i=1}^{n}\int_0^{\tilde{t}^D}\left(Z_i^D - \bar{Z}(s,\hat{\beta}^l)\right)^{\otimes 2}\tilde{Y}_i^D(s)e^{Z_i^D\hat{\beta}^l}\lambda_0(s)ds$$
    where $\bar{Z}(s,\hat{\beta}^l) = \frac{n^{-1}\sum_{q:\tilde{X}_q^D\geq\tilde{X}_i^D} Z_q^D exp\{Z_q^D\hat{\beta}^l\}}{n^{-1}\sum_{q:\tilde{X}_q^D\geq\tilde{X}_i^D} exp\{Z_q^D\hat{\beta}^l\}}$
end for
Set $\hat{\beta} = \frac{1}{L}\sum_{l=1}^{L}\hat{\beta}^l$ and $\hat{V}(\hat{\beta}) = \frac{1}{L}\sum_{l=1}^{L}\hat{V}_{SI}(\hat{\beta}^l) + \frac{L+1}{L}\left(\frac{\sum_{l=1}^{L}(\hat{\beta}^l-\hat{\beta})^2}{L-1}\right)$
```

3.2.3 Asymptotic Distribution

To derive the asymptotic distribution of $\hat{\beta}$, we need to consider two pieces: 1) the asymptotic distribution of the estimator for the l^{th} imputation data set, and 2) the distribution of the multiple imputation corrected estimators of the mean and variances of $\hat{\beta}$. For (1), we follow arguments similar to those by Lu and Tsiatis (2001) [48] to establish asymptotic properties of the $\hat{\beta}^{(l)}$. For (2), along with some technical conditions, if the imputation method is proper for the complete-case data, then the results converge to a normal distribution [47]. We assume the prediction model for return-to-clinic time is correctly specified, which implies that the imputed times will not impact the convergence of the estimating equations.

For multiple imputation to maintain convergence properties and valid inference, three conditions must hold: 1) the imputation method must be proper for the complete-case distribution, 2) the complete-data analysis must be valid, and 3) the sample size much be large enough to appeal to asymptotic arguments. A proper imputation method satisfies three conditions: 1) the imputation method should effectively be drawing random samples from the complete-data distribution, 2) the within-imputation variance estimates are centered at the corresponding complete-data statistics, and 3) the between-imputation variation is controlled. Under the assumption that the prediction model is correctly specified, the MA-Cox model randomly selects a return-to-clinic time from the complete-case distribution. Since the true diagnostic times can only fall in a finite range, $(X_{iK_{i-1}}, X_{iK_i}]$, the complete-case distribution need only contain dense enough points within that range for each person. The correctly-specified return-to-clinic prediction model will provide dense enough times to allow for proper imputation. This also implies that the variance estimates from the Cox model within each imputation set is approximately centered at the complete-data variance estimate. Since our imputation method results in imputations based on the complete-case data, the resulting statistics will be centered around the complete-case statistics. Third, we assume the variation between imputation data sets is finite. With a correctly-specified return-time prediction model, each imputation data set will result in proper estimates of the coefficients. Thus, the between imputation set variation will be controlled. Therefore, the MA-Cox model's imputation step is proper for the complete-case distribution. Further, the Cox model is a valid method in the complete-data analysis. The MA-Cox model is also valid because if there is no SM, then the estimated probabilities of SM will be small, and the method will not impute new times for patients. These results hold in simulation settings too where no SM was included. Finally, we assume sample sizes are large enough for asymptotic arguments. This implies that the multiple imputation within the MA-Cox model is proper and the multiple imputation corrections will maintain the properties that occur within each imputation data set.

3.2.4 Asymptotic Distribution Derivation

We define the filtration at time t for individual i, $\mathcal{F}_i(t)$, to be the history of events, visit times, and covariates up to (and including) time t. That is, $\mathcal{F}_i(t^D) = \sigma\{N_i^D(s), Y_i^D(s^+), Z_i^D; s \in (0, t^D]\}$. It is important to note that 1) the imputation model only considers patients who have a diagnosis as at-risk for SM, and 2) the imputation method only changes observed times, t^D for a subset of patients in sample. Our proof follows closely to the work of Lu and Tsiatis (2001) [48].

We consider the contribution from the l^{th} imputed data set. The counting process for the l^{th} imputation data set is defined as:

$$N_i^l(\tilde{t}^D) = I(\tilde{X}_i^D \leq \tilde{t}^D).$$

Recall that superscript D refers to variables for the diagnosis time, X_i is the observation time for subject i, and $I(\cdot)$ is the indicator function. The coefficients, $\hat{\beta}^l$ are estimated by solving the score function set equal to 0:

$$U^l(\beta^l) = \sum_{i=1}^{n} \int\limits_{\tilde{t}^D=0}^{\infty} \left(Z_i^D - \frac{n^{-1}\sum\limits_{q=1}^{n} \tilde{Y}_q^D(\tilde{t}^D) Z_q^D exp\{Z_q^D \beta^l\}}{n^{-1}\sum\limits_{q=1}^{n} \tilde{Y}_q^D(\tilde{t}^D) exp\{Z_q^D \beta^l\}} \right) dN_i^l(\tilde{t}^D) \triangleq 0. \tag{3.4}$$

Note that we can rewrite $U^l(\beta^l)$ by replacing $dN_i^l(\tilde{t}^D)$ with the martingale increment $dM_i^l(\tilde{t}^D)$ where

$$dM_i^l(\tilde{t}^D) = dN_i^l(\tilde{t}^D) - \lambda_0(\tilde{t}^D) e^{Z_i^D \beta^l} \tilde{Y}_i^D(\tilde{t}^D) d\tilde{t}^D.$$

The concave function argument in Andersen and Gill (1982) [49] can be used to establish that $\hat{\beta}^l \xrightarrow{P} \beta_0$. Thus, by the weak law of large numbers, it can be shown that

$$\frac{n^{-1}\sum\limits_{q=1}^{n} \tilde{Y}_q^D(\tilde{t}^D) Z_q^D exp\{Z_q^D \beta_0\}}{n^{-1}\sum\limits_{q=1}^{n} \tilde{Y}_q^D(\tilde{t}^D) exp\{Z_q^D \beta_0\}} \rightarrow \mu(t, \beta_0).$$

Now that this term converges to some value, using similar arguments to those in Tsiatis (1981) [50], we have

$$n^{-1/2} \sum_{i=1}^{n} \int \left(\frac{n^{-1}\sum\limits_{q=1}^{n} \tilde{Y}_q^D(\tilde{t}^D) Z_q^D exp\{Z_q^D \beta_0\}}{n^{-1}\sum\limits_{q=1}^{n} \tilde{Y}_q^D(\tilde{t}^D) exp\{Z_q^D \beta_0\}} - \mu(t, \beta_0) \right) dM_i^l(\tilde{t}^D) \xrightarrow{P} 0,$$

which means

$$n^{-1/2}U^l(\beta_0) = n^{-1/2}\sum_{i=1}^{n}\phi_i^l(\beta_0) + o_p(1),$$

where $\phi_i^l(\beta_0) = \int(Z_i^D - \mu(t,\beta_0))dM_i^l(\tilde{t}^D)$.

Since $\phi_i^l(\beta_0)$ is a martingale increment, it is mean zero. Note then that this is a normalized sum of independent and identically distributed random variables with mean zero. Thus, by the central limit theorem, $U^l(\beta_0)$ is asymptotically normal with with asymptotic variance of

$$V_{SI} = \int \mathrm{E}\left[(Z_i^D - \mu(s,\beta_0))^{\otimes 2} e^{Z_i^D\beta_0} Y_i(s)\right]\lambda_0(s)ds.$$

The multiple imputation estimator, $\hat{\beta}$, is a mean of consistent estimators from the single imputation data sets, thus $\hat{\beta}$ is consistent for the same quantity, β_0. Further, $n^{1/2}(\hat{\beta} - \beta_0) = V_{SI}^{-1}n^{-1/2}U(\beta_0) + o_p(1)$, where $n^{-1/2}U(\beta_0)$ is asymptotically equivalent to

$$n^{-1/2}\sum_{i=1}^{n}\left(L^{-1}\sum_{l=1}^{L}\phi_i^l(\beta_0)\right).$$

Again, this is a normalized sum of independent and identically distributed random variables with mean zero. Thus, we have asymptotic normality by the central limit theorem. We estimate the asymptotic variance with the typical estimator

$$\hat{V}(\hat{\beta}) = \frac{1}{L}\sum_{l=1}^{L}\hat{V}_{SI}(\hat{\beta}^l) + \frac{L+1}{L}\left(\frac{\sum_{l=1}^{L}(\hat{\beta}^l - \hat{\beta})^2}{L-1}\right).$$

3.3 Simulations

We assess the MA-Cox model via simulation with varying degrees of SM. We consider two scenarios of missingness in the "missing at random" (MAR) category. *MAR1* refers to system migration that depends on covariates indirectly through the time-to-diagnosis; and *MAR2* refers to system migration that depends on covariates directly (matching the traditional sense of MAR data). Since people with shorter times-to-event are less likely to migrate out of system missing completely at random (MCAR) data would refer to the setting where the time-to-diagnosis and system migration are both completely independent of all covariates. We found this to be unlikely in pracitce and do not consider it further.

3.3.1 General Simulation Settings

We consider a correctly specified Cox proportional hazards model depending on three predictors: one continuous ($Z_1 \sim \text{Normal}(0,1)$) and two binary ($Z_2 \sim \text{Bernoulli}(0.5)$ and $Z_3 \sim \text{Bernoulli}(0.15)$). All results focus on the coefficient estimates for the Z_2 (results for the other predictors are similar). Event times, T_i, are sampled from a exponential distribution with rate $\lambda(t) = \lambda_0 e^{\beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_3}$ where $\lambda_0 = 0.1$, $\beta_1 = \log(1.8)$, $\beta_2 = \log(1.2)$, and $\beta_3 = \log(1.6)$. We simulate data with samples sizes of 250 and 1000 subjects. Each simulation setting is simulated 1000 times.

After we sample the dementia diagnosis times, right censoring times, C_i, are drawn from uniform distributions with limits specified to achieve the desired censoring rates. We define $X_i^D = \min(T_i^D, C_i)$, where X_i^D is the observed time for subject i. For all subjects, we simulate system visit times, X_{ik}^V from $\text{Uniform}(0.1, 2.01)$, so that each subject has an average of a one year between routine visits. To simulate SM, we simulate a probability of SM as $p_i^{SM} = \text{expit}(\cdot)$, where $\text{expit}(\cdot)$ is the expit function and the input is defined under the appropriate missingness mechanism (MAR1 or MAR2). Migration status is simulated via $M_i \sim \text{Bernoulli}(p_i^{SM})$ to identify which patients migrate out-of-system. We then simulate times of migrating out-of-system as $t_i^{\text{leave}} \sim \text{Exponential}(\text{rate} = \lambda_i^{\text{leave}})$ and the length of time out-of-system, as $t_i^{\text{length}} \sim \text{Exponential}(\text{rate} = \lambda_i^{\text{length}})$, where λ_i^{leave} and $\lambda_i^{\text{length}}$ are specified under the appropriate missingness mechanism.

For patients with $M_i = 1$, we apply SM by removing all system visits such that $t_i^{\text{leave}} \leq X_{ik}^V \leq t_i^{\text{leave}} + t_i^{\text{length}}$. For the cases where a subject is censored (i.e. $X_{iK_i}^V = X_i^D = C_i$) and $t_i^{\text{leave}} \leq C_i \leq t_i^{\text{leave}} + t_i^{\text{length}}$, then we leave $X_i^D = C_i$ and only remove system visits such that $t_i^{\text{leave}} \leq X_{ik}^V \leq X_i^D$. For the cases where $t_i^{\text{leave}} \leq X_i^D \leq t_i^{\text{leave}} + t_i^{\text{length}}$ and $\delta_i^D = 1$, we adjust the observed event time, X_i^D, to be $t_i^{\text{leave}} + t_i^{\text{length}}$. While this case assumes that patients would be diagnosed upon return to the system, this also represents the setting where SM would have the smallest impact on coefficient estimates. Note that t_i^{leave} and t_i^{length} do not impact the data when $M_i = 0$ or when $X_i^D < t_i^{\text{leave}}$.

3.3.2 MAR1 Simulation Settings

Under the MAR1 setting, we treat SM as only related to the covariates indirectly through the diagnosis time. That is, patients with later diagnosis times have more opportunity to migrate out of system. In this setting, we define the probability of SM as $P(\text{SM}) = \text{expit}(\beta_{\text{SM}})$, where $\beta_{\text{SM}} = \{-2.2, 0, 2.2\}$ for migration probabilities of approximately 0.1, 0.5, and 0.9. The time of leaving the system (t_i^{leave}) and length of system migration (t_i^{length}) are drawn from exponential distributions with rates $\lambda_{\text{leave}} = \log(-\log(0.1)/4)$ and $\lambda_{\text{length}} = \log(-\log(0.75)/2)$.

3.3.3 MAR2 Simulation Settings

In the MAR2 setting, probability for SM depends on the covariates directly. We define $P(\text{SM}) = \text{expit}(Z * \beta_{\text{SM}})$, where Z is the design matrix of covariates (including a column of 1 for the baseline hazard), and $\beta_{\text{SM}} = [b_0, 0, -0.3, 0.8]^T$, with $b_0 = \{-2.2, 0, 2.2\}$ for migration probabilities of approximately 0.1, 0.5, and 0.9. Similarly, the leave time (t_{leave}) and length of the SM (t_{length}) depend on covariate values also. They are drawn from exponential distributions with rates:

$$\lambda_{\text{leave}} = e^{Z\beta_{\text{leave}}}, \text{ where } \beta_{\text{leave}} = [\log(-\log(0.1)/4), \log(0.8), \log(0.4), \log(1.6)]^T$$
$$\lambda_{\text{length}} = e^{Z\beta_{\text{length}}}, \text{ where } \beta_{\text{length}} = [\log(-\log(0.75)/2), \log(0.6), 0, \log(0.4)]^T.$$

3.3.4 MAR1 Simulation Results

Figures 3.2 and 3.3 depict the effect of SM on coefficient estimates when using three correctly specified Cox models. Each figure is broken into three plots representing that cases where 1) the traditional Cox model is fit (top forest plot), 2) a Cox model is fit to data with single imputation of the midpoint of the SM gap for patients whose diagnosis time is displaced due to SM (middle forest plot), and 3) the MA-Cox model is fit to adjust for SM. The vertical lines in each plot represents the true coefficient for Z_2, and the horizontal lines represent the 95% confidence intervals on the mean coefficient estimate with the empirical standard error. The columns to the left represent the right censoring percent (left), the percent of patients with potential SM (percent of patients where SM impacted their diagnosis time among those with dementia diagnoses; middle), and the mean of the hazard ratio estimates over the 1000 simulations (percent bias from the truth; right).

When SM is unaccounted for the resulting estimates show greater than 10% bias toward the null with as little as 17% of the subjects having migration periods that impact diagnosis times for sample sizes of 250 and 1000 (row 1 of Figure 3.2 and Figure 3.3). As the rates of censoring and/or SM increase so too does the bias.

In the second row of plots in Figures 3.2 and 3.3, we attempt to account for the bias by imputing the midpoint of the SM gap for the patients whose diagnostic times change due to SM. The bias is reduced through midpoint imputation, but not completely removed in the most extreme settings. This method also requires knowledge of which patients migrated across systems.

In the third row of plots in Figures 3.2 and 3.3, we use the proposed MA-Cox model to identify potential cases of SM and impute diagnostic times for those patients. We observe that the proposed method reduces bias to less than 10% bias in each setting. With a smaller sample size of 250, MA-Cox model reduces bias to less than 10% in all settings. Further, with higher rates of censoring, the method reduces bias to almost zero. In Figure 3.3, the settings with the most extreme SM exhibit the most bias ($\sim 7-9\%$), regardless of the amount of right censoring. However, these cases also have 90% of patients with the possibility of migrating out-of-system, which averages around 40% of observed diagnostic times being delayed due to SM. The bias from the MA-Cox model is 35% lower in the 35% right censoring setting and is 80% lower in the 70% right censoring setting. We also note that the proposed method does not induce bias in the cases where there is no SM.

MAR2 Simulation Results

The simulation settings for MAR2 are set to bias the estimates away from the null because the MAR1 setting will always bias toward the null. The first row of Figure 3.4, depicts that the Cox model has 10% bias in the setting with the least SM. As censoring and the amount of SM increases, the estimated log-hazard ratios exhibit more bias, reaching over 85% bias in the most extreme setting we consider. Further, even knowing exactly who migrates out-of-system and imputing an earlier event time (midpoint of the SM gap) is not sufficient to completely remove the bias in the estimates. The bias still ranges from 10% to 50% when there is at least 50% of subjects capable of SM. The last row of 3.4 depicts that the MA-Cox model reduces bias drastically in almost all cases. Only the most extreme case has over 15% bias in the same direction as the Cox model that does not account for SM. Further, in terms of raw difference between the estimates and the truth, the MA-Cox model estimates are on

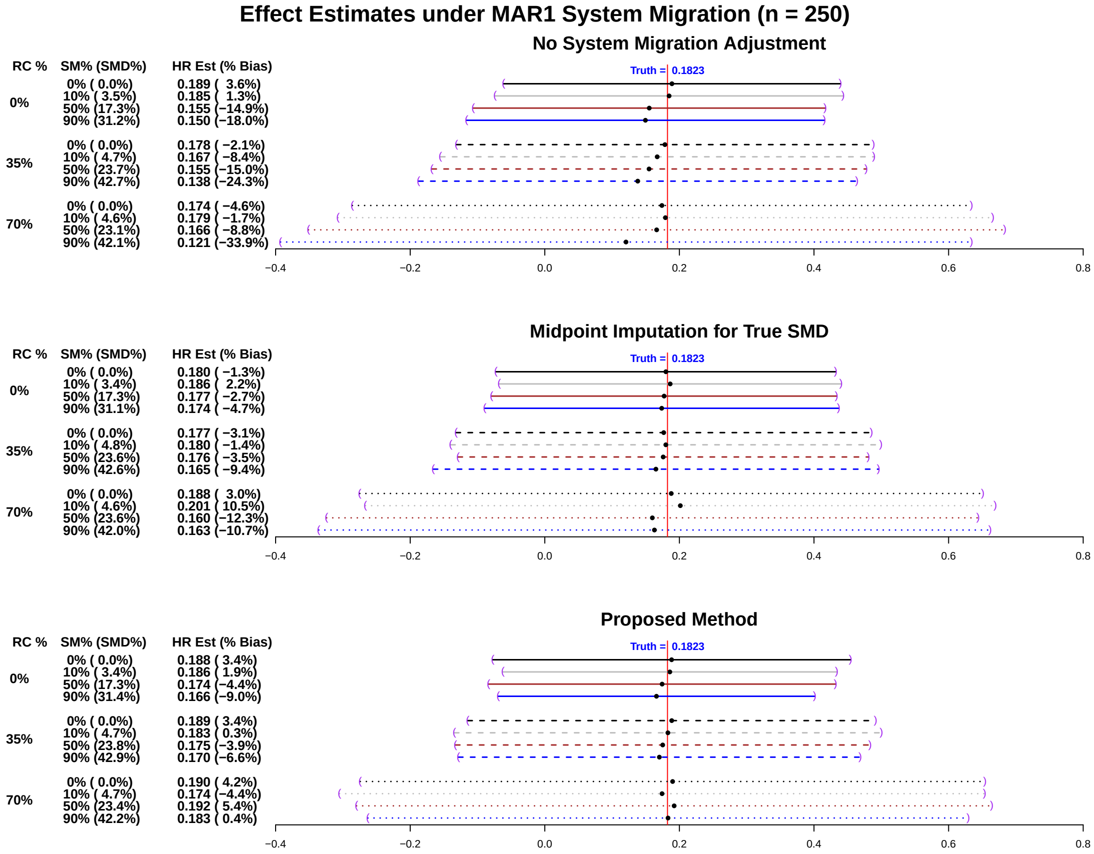


Figure 3.2: Forest plots for Cox model estimates of a binary covariate with 50% successes. The effect size is log(1.2). The MAR1 setting occurs when SM depends on covariates indirectly through the time-to-diagnosis. The columns are defined from left to right: “RC %” refers to right censoring exclusively. The “SM % (SMD %)” refers to the percent of subjects who could have migrated out-of-system (percent of subjects whose migration period impacts their diagnosis time among subjects with a diagnosis). The “HR Est (% Bias)” refer to the mean coefficient estimates (percent bias) from 1000 simulations. The sample size is 250 subjects in each of these simulations.

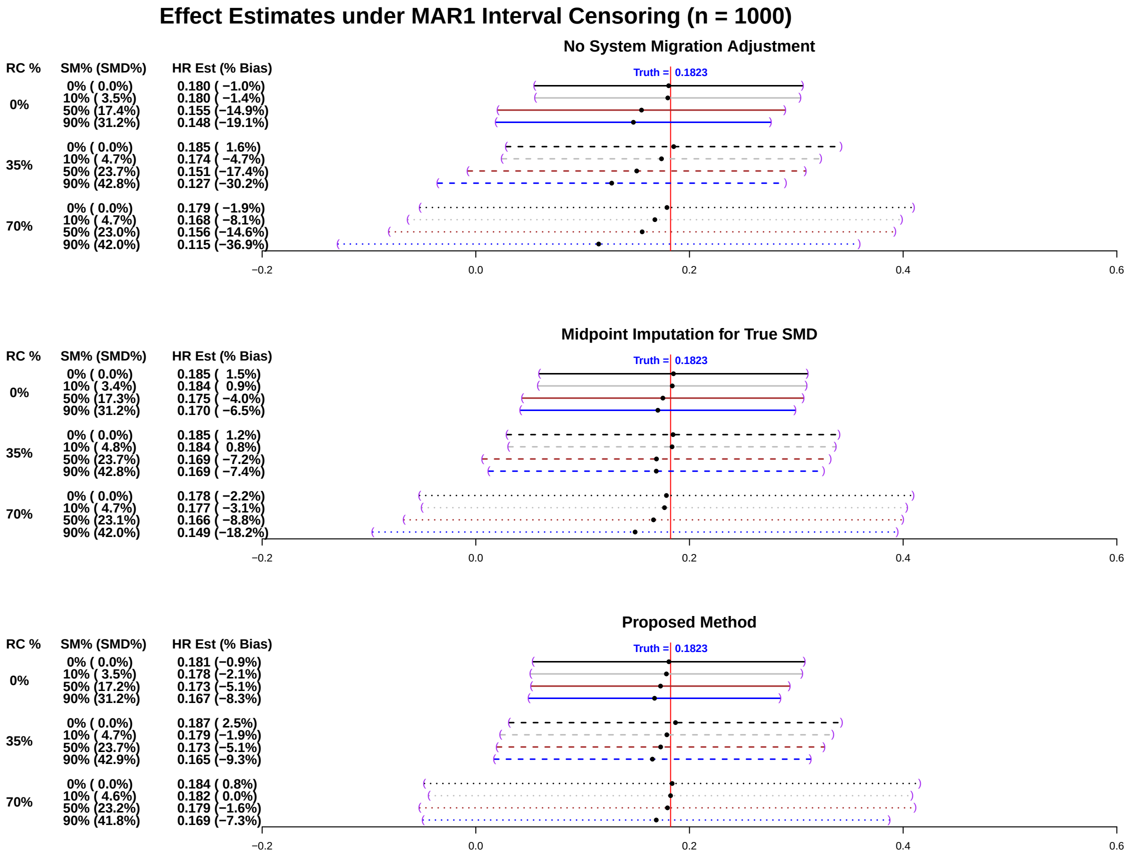


Figure 3.3: Forest plots for Cox model estimates of a binary covariate with 50% successes. The effect size is log(1.2). The MAR1 setting occurs when SM depends on covariates indirectly through the time-to-diagnosis. The columns are defined from left to right: “RC %” refers to right censoring exclusively. The “SM % (SMD %)” refers to the percent of subjects who could have migrated out-of-system (percent of subjects whose migration period impacts their diagnosis time among subjects with a diagnosis). The “HR Est (% Bias)” refer to the mean coefficient estimates (percent bias) from 1000 simulations. The sample size is 1000 subjects in each of these simulations.

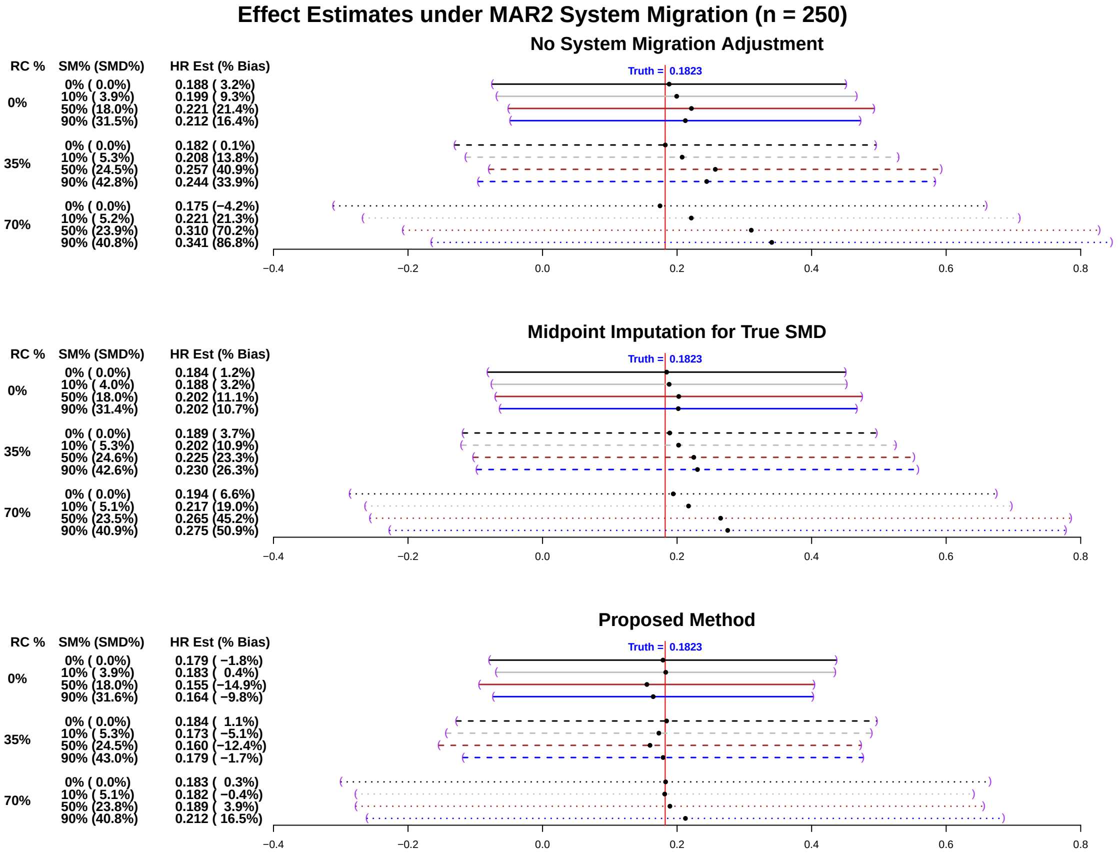


Figure 3.4: Forest plots for Cox model estimates of a binary covariate with 50% successes. The effect size is log(1.2). The MAR2 setting occurs when SM depends on the same covariates as diagnosis time for subjects (MAR2). The columns are defined from left to right: "RC %" refers to right censoring exclusively. The "SM % (SMD %)" refers to the percent of subjects who could have migrated out-of-system (percent of subjects whose migration period impacts their diagnosis time among subjects with a diagnosis). The "HR Est (% Bias)" refer to the mean coefficient estimates (percent bias) from 1000 simulations. The sample size is 250 subjects in each of these simulations.

average within 0.03 of the true coefficient value.

In Figure 3.5, we see similar results as in the case with Figure 3.4. When not accounting for SM, the Cox model shows greater than 15% in all but one setting with SM. In the most extreme settings, the bias is over 30% for 35% right censoring and over 75% for 70% right censoring. Further, row two of Figure 3.5 depicts that knowing exactly whose diagnostic times are delayed due to SM is not sufficient to remove bias with midpoint imputation. In row three of Figure 3.5, the MA-Cox model reduces bias in most settings, especially when there is more censoring. In the 70% right censoring setting, estimates with low and moderate amounts of SM are unbiased. The setting with 70% right censoring and 90% of subjects with possibility of SM shows 19% bias. While still biased, the estimate is 75% lower in bias than Cox model. In the settings with no censoring and 35% censoring the MA-Cox model over-corrects for SM and estimates a coefficient that is 5-15% lower than the truth. On average, the MA-Cox model coefficient estimates within 0.03 of the true value in terms of raw difference. We also note that the proposed method does not induce bias in the cases where there is no SM.

3.4 Application

For privacy concerns related to IHS data, we consider simulated data using previously published statistics from the IHS data and IHS website. We apply the MA-Cox model to data simulated off of what we might see in the Indian Health Services data. Data is simulated from previously observed results from IHS papers [43, 51, 52, 53, 54]. We simulate several variables: baseline age, sex, and indicators of depression, vascular disease, cerebrovascular disease, diabetes, and health coverage status (Medicare, Medicaid, and private) to influence the time of a dementia diagnosis [55, 56, 57, 58]. We simulate additional variables: body mass index (BMI), number of prescriptions, and indicators of amputation, liver disease, chronic kidney disease, hypertension, cancer, cardiovascular disease (CVD), mental disorders, alcohol-use disorder, drug-use disorder, and tobacco-use disorder to influence the times between observations (or IHS system visits). All variables are simulated from distributions observed previously and are simulated as independent of one another due to lack of information on inter-variable relationships. The time of dementia and the time-between-visits are simulated from exponential distributions with rates determined loosely based on previous knowledge of the direction of the relationship. Censoring is simulated uniformly for all pa-

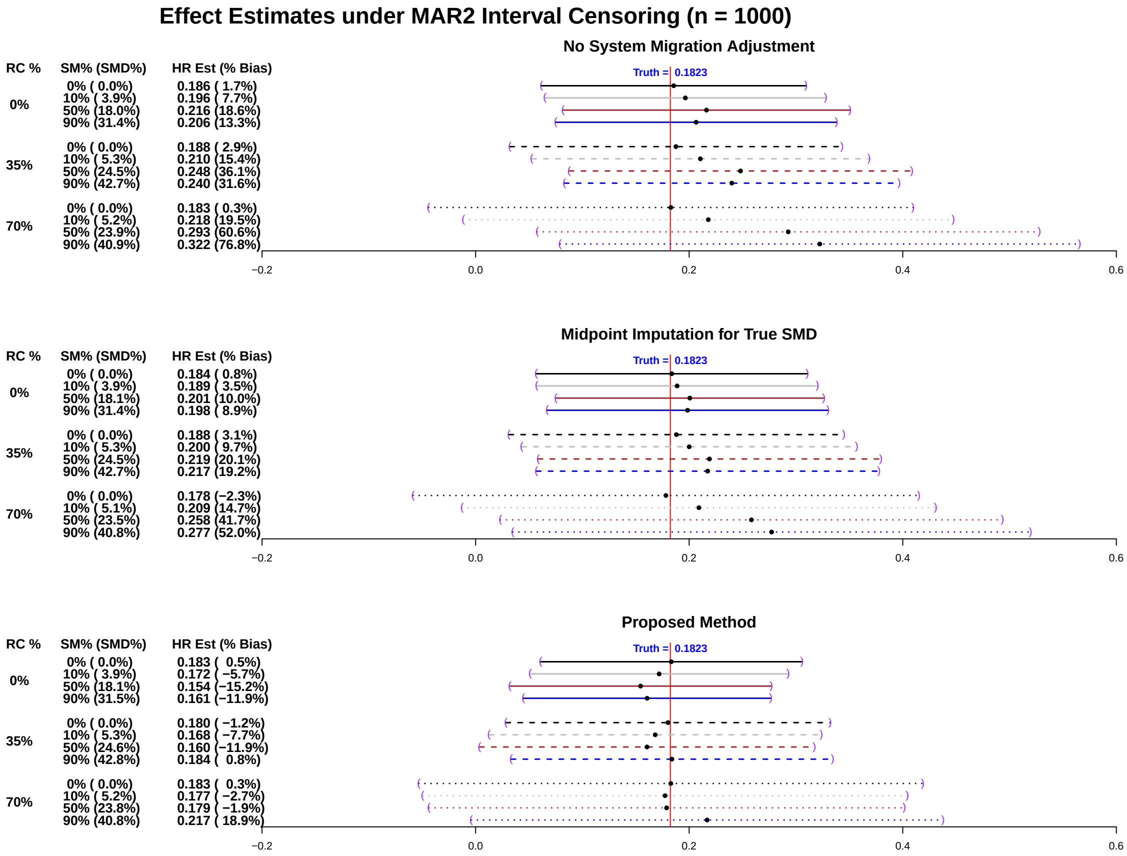


Figure 3.5: Forest plots for Cox model estimates of a binary covariate with 50% successes. The effect size is log(1.2). The MAR2 setting occurs when SM depends on the same covariates as diagnosis time for subjects (MAR2). The columns are defined from left to right: “RC %” refers to right censoring exclusively. The “SM % (SMD %)” refers to the percent of subjects who could have migrated out-of-system (percent of subjects whose migration period impacts their diagnosis time among subjects with a diagnosis). The “HR Est (% Bias)” refer to the mean coefficient estimates (percent bias) from 1000 simulations. The sample size is 1000 subjects in each of these simulations.

tients to achieve $\sim 10\%$ event rate. Each dementia patient is assigned a probability of SM based on covariates possibly related to system usage. The weight of those factors is chosen somewhat arbitrarily to establish $\sim 50\%$ SM among the patients with a dementia diagnosis.

We hypothesize that the coefficient estimates of health coverage status variables would be most impacted by system migration because they are a direct indicator of a patient's ability to use multiple health systems, but they also serve as an important proxy for underlying health conditions (among other things) that may be associated with dementia diagnoses. Despite simulating data for four health coverage groups, we create mutually exclusive groups based on those variables and how we would approach health coverage groups in an actual analysis with the data. Each patient is simulated to have a three indicators with respect to access, one representing each of Medicare coverage, Medicaid coverage, and private coverage. This results in patients possibly belonging to one of eight groups. In an actual analysis, we would likely condense the groups to four mutually exclusive groups: 1) any patient with private coverage, 2) any patient with Medicaid coverage but not private coverage, 3) patients with only Medicare coverage, and 4) patients with no coverage. Given that the outcome of interest is dementia diagnosis, the vast majority of this population should have access to Medicare coverage, thus we combine groups 3 and 4 to create a Medicare and IHS-only group.

3.4.1 Study Population

The Indian Health Service (IHS) is an agency within the department of human and health services which is responsible for providing culturally appropriate federal health services to AI/AN peoples [13]. The IHS primarily provides health services at IHS, Tribal, and Urban Indian health facilities to approximately 2.6 million AI/AN individuals spread across 37 states and 574 federally recognized tribes [59]. For more information on the I/T system data see [16].

The analytical data set consists 3,680 simulated IHS patients aged 65+ years old. There are 454 incident dementia cases among the 3,680 patients. Note that in real IHS data adults without incident dementia includes potential undiagnosed dementia cases. In this simulation, we assume all incident dementia is diagnosed. The true number of patients who migrate out of system is 1814. The number of visits for a given patient ranges from 1 to 12, with a median number of visits of 2. There is a median of 113 days between visits (mean 158 days, standard deviation 171 days).

The study population is simulated via the following steps. First, baseline age is simulated for 4,000 subjects from a normal distribution centered at 75 with a standard deviation of 7. All subjects with a baseline age less than 65 are removed. Characteristics for the remaining subjects are simulated with an independent covariate structure. The variables that are binary: sex, health coverage status (Medicaid, Medicare, private), and all of the history of comorbidities are simulated with rates taken or approximated. BMI is simulated from [53] split into the common categories as specified by the Centers for Disease Control and Prevention [60]. The number of prescriptions is simulated via a zero-inflated Poisson with a rate approximated from [54] and based on the number of comorbidities for the subject. All variables are simulated to be time-invariant.

Factors in the dementia diagnostic model are given hazard ratios approximated from sources or hazard rates in the direction noted previously [55, 56, 57, 58]. Variables associated with the hazard of a dementia diagnosis include: baseline age, sex, BMI, health coverage status, and histories of depression, vascular disease, cerebrovascular disease, and diabetes. Censoring rate is determined from a Uniform distribution to achieve approximately a 10% event rate.

Factors determined to be associated with the time between clinic visits are determined through a thought experiment without consideration of SM. These rates are considered by asking: what would cause someone to come back to the doctor more or less often? Variables associated with shorter time-to-return: obesity, Medicaid health coverage, and histories of:

cardiovascular disease (CVD), mental or cognitive disorders, hypertension, diabetes, cancer, and renal disease. Variables associated with longer time-to-return: prescription quantity, Private health coverage, normal BMI, and histories of drug, alcohol, and tobacco use. Note that variables such as drug and alcohol use may be related to more health problems in the real world, but are simulated to be associated with longer time-to-return because people may choose to self-medicate instead of visiting a doctor. Clinic visit times are simulated from an exponential distribution. The number of visits a patient has depends on the time of their censoring or dementia diagnosis. When the cumulative sum of the draws from the time-between-visits distribution is greater than the censoring/dementia diagnosis time, then no more clinic visits are simulated.

To include SM, each subject receives an indicator of SM drawn from a Bernoulli distribution with individual probabilities of SM. Patients with a dementia diagnosis and system migration indicator are flagged to have a system migration visit before their diagnosis time. The SM is induced by extending the event time by another clinic visit draw. Thus, it is as if the visit where the patient should have be diagnosed occurs at another system. The patient is then diagnosed upon return to the "IHS" system at their next visit.

Factors associated with SM are determined through thought experiment considering the available variables. Factors that increase the number of preventable hospital visits are associated with a lower probability of SM. These factors are age, Medicaid coverage, an interaction between Medicaid and age, an interaction between private health coverage and age, and histories of CVD and renal disease. Factors associated with an increase in the probability of SM are Medicare health coverage, history of cancer, and an interaction between age and Medicare health coverage. These are chosen because patients with Medicare likely have access to multiple health systems, and cancer requires specialty care that may not be available at all IHS locations.

3.4.2 Measures

We use the Cox model and the Migration-Adjusted Cox model, both using time-on-study as the time scale and adjusted for baseline age as a covariate [34, 38].

Our sample consists of 3,680 subjects, 454 (12%) of whom have an incident dementia diagnosis. 79% of subjects fell into the Medicare and IHS-only group, meaning that they only have access to IHS and Medicare health coverage. 12% of the subjects had Medicaid coverage, possibly in addition to Medicare coverage. 9% of the subjects have private health coverage, regardless of the status of other coverages. Subjects with dementia have an older average age by 9 years (84 vs 75). 60% of the participants are female, 51% have diabetes, 21% have depression, 58% have CVD, and 75% have hypertension. 64% of the subjects with incident dementia migrate out-of-system. Most comorbidities are balanced between patients with incident dementia and those without incident dementia. Covariate distributions can be found in Table 3.1 and history of disease distributions can be found in Table 3.2.

3.4.3 Results

We fit the MA-Cox model and the Cox model to the simulated data. Our predictor of interest is health coverage status because we believe it to be an important confounder in the relationship between the time of dementia diagnoses and system migration. We adjust for sex, baseline age, BMI, and histories of depression, vascular disease, cerebrovascular disease, and diabetes as important adjustment variables in this relationship. Our interest in the analysis of the simulated data set is to compare the estimates of the covariates between the traditional Cox model (our “naive” model) and the MA-Cox model.

With our simulated data, we expect patients with Medicare health coverage and history of cancer to increase the probability of system migration. We are interested in how the

	No Dementia n = 3226 (87.7%)	Dementia n = 454 (12.3%)	Total n = 3680 (100%)
System Migration			
No Migration	1701 (52.7%)	165 (36.3%)	1866 (50.7%)
Migration	1525 (47.3%)	289 (63.7%)	1814 (49.3%)
Baseline Age	74.93 (5.45)	84.35 (3.82)	76.09 (6.12)
Health Coverage Status			
Medicare and IHS-only	399 (12.4%)	55 (12.1%)	454 (12.3%)
Medicaid	2539 (78.7%)	363 (80%)	2902 (78.9%)
Private	288 (8.9%)	36 (7.9%)	324 (8.8%)
Sex			
Male	1294 (40.1%)	165 (36.3%)	1459 (39.6%)
Female	1932 (59.9%)	289 (63.7%)	2221 (60.4%)
Body Mass Index	28.76 (4.8)	28.18 (4.98)	28.69 (4.83)
Categorical BMI			
Underweight	111 (3.4%)	24 (5.3%)	135 (3.7%)
Normal	546 (16.9%)	86 (18.9%)	632 (17.2%)
Overweight	1017 (31.5%)	152 (33.5%)	1169 (31.8%)
Obese	1552 (48.1%)	192 (42.3%)	1744 (47.4%)
Prescription Quantity	1.43 (2.17)	1.5 (2.16)	1.44 (2.16)

Table 3.1: This table depicts the distributions of system migration, the covariate of interest (health coverage status), and covariates. We report the mean (standard deviation) for continuous variables and count (percent) for categorical variables.

Disease History	No Dementia n = 3226 (87.7%)	Dementia n = 454 (12.3%)	Total n = 3680 (100%)
Diabetes			
No Diabetes	1580 (49%)	218 (48%)	1798 (48.9%)
Diabetes	1646 (51%)	236 (52%)	1882 (51.1%)
Depression			
No Depression	2558 (79.3%)	357 (78.6%)	2915 (79.2%)
Depression	668 (20.7%)	97 (21.4%)	765 (20.8%)
Hypertension			
No Hypertension	805 (25%)	114 (25.1%)	919 (25%)
Hypertension	2421 (75%)	340 (74.9%)	2761 (75%)
Vascular Disease			
No Vascular Disease	2627 (81.4%)	358 (78.9%)	2985 (81.1%)
Vascular Disease	599 (18.6%)	96 (21.1%)	695 (18.9%)
Cerebrovascular Disease			
No Cerebrovascular Disease	2796 (86.7%)	394 (86.8%)	3190 (86.7%)
Cerebrovascular Disease	430 (13.3%)	60 (13.2%)	490 (13.3%)
Renal Disease			
No Renal Disease	2369 (73.4%)	351 (77.3%)	2720 (73.9%)
Renal Disease	857 (26.6%)	103 (22.7%)	960 (26.1%)
Cardiovascular Disease			
No CVD	1348 (41.8%)	206 (45.4%)	1554 (42.2%)
CVD	1878 (58.2%)	248 (54.6%)	2126 (57.8%)
Cancer			
No Cancer	2948 (91.4%)	423 (93.2%)	3371 (91.6%)
Cancer	278 (8.6%)	31 (6.8%)	309 (8.4%)
Tobacco-use Disorder			
No Tobacco-use Disorder	3124 (96.8%)	436 (96%)	3560 (96.7%)
Tobacco-use Disorder	102 (3.2%)	18 (4%)	120 (3.3%)
Alcohol-use Disorder			
No Alcohol-use Disorder	3001 (93%)	421 (92.7%)	3422 (93%)
Alcohol-use Disorder	225 (7%)	33 (7.3%)	258 (7%)
Drug-use Disorder			
No Drug-use Disorder	3141 (97.4%)	446 (98.2%)	3587 (97.5%)
Drug-use Disorder	85 (2.6%)	8 (1.8%)	93 (2.5%)

Table 3.2: This table shows the distributions of comorbidities. History of disease was simulated to be time-invariant. We report the mean (standard deviation) for continuous variables and count (percent) for categorical variables.

hazard ratios for Medicaid relative to Medicare, and Private relative to Medicare may change between the two models. As a note, most of the other covariates have fairly similar estimates and draw the same conclusions between the two models. This is expected because those covariates do not impact SM. The small changes may be due to the relationship those variables have with the time-to-return to clinic. BMI is the only covariate that has different conclusions between the two models. This may be because the data is simulated such that a normal BMI is associated with a longer time between visits, an obese BMI is simulated to have a shorter time between visits, and overweight and underweight were simulated to have no impact on the time between visits. Since normal BMI is the reference group, the imputation model may have picked up some of this relationship and adjusted for it in the imputation phase of the MA-Cox model. As for the predictors of interest, we observe a large change in the estimated hazard ratios for subjects with Medicaid (relative to Medicare) coverage and for subjects with Private (relative to Medicare) coverage. The hazard ratios change from outrageously large estimates of 6-7 fold higher risk of a dementia diagnosis to more common 93% higher risk (for Medicaid coverage) and 55% higher risk (for Private coverage). While the traditional Cox model's estimated hazard ratios are likely higher than would be observed in a real data set, the reduction of estimated hazard ratios under the MA-Cox model are what we would expect to see in an analysis with real data. This reduction is due to the subjects with Medicare coverage being more likely to migrate to other healthcare systems, resulting in observed dementia diagnosis times that are delayed. The MA-Cox model is able to identify the delayed diagnostic times and impute earlier times, which reduces the hazard ratio estimates.

	Naive Results		MA-Cox Results	
Covariate	HR (95% CI)	p-value	HR (95% CI)	p-value
Health Coverage Status				
Medicare and IHS-Only	**Referent**	**-**	**Referent**	**-**
Medicaid	**6.87 (4.95, 9.54)**	**<0.001**	**1.93 (1.45, 2.57)**	**<0.001**
Private	**6.20 (4.21, 9.12)**	**<0.001**	**1.55 (1.11, 2.17)**	**0.01**
Baseline Age	1.15 (1.12, 1.19)	<0.001	1.22 (1.20, 1.25)	<0.001
Sex (reference Male)				
Female	1.03 (0.82, 1.28)	0.826	1.18 (0.92, 1.52)	0.19
History of				
Depression	0.99 (0.73, 1.34)	0.935	1.25 (0.89, 1.75)	0.208
Vascular Disease	1.08 (0.82, 1.41)	0.578	1.11 (0.81, 1.52)	0.526
Cerebrovascular Disease	1.05 (0.76, 1.47)	0.755	0.96 (0.66, 1.40)	0.829
Diabetes	1.00 (0.80, 1.27)	0.971	1.06 (0.78, 1.45)	0.829
Body Mass Index				
Underweight	2.93 (1.72, 5.01)	<0.001	1.41 (0.78, 2.53)	0.252
Normal	Referent	-	Referent	-
Overweight	1.80 (1.19, 2.72)	0.005	1.69 (0.92, 3.08)	0.089
Obese	1.72 (1.17, 2.52)	0.006	1.41 (0.78, 2.53)	0.252

Table 3.3: Results from the Traditional Cox model compared to the MA-Cox model for the simulated IHS data set. Hazard ratio (HR) estimates, 95% Wald-based confidence intervals and p-values are provided for each variable in each model.

3.5 Discussion

In EHR data, patients who have access to multiple health systems may seek treatment for certain conditions at specific health systems while seeking treatment for other conditions at other health systems. We call this phenomenon system migration. In this chapter, we investigate how SM may impact coefficient estimates from the Cox model, and develop a novel method, a Migration Adjusted Cox model, for reducing the bias in these settings. System migration is shown to bias coefficient estimates from the Cox model by greater than 15% in settings with as little as 17% of diagnosed patients migrating out-of-system. Further, our simulations consider the setting where subjects are diagnosed within the system immediately upon return-to-system. We hypothesize that the bias would be even larger if patients are not diagnosed within system at their first return-to-system visit.

We consider simulation settings with right censoring rates of 0%, 35%, and 70%, each with system migration impacting the diagnosis times ranging from 0% to 40%. Under the MAR1 missingness mechanism, we observe that coefficient estimates from the Cox model are biased toward the null when SM is present. Under the MAR2 missingness mechanism, we simulate data such that the coefficient estimates from the Cox model are biased away from the null. Here, we observe estimates with bias ranging from 7% to 77%. In most of the simulation settings, the MA-Cox model reduces the bias to nominal levels. The bias in the MAR2 settings is away from the full in our settings, but can be toward the null also. In particular, when subjects with earlier diagnosis times are the patients migrating out-of-system, then the observed diagnostic times are delayed. This would result in the SM patients appearing to have similar diagnostic times to the patients who do not migrate out-of-system, which results in estimated hazard ratios that are biased toward no difference between the groups.

While the MA-Cox model reduces bias in most settings, the model over-corrects for the bias and under-estimates the true coefficient value in a few settings where there is relatively low amounts of SM that impact the diagnostic times for subjects. This over-correction may improve with the use of more sophisticated prediction algorithms. Some machine learning methods, or methods developed specifically for model building in recurrent event survival analysis, may improve the ability of the MA-Cox model to identify subjects who migrate out-of-system, and thus improve the imputation step. Further, the asymptotic results of the coefficients rely on a correctly-specified prediction model in step 1 of the proposed method. Thus, it is necessary to use a model building technique that estimates an honest assessment of out-of-sample prediction error to ensure valid results from the MA-Cox model. Another limitation for the MA-Cox model is the reliance on information of patient system-usage to estimate the probabilities of SM. While EHR data generally has recurrent visits for patients, it is necessary to have at least a few visits per patient to use this method.

In this chapter, we propose a novel method for estimating time-to-event outcomes in the setting where SM may be of concern. The MA-Cox model is shown to reduce bias in the presence of SM across several levels of SM and right-censoring. We recommend the proposed method in settings where patients may have access to multiple healthcare options, even if system migration is not of concern as the method does not induce bias when system migration is not present. This method is applied to simulated AI/AN health care data to demonstrate the potential impact of SM on covariates. The MA-Cox model adjusts coefficient estimates for covariates that are related to SM while not changing the coefficient estimates for the covariates that do not contribute to SM.

More work is needed to consider SM bias and the MA-Cox model when the proportional hazards assumption is broken. Future work can involve developing a censoring-robust MA-Cox model, similar to the work of Nuño and Gillen [61] in the nested case-control setting. Also, estimating the probability of migrating out-of-system for each subject is vital to the MA-Cox model's imputation method, so improving the method of selecting the best subset of covariates for predicting time-to-return-to-system may make this method more reliable.

Chapter 4

Receiver Operating Characteristic Curves for Time-Dependent Recurrent Event Survival Data

4.1 Introduction

In many healthcare related studies, predicting the occurrence of some event is oftentimes of interest. In the cross-sectional setting with a single event the receiver operating characteristic (ROC) curves have been established as a powerful tool for assessing a model's predictive ability [62]. ROC curves are a plot of the sensitivity against (1 - specificity) across a range of values for some continuous risk score [62]. The area under the ROC curve (AUC) is a common summary metric for the predictive ability of a marker marginalized over all possible values of the risk score. The AUC can be understood as an estimate of the probability that the risk score for a randomly sampled case is higher than the risk score for a randomly sampled control [62]. These traditional metrics for ROC curves and the AUC are not appropriate when disease status can change over time, as in the case of survival analysis [32]. Heagerty et al. (2000) [30] proposed a method for constructing ROC curves and the calculation of the AUC when disease status is time-dependent. Heagerty and Zheng (2005) [31] and Chambless and Diao (2006) [32] further proposed methods for estimating sensitivity and specificity using risk scores derived from the linear predictors of a Cox proportional hazards (PH) model.

To our knowledge, no one has considered time-dependent ROC curves in the setting of recurrent event survival analysis. Recurrent event survival analysis occurs when subjects may experience the event of interest multiple times. In these settings, it may be of interest to consider what factors predict the occurrence of a next event. For example, the National Alzheimer's Coordinating Center's (NACC) Uniform Data Set (UDS) is a collection of data on study participants enrolled at Alzheimer's Disease Research Centers (ADRCs). The data is collected longitudinally to study the onset and progression of Alzheimer's disease and related dementias (ADRD) [63]. When conducting longitudinal studies, it is imperative to retain patients for the duration of the study. Attrition of research participants can lead to lower power, reduced generalizability, and biased results [64]. Previous research has suggested that American Indian and Alaska Native (AI/AN) research participants in the NACC UDS were less likely to be retained than non-Hispanic White (NHW) participants [65]. A natural next question is to explore what factors predict retention in AI/AN research participants to identify potential areas of intervention for increasing retention. Current time-dependent sensitivity, specificity, and AUC estimators will weight each event the same. This will result in estimates that generalize to individuals with more events. Essentially, participants with more events have contribute more information due to having more follow-up visits. Here we propose a method estimating time-dependent sensitivity, specificity, and AUC when subjects can have more than one event.

In this chapter, we start with a background of some of the methods mentioned in Chapter 2. We then develop a method for calculating sensitivity, specificity, and AUC in the recurrent event setting. We provide simulations to illustrate the performance of the estimator compared to the Chambless and Diao (2006) [32] estimator. We then provide an application of the method to the NACC UDS for identifying predictors of retention for AI/AN study participants.

4.2 Background

Heagerty et al. (2000) [30] proposed time-dependent ROC curves to estimate sensitivity, specificity, and AUC at some meaningful (i.e. clinically relevant) time. Estimating these quantities in a time-dependent setting requires specification of which subjects are controls and which are cases at each time t. Heagerty et al. [31] specify that controls can be either static or dynamic. Static controls means that everyone who is disease-free until some time t^* is considered a control. This means *static specificity*, $\text{Spec}_{\bar{D}}(c, t^*) = Pr(M_i \leq c \mid T_i > t^*)$, is defined as the probability that a subject's risk score, M_i, is smaller than some threshold, c, given that their event time, T_i, is greater than some fixed time t^*, where t^* is usually taken be some time equal to or greater than the time of interest in the study and where $\bar{D}$ is used to denote "static." Likewise, the dynamic control specification allows subjects to contribute as controls up until their event time. *Dynamic specificity* is similarly defined as $\text{Spec}_D(c,\ t) = Pr(M_i \leq c \mid T_i > t)$. Additionally, Heagerty et al. [31] similarly note that cases can be incident or cumulative. Incident cases means that cases are only the subjects who experience the event at time t. This gives rise to *incident sensitivity*, $\text{Sens}_I(t, c) = P(M_i > c \mid T_i = t)$. Moreover, cumulative cases are defined such that cases at each time t consist of all subjects who experience the event up until (and including) time t. Thus, *cumulative sensitivity* is $\text{Sens}_C(t, c) = Pr(M_i > c \mid 0 < T_i \leq t)$. From these definitions, time-dependent estimators of sensitivity, specificity, and AUC can be situated into the following categories: cumulative sensitivity and dynamic specificity (C/D), incident sensitivity and static specificity ($I/\bar{D}$), or incident sensitivity and dynamic specificity (I/D) [31]. In the recurrent event setting, the C/D estimator of Chambless and Diao (2006) [32] is the most natural because it allows participants to contribute to the specificity as controls, and allows participants to contribute multiple times to the sensitivity as they accumulate more events.

Chambless and Diao [32] developed sensitivity and specificity estimators in the C/D framework, where the linear predictors from a Cox model serve as the risk score. That is, let $M = Z\hat{\beta}$, where Z are covariates and $\hat{\beta}$ are estimated regression parameters. This framework classifies every subject as either a case or a control for every fixed time t, so that subjects contribute to the estimation of the specificity prior to their event time and contribute to the estimation of the sensitivity at all points after their event time. Chambless and Diao [32] denote $\text{Sens}_{CD}(t,c) = \frac{E(1-S(t|M))I(c<M)}{E(1-S(t|Z))}$ and $\text{Spec}_{CD}(t,c) = \frac{E(S(t|M)I(c>M))}{E(S(t|M))}$, where $E(\cdot)$ is the expectation, $M = Z\beta$ are the linear predictors from a Cox model, and $S(t \mid M)$ is the survival function. These quantities can be estimated by:

$$\widehat{\text{Sens}}_{CD}(t,c) = \frac{\frac{1}{n}\sum_{i=1}^{n}[(1-\hat{S}(t \mid M_i)I(M_i > c)]}{\frac{1}{n}\sum_{i=1}^{n}[1-\hat{S}(t \mid M_i)]}$$

and

$$\widehat{\text{Spec}}_{CD}(t,c) = \frac{\frac{1}{n}\sum_{i=1}^{n}[\hat{S}(t \mid M_i)I(M_i < c)]}{\frac{1}{n}\sum_{i=1}^{n}\hat{S}(t \mid M_i)}$$

where $\hat{S}(t \mid M)$ is the estimated survival function from a Cox model which is estimated via $\hat{S}(t \mid M) = exp\left\{-\hat{\Lambda}_0(t)exp(Z\hat{\beta})\right\}$, where $\hat{\Lambda}_0(t)$ is the Breslow estimator for the baseline hazard [33] and $\hat{\beta}$ is estimated via the Cox model. Further, they provide an estimator for the area under the ROC curve created by the above estimators for sensitivity and (1 - specificity) across all potential thresholds c. Let M_i and M_j be independent observations of the marker, M, then

$$\text{AUC}_{CD}(t) = \frac{E[(1-S(t \mid M_i))S(t \mid M_j)I(M_j < M_i)]}{E[1-S(t \mid M)]E[S(t \mid M)]}.$$

This AUC can be estimated by replacing M with linear predictors from a Cox model and taking a mean over the distinct (M_i, M_j) pairs:

$$\widehat{\text{AUC}}_{CD}(t) = \frac{\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}[(1-\hat{S}(t \mid M_i))\hat{S}(t \mid M_j)I(M_j < M_i)]}{\frac{1}{n}\sum_{i=1}^{n}[1-\hat{S}(t \mid M_i)]\frac{1}{n}\sum_{i=1}^{n}\hat{S}(t \mid M_i)},$$

where $\hat{S}(\cdot)$ is estimated the same way as above. More information can be found in [32].

4.3 Methods

As previously noted, the Chambless and Diao [32] estimator will give equal weight to each event in the recurrent event setting. This means individuals with more events will have more weight in $\widehat{\text{Sens}}_{CD}(t, c)$ and $\widehat{\text{Spec}}_{CD}(t, c)$. We consider reweighting these quantities by a subject's number of events. That is, sensitivity and specificity in the setting of right-censored recurrent event data can be estimated with:

$$\widehat{\text{Sens}}_{CD}(t, c) = \frac{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{K_i}\sum_{k=1}^{K_i}[(1-\hat{S}(t \mid M_{ik})I(M_{ik} > c)]}{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{K_i}\sum_{k=1}^{K_i}[1-\hat{S}(t \mid M_{ik})]}$$

and

$$\widehat{\text{Spec}}_{CD}(t, c) = \frac{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{K_i}\sum_{k=1}^{K_i}[\hat{S}(t \mid M_{ik})I(M_{ik} < c)]}{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{K_i}\sum_{k=1}^{K_i}\hat{S}(t \mid M_{ik})},$$

where M_{ik} is the k^{th} event for subject i and K_i is the total number of events for subject i.

Similarly, the AUC can be estimated via reweighting each individual's contribution by their number of events.

$$\widehat{\text{AUC}}_{CD}(t) = \frac{\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{1}{K_i}\sum_{k=1}^{K_i}[(1-\hat{S}(t \mid M_{ik}))\bar{S}(t \mid M_j)]}{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{K_i}\sum_{k=1}^{K_i}[1-\hat{S}(t \mid M_{ik})]\frac{1}{n}\sum_{i=1}^{n}\frac{1}{K_i}\sum_{k=1}^{K_i}\hat{S}(t \mid M_{ik})},$$

where M_{ik} is the linear predictor for the k^{th} event for subject i, $M_j = M_1, \ldots, M_{K_j}$ is a vector of subject j's linear predictors at each of their observations, and $\bar{S}(t \mid M_j) =$

$\frac{\sum_{k'=1}^{K_j} \hat{S}(t|M_{jk'})I(M_{jk'}<M_{ik})}{\sum_{k'=1}^{K_j} I(M_{jk'}<M_{ik})}$. Note that $\bar{S}(t \mid M_j)$ is an average over only the cases where subject j's linear predictors are smaller than than subject i's linear predictors in the setting where covariate values change with events. In practice, the way to implement this is through reweighting a subject's contribution to the score of the Cox model by their number of events. In some programs, this may require reweighting the subject's contribution to the baseline hazard as well.

4.4 Simulation Studies

We investigate the performance of the proposed estimator compared to the Chambless and Diao [32] estimator. We perform 1000 simulations of 1000 subjects coming from two subpopulations, $\text{subpop}_i \sim \text{Bernoulli}(p)$, where $p \in \{0.5, 0.6, 0.7, 0.8\}$ and where $\text{subpop}_i = 1$ indicates belonging to subpopulation B. We consider the setting of predicting event status based on a single predictor of interest, $X_i \sim \text{Normal}(0.5, (0.2)^2)$. We simulate event times, t_i from an exponential distribution with rate $\lambda_i = \lambda_0 \exp\{\beta_1 x_i + \beta_2 \text{subpop}_i + \beta_3 \text{subpop}_i * x_i\}$, where we fixed $\lambda_0 = 0.5$, $\beta_1 = \log(1.3)$, and $\beta_3 = 1.5$. We use β_2 to determine the ratio of the median number of events between the two subpopulations. Comparing subpopulation B:A, we fix the ratio of events as 1:1, 2:1, 3:1, and 5:1 by setting $\beta_2 \in \{-0.7, -0.15, 0.3, 055\}$, respectively. We consider the four prediction times 1, 3, 5, and 8 to be of interest. The maximum follow-up time is $\tau = 10$ with uniform censoring such that $C_i \sim \text{Uniform}(2, 10)$. To simulate recurrent events, we draw K_i times for subject i from the exponential distribution above such that $\sum_{j=1}^{K_i-1} t_j \leq C_i < \sum_{j=1}^{K_i} t_j$. Thus, the observed times for subject i are $\vec{t_i} = (t_1, t_2, \ldots, t_{K_i-1}, C_i)$.

Figures 4.1 and 4.2 depict the estimated AUCs from the simulated data for the CD estimator and the proposed estimator, respectively. The figures consist of four plots representing balances in the proportions subjects from subpopulations A and B. Within each of the plots

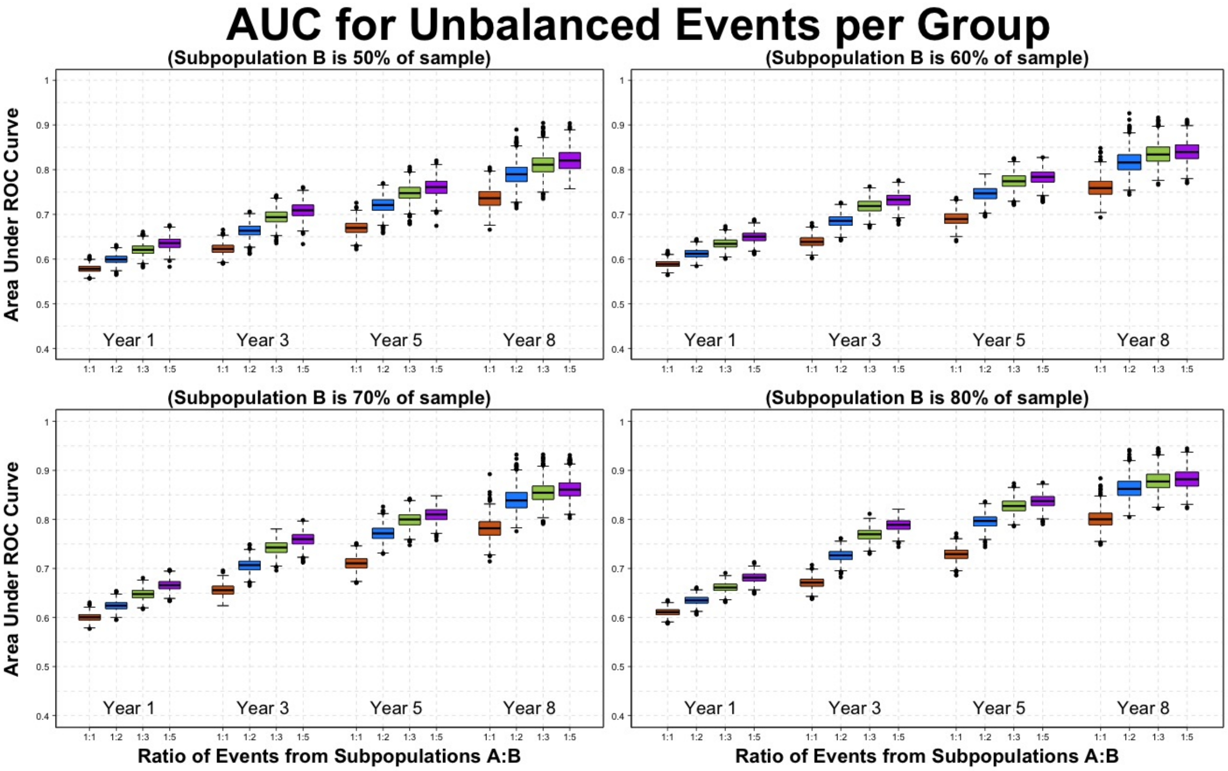


Figure 4.1: AUC estimates from the Chambless and Diao estimator for the simulated data with unbalanced number of subjects between subpopulations and with differing rates of events between the subpopulations.

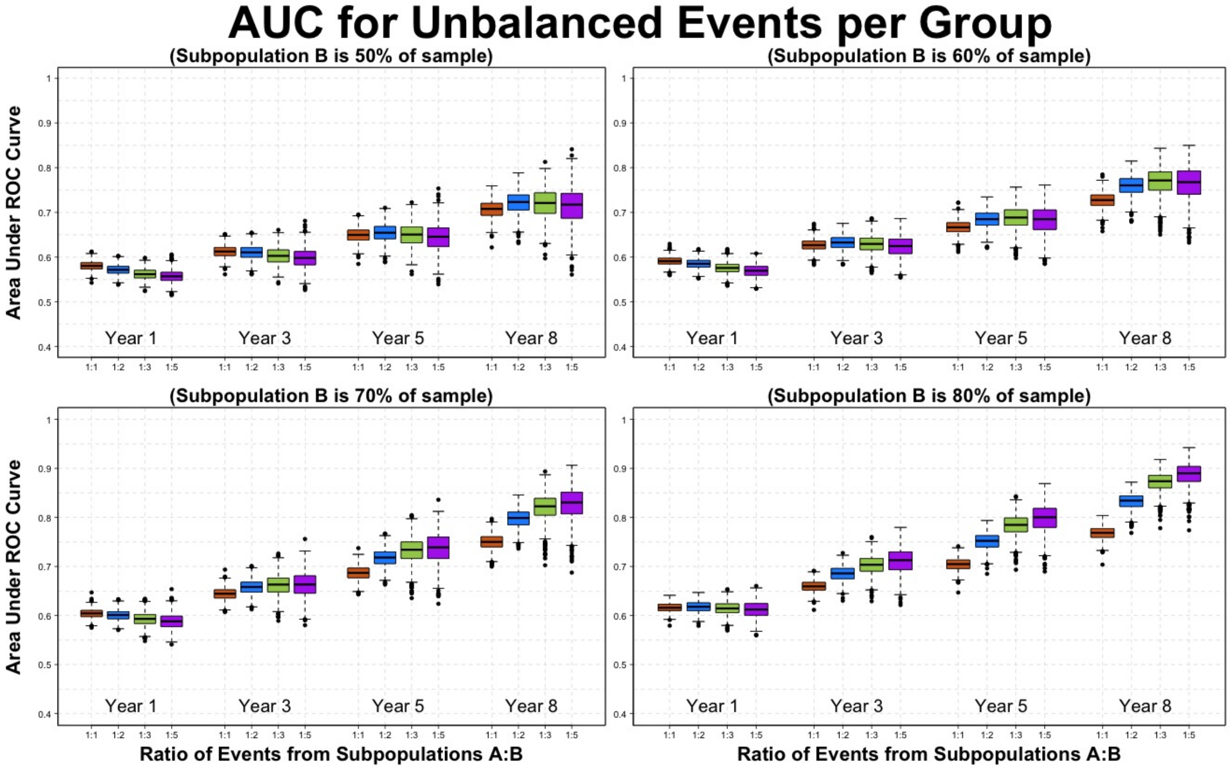


Figure 4.2: Estimated AUCs from the proposed estimator for the simulated data with unbalanced number of subjects between subpopulations and with differing rates of events between the subpopulations.

are boxplots of the AUCs for various ratios of events for the subpopulations at the times of interest.

As stated before, the CD estimator weights each event the same, so that the AUC estimates are biased toward the subpopulation with more events. We observe this trend in the increasing of the AUC estimates within a given year for each of the plots in Figure 4.1. Conversely, the proposed method estimates approximately the same AUC within years when the subpopulations are relatively balanced (50/50 and 40/60 for A to B, respectively). The proposed estimator also estimates approximately the same AUC in the early years of the unbalanced subpopulation settings (70% and 80% of participants from subpopulation B). In the later years, however, the proposed method exhibits a similar increasing trend in the AUCs as we observe for the the CD estimator. This occurs because there are no subjects from subpopulation A in the risk sets at those later times. With no subjects from subpopulation A in the risk sets, there is no way to reweight the contributions from subjects in subpopulation B back to even with subpopulation A.

4.5 Application

We apply the proposed method to identifying predictors potentially associated with retention of American Indian and Alaska Native (AI/AN) participants in the NACC UDS. The NACC UDS is a collection of longitudinal data from research participants at National Institute on Aging-funded Alzheimer's Disease Research Centers (ADRCs). Each ADRC has their own protocol for the recruitment and retention of participants. The NACC UDS is primarily a repository for data on subjects enrolled in Alzheimer's and related dementia research. A standardized set of questions and assessments are collected from each participant to be submitted to NACC. It was previously reported that AI/AN participants in the NACC data were retained at lower rates than non-Hispanic White participants [65]. In this prediction analysis, we consider data from 43 ADRCs collected between September 2005 and November 2021 to develop a model for predicting retention to the next follow-up visit for AI/AN participants in the NACC UDS [66].

We follow the same protocol as Coniff et al. [65] to develop a data set with the potential predictors: baseline diagnostic status (categorical: "Normal cognition", "Impaired-not-Mild Cognitive Impairment", "Mild cognitive impairment", and "Dementia"), baseline age (continuous), binary sex (categorical: male and female), education level (categorical: "Less than high school", "High school diploma", "Some college", "4-year degree", and "Greater than 4-year degree"), smoking status (categorical: "Never smoked", "Previously smoked", and "Currently smokes"), independence (categorical: "Able to live independent", "Requires assistance with complex tasks", "Requires assistance with basic activities", "Completely dependent"), marital status (categorical: "Married/partnered", "Previously married", "Never married"), co-participant sex (categorical: "Male", "Female"), body mass index (continuous), categorical body mass index (categorical: "Normal", "Overweight", "Obese"), and Hispanic ethnicity (categorical: "Yes", "No"). Details of data cleaning can be found in [65]. We consider retention as having a next visit, regardless of the amount of time between visits. Participants are censored at the minimum date of 1) end of follow-up (November 21, 2021), 2) 180 days after their last observed visit, 3) death, or 4) study discontinuation.

The analysis data set consists of 197 AI/AN participants from 24 ADRCs who attended between 0 and 11 follow-up visits. The median number of follow-up visits was one, with 84% of participants completing three or fewer follow-up visits. The average time between visits was 14.5 months. Information on covariate distributions can be found in Table 4.1.

We compare the models that maximize AUC for predicting retention to the next visit between the proposed method (M_{prop}) with weights set as each individual's number of visits to the CD method (M_{CD}), which gives equal weight to each event, regardless of the number of visits for a given individual. The recurrent event Cox model for M_{Prop} follows the Andersen and Gill (1982) [49] specification. Since the specification from [49] does not work with the Chambless and Diao [32] estimator, we specify a Cox model that predicts the gap time between events, similar to the gap time model proposed by Prentice et al. (1981) [67] except that we do not stratify by event number in this example. We estimate the AUCs at 1,500 days post-enrollment. The model scope consists 2,047 models with only first order covariates and no interactions. The data set also does not include any comorbidities, which may impact retention.

Table 4.2 depicts the selected models for each method. The two methods identify the same eight covariates as predictive of retention. M_{Prop} identifies two additional covariates that M_{CD} does not: 1) continuous BMI and 2) Hispanic ethnicity. We note that these two predictors are likely not indictive of retention, but rather are likely proxies for comorbididies

Covariate	AI/AN (n = 197)
Age	69 (10.8)
Sex:	
Female	125 (63%)
Male	72 (37%)
Co-participant Sex:	
Female	141 (72%)
Male	56 (28%)
Hispanic Ethnicity:	
Yes	30 (15%)
No	167 (85%)
Cognitive Status:	
Normal	76 (39%)
Impaired not-MCI	8 (4%)
MCI	37 (19%)
Dementia	76 (39%)
Independence Level:	
Completely Independent	123 (62%)
Slightly Dependent	53 (27%)
Dependent	21 (11%)
Education Level:	
< High School	43 (21%)
High School Diploma	75 (38%)
Some College	38 (19%)
4-year Degree	19 (10%)
More than 4-year Degree	22 (11%)
Marital Status:	
Married	114 (58%)
Previously Married	75 (38%)
Never Married	8 (4%)
Smoking Status:	
Never	87 (44%)
Previous	88 (45%)
Current	22 (11%)
BMI (continuous)	30.3 (6.9)
Normal	48 (24%)
Overweight	65 (33%)
Obese	84 (42%)

Table 4.1: Descriptive statistics on the 193 American Indian and Alaska Native participants in the NACC UDS who are contribute to the development of the prediction models. Continuous covariates display the mean (standard deviation). Categorical covariates display count (percent). AI/AN stands for American Indian and Alaska Native. BMI stands for Body Mass Index. MCI stands for Mild Cognitive Impairment.

Proposed Method (M_{Prop})	CD Method (M_{CD})
Diagnostic status	Diagnostic status
Baseline Age	Baseline Age
Independence	Independence
Sex	Sex
Education	Education
Marital Status	Marital Status
Co-participant Sex	Co-participant Sex
Smoking status	Smoking status
BMI	
Hispanic Ethnicity	

Table 4.2: The covariates selected as most predictive of retention for AI/AN research participants in the NACC UDS as determined by maximizing AUC. The proposed method uses the Andersen and Gill (1982) [49] specification of the Cox model for recurrent events. The CD method uses the Prentice et al. [67] gap time model without stratification for the Cox model for recurrent event.

	Selected Model	
	M_{Prop}	M_{CD}
AUC from M_{Prop}	0.6457	0.6439
AUC from M_{CD}	0.8040	0.8047

Table 4.3: The AUC estimates for each model. The diagonal elements represent the AUCs from the models in Table 4.2. The top-right off-diagonal is the AUC estimates from the proposed method for the model in column 2 of Table 4.2. The bottom-left off-diagonal is the AUC estimates from the CD method for the model in column 1 of Table 4.2.

that were not included in the model space. We also note that the AUC estimates for M_{Prop} are generally lower than for the M_{CD}. Table 4.3 shows the AUC estimates from each method for each of the selected models. Similar to what the simulations show, the Chambless and Diao estimator has higher AUCs than the proposed method. The higher AUC estimates indicate that the M_{CD} will generalize better to the participants with more events. The M_{Prop} model, however, suggests that the predictive ability of these models is lower when the goal is to predict retention for all participants.

The M_{Prop} method estimates lower AUCs across almost all 2,047 models compared to the M_{CD} method. Figure 4.3 shows the AUC estimates for both methods across the entire model space. The AUC estimates for M_{Prop} are higher for the models with fewer covariates, but

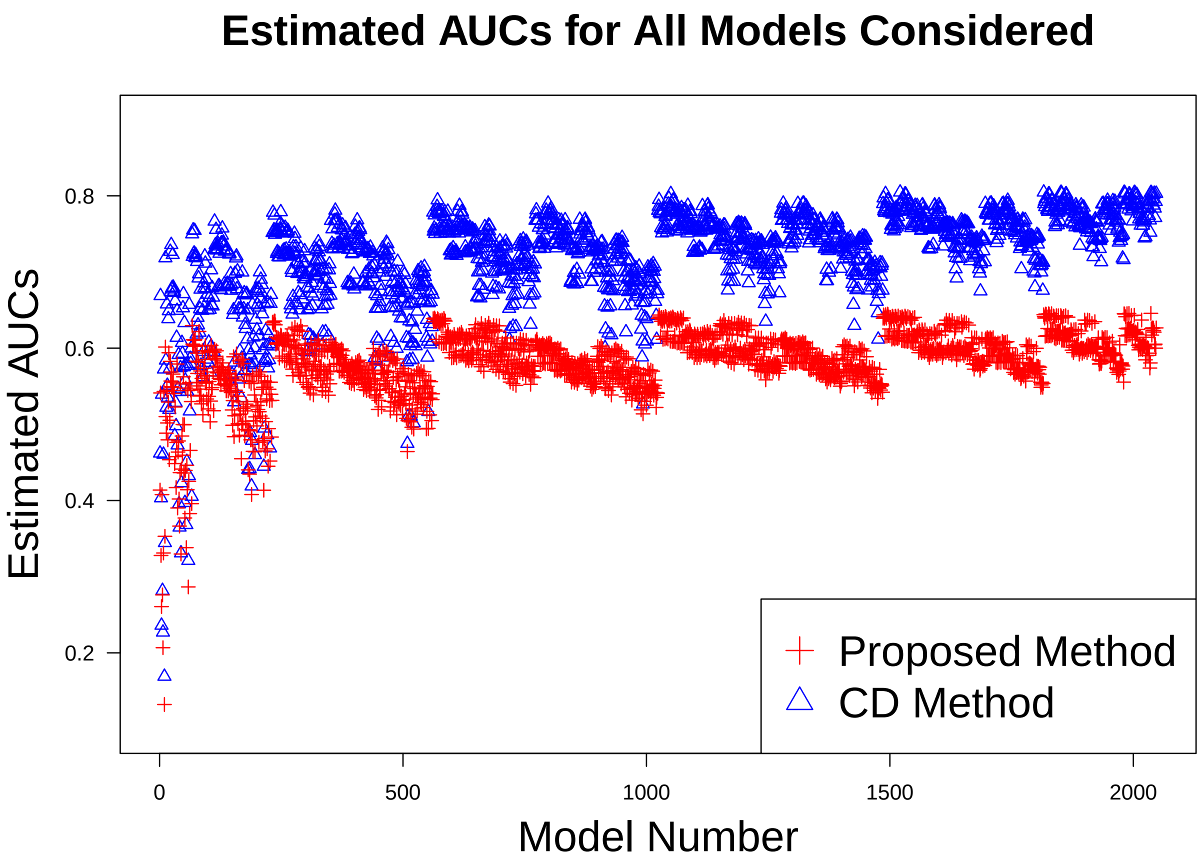


Figure 4.3: Shows the AUC estimates for each of the 2,047 models in the model space. The red crosses are AUC estimates from M_{Prop}. The blue triangles are AUC estimates from M_{CD}. Note that M_{Prop} estimates lower AUCs in most settings, especially outside of the first few models, which were the models with the fewest covariates.

as the model complexity increases, the M_{CD} method estimates consistently higher AUCs than does M_{Prop}. The AUCs are lower for the M_{Prop} because the selected covariates predict retention to the next visit for people with more visits. In Figure 4.3, some of the simplest models have higher AUC estimates from the M_{Prop} method because the factors better predict retention for participants with fewer return visits.

4.6 Discussion

In this chapter, we develop an estimator for time-dependent sensitivity, specificity, and area under the ROC curve in the recurrent event setting. Previous methods of estimating time-dependent sensitivity, specificity, and AUC applied to the recurrent event setting gives equal weight to every event, both within and across individuals. This results in AUC estimates that depend on the ratio of events across subpopulations as well as the balance among the subpopulations within the sample. To account for this dependence, we propose reweighting each subjects contribution to the Chambless and Diao [32] estimators by their number of events. We demonstrate through simulation how this adjustment results in estimating similar AUCs for settings with relative balance between the subpopulations across varying ratios of events between subjects in the subpopulations. We also show how this estimator may be broken when the subpopulations are severely unbalanced in the distribution of subjects between subpopulation and in the ratio of events.

We apply the method to an analysis of the retention of American Indian and Alaska Native research participants in the National Alzhiemer's Coordinating Center Uniform Data Set. Our results show that the CD estimator of AUC and the proposed estimator of AUC can identify different models as the being the most predictive of retention to next visit. Further, the AUC estimates may be different, resulting in different levels of confidence in the predictive ability of the model. This occurs because the CD estimator will generalize to subpopulations who have more events, whereas the proposed estimator will generalize to a broad population. Based on this, the CD estimator is appropriate to use in the recurrent event setting when the goal is to predict more events, regardless of whether the events are cluster-based or not. For example, consider the setting where the cost of treating each event is the same, regardless of whether it is an individual's fifth event or their first event. It is appropriate to use the CD estimator for predicting the next event because each event has the same cost. Weighting each event the same and understanding factors that lead to future events can help reduce costs. The CD estimator will also likely have better precision in its estimates because it treats each event as independent. The proposed estimator is appropriate to use when the goal is to generalize to a broader population than the subpopulation with the most events. For example, in longitudinal research studies it is imperative to retain participants for the duration of the study to avoid potential loss of power, reduced generalizability, and biased results [64]. The proposed method can identify covariates that predict retention to the next visit for all participants, which provides options for intervention. We further note that the simulation settings and application show lower AUC estimates for the proposed method compared to the CD method. This is because the predictors in both settings are predictive of more events. The proposed method can estimate higher AUCs if the predictors are indicative of a fewer number of events.

The simulation results show that the proposed estimator does not fully account for differences in the rates of events between two subpopulations when the subpopulations are severely different. When 80% of patients are from the subpopulation that experiences more events, we observe AUC estimates that depend on the number of events per subject. This occurs because there are no subjects from subpopulation A remaining in the risk set at the later times, thus the reweighting scheme fails to balance the contributions between the subpopulations. This can likely be accounted for by additionally reweighting a subject's contribution to the AUC by the probability of being censored. Another limitation is that the simulations only allow for reweighting by the number of visits for each subject. In some settings, the events may be correlated, and it may be of interest to reweight an individual's contribution to the AUC by some quantity between the CD estimator and the proposed estimator. This can be accomplished by allowing a more generalized weighting scheme.

The proposed method can be used to develop a prediction model for recurrent event survival data. This can include developing a risk score for future events, or for attempting to predict the time until a next event. We demonstrate one potential application in retention analyses for longitudinal studies.

Future work on this topic to be included in the related manuscript include deriving an analytical variance, developing a bootstrap estimator of the variance, and allowing for reweighting by different quantities. Other future work includes extending the censoring robust time-dependent AUC estimator developed by Nuño and Gillen (2021) [68].

Chapter 5

An Investigation of the Choice of Time Scale under Misspecification of Time

5.1 Introduction

Survival analysis models consider the time until an event occurs. The time component defines the ordering of event times as well as the risk sets at each event time [34]. Time is measured from some origin, called the *time origin*, which is the point at which subjects are initially considered at-risk for the event. The time origin defines how time is measured in the model, which is called the *time scale*. In practice, there are two common time scales: age and time-on-study [19]. In the age time scale, the time origin is birth and the study measures the risk of experiencing the event at a particular age. No covariate adjustment for age is included when fitting a model under this time scale [35]. In the time-on-study time scale, the time origin is the first visit, or examination, or study enrollment. This time scale measures the risk of experiencing the event from the time observation begins. When fitting a model with this time scale, baseline age or age-at-entry to the study should be included as an adjustment covariate [34, 35].

The choice of time scale and time origin can be straightforward in some settings. For example, when running a clinical trial to test the effectiveness of a new drug, it makes sense to measure time from the date of randomization [35]. That date is when participants are given the treatment (be it placebo or the experimental drug), and thus is the earliest possible time for an effect to start taking place. In some settings, the time origin is less clear. For example, in epidemiological cohort studies, we are interested in learning about risk factors related to some condition. For a condition like dementia, a time origin related to study start is less meaningful because the study start date is not related to dementia risk. Dementia risk is more related to age, which makes the age time scale a more natural choice. However, the risk of dementia does not really start at birth. In the general population, the risk of dementia is effectively 0 at birth, and remains that way for many years [69]. While dementia symptoms may begin in some populations as early as the late-teens or early twenties, this is not the case for most people [70]. In reality, the risk for dementia onset in the general population does not begin until sometime in the 40s, or 50s [69].

Several criterion have been considered previously to help with the choice of time scale. Korn et al. (1997) [36] wrote the first major paper considering the appropriateness of the time-on-study time scale for epidemiological studies. They note that most time-to-event analyses used time-on-study as the time scale and adjusted for baseline age as a covariate. They generally recommend using a Cox model stratified on birth cohort with age as the time scale (and thus no covariate adjustment for baseline age). Their reasoning is largely heuristic in that age is a more appropriate setting for considering the time-to-onset of conditions that do not have a clear exposure date. They further consider two conditions for when incorrectly using the time-on-study time scale would be equivalent (heuristically) to the correct model using age as the time scale. First, when the baseline hazard takes an (approximately) exponential form, they argue that the models are mathematically equivalent and thus should result in similar coefficient estimates. Second, they argue state that the coefficient estimates should be the same when the covariate-of-interest and the age-at-entry to the study are independent. These conditions have been assessed by others through simulation [34, 35, 37, 38]. The results have been mixed. For the second condition, Thiebaut and Benichou (2004) [35] and Chalise et al (2016) [71] report that independence is not a sufficient condition for equivalence of the coefficient estimates. In contrast, Pencina et al. (2007) [37] suggests that the second condition does hold in simulation. Chalise et al. (2012) [34] further consider the conditions from the standpoint of estimating procedures for the Cox model. They note that models using the age time scale will not include an age covariate in the partial likelihood, thus meaning that the distribution of age will not influence the estimated hazard ratios as it will in the time-on-study model. They conclude that the results from the two time scales will only be the same when the age-at-entry drops out of the partial likelihood in the time-on-study model. This happens when all subjects enter the study at the same age, because the event times and risk sets across models will be the same at all times t. On the other hand, if the variation in age-at-entry is non-zero, then the risk sets will likely be different, meaning the two models will be comparing cases to a different set of controls. This will result in different estimates of the coefficients.

We further note that under condition 2, baseline age is either a *precision variable* or a *nuisance variable*. We define a precision variable as a variable associated with the outcome-of-interest and uncorrelated with the predictor of interest. We note that precision variables should be adjusted for in the model because they reduce the variance of the estimated coefficients for the parameter of interest while not impacting that estimate. In the Cox model adjustment for precision variables will generally result in different coefficient estimates than when marginalizing over the precision variables. This is because the Cox model is generally non-collapsible [72, 73]. Further, if age-at-entry is a precision variables then the model using the age time scale is misspecified. Model misspecification in the Cox model results in estimating equations that are consistent for a quantity that depends on the censoring distribution [74]. With one model misspecified and the other correctly specified, the models will likely estimate different coefficient values. We define a nuisance variable as a variable unrelated to both the outcome-of-interest and the covariate-of-interest. Adjustment for nuisance variables should not bias the coefficient estimates for the covariate-of-interest because the estimated log hazard ratio for all subpopulations of the nuisance variable should be zero.

In the present chapter, we consider the choice of time scale in epidemiological studies where the outcome-of-interest is the time of dementia diagnosis. A common model for covariate adjustment in time-to-dementia diagnosis studies is the Cox Proportional Hazard model (Cox model) [18]. In general, the Cox model uses the time scale to order the events and determine risk sets at each event time [34]. The time of the events is not accounted for in the estimation of coefficients. Time-origins should be carefully considered in the design and analysis of time-to-event studies because they directly impact the results of analyses through defining the risk-sets [34, 35].

The choice of time scale for the Cox model has previously been assessed in simulation studies comparing three common specifications of the Cox model: the the age model, the left truncated age model, and time-on-study model. Let a_{0j} be the baseline age for subject j, let t_j be the observed time of dementia diagnosis or censoring for subject j, let a_j be the observed age of dementia diagnosis or censoring for subject j, and let x_j be the one-dimensional covariate value for subject j. Consider the following models:

M1) age model

$$\lambda_j^A(a \mid x_j) = \lambda_0^A(a_j)e^{\beta x_j},$$

M2) left truncation age model

$$\lambda_j^A(a \mid x_j, a_{0_j}) = \lambda_0^A(a_j \mid a_{0_j})e^{\beta x_j},$$

M3) time-on-study model

$$\lambda_j^T(t \mid x_j, a_{0_j}) = \lambda_0^T(t)e^{\beta x_j + \gamma a_{0_j}}.$$

Thiebaut and Benichou (2004) [35] consider simulation studies of the performance of M1 and M3 under the setting where the true time scale follows age from birth. They assess 1) Korn's second condition, and 2) the amount of bias for different degrees of association between covariates and baseline age when the time-on-study time scale is incorrectly used. Their results show some bias, but only for large effect sizes (i.e. $\beta \in \{\ln(5), \ln(10), \ln(50)\}$). Further, the report results from their simulations depict bias in the third decimal place for a log-hazard ratio of ln(5). They conclude that the age time scale should be chosen in general for epidemiological studies using the Cox model to study time-to-event outcomes.

In contradiction to Thiebaut and Benichou (2004) [35], Chalise et al. (2016) [71] consider the performance of all three models above in the settings of Korn's conditions. Further, Chalise et al. (2012) [34] and (2013) [38] consider the robustness of M1, M2, and M3 in the presence of time scale misspecification, and the predictive abilities of the models. The three Chalise papers consider the performance of the models under each time scale. In [34], they show that M3 out-performs M1 and M2 when time-on-study is the correct time scale. Further, when age is the correct time scale, they show that M3 performs about the same as M1 and M2. They conclude that the time-on-study model is more robust to use when the time scale is unknown or unclear. In [38], they study the bias of M1, M2, and M3, as well as their predictive ability (measured by concordance index) under each time scale. They conclude similarly to [34] on the bias aspect. For predictive ability, they observe that M3 outperforms M1 and M2 regardless of the true time scale. They again conclude that M3 is a more robust model to misspecification of the time scale and that M3 is the better model in settings where prediction of the outcome is of interest.

In this chapter, we note that these papers only consider the time-scale settings of time-on-study or age. However, in many late-life conditions, the risk of disease is likely not continuous and proportional from birth to old age. For example, in the setting of dementia, the risk of developing dementia is small for the first few decades for most subpopulations of people. The risk of dementia is generally negligible for the first several decades of life. We hypothesize that the true time scale for dementia lies somewhere between age and time-on-study. Under this framework, the true time origin for the time of dementia onset begins at some unknown age, $T_0 > 0$. In this chapter, we assess the performance of the three models presented above. Our simulations consider four cases with a single, binary predictor of interest. We consider the performance of each time scale when the predictor of interest is uncorrelated with age-at-entry and correlated with age-at-entry. For the age time scale, we simulate dementia onset to follow an age time scale, where risk of dementia onset is nearly 0 until age 50 and grew log-linearly in time after age 50. For the time-on-study time scale, we simulate dementia onset to grow log-linearly after some age T_0.

In Section 5.2, we provide a detailed description of the simulation settings and the results of the simulations. We end with a discussion of the results in section 5.3.

5.2 Simulation Studies

We run 10,000 simulations for 10,000 subjects under both age and time-on-study related time scales. Since our time scale assumes there is some age where risk for the disease starts, we set T_0 as the age at which risk starts to grow. In the age time scale, we let risk grow slowly from birth to T_0 by simulating from an Exponential distribution. After T_0 we simulate event times from a Weibull distribution. For the time-on-study time scale, we let risk start at T_0. Under each of these time scales, we consider the setting where 1) time-invariant age or age-at-entry is independent with a binary predictor of interest and 2) time-invariant age or age-at-entry is associated with a binary predictor of interest.

In all four simulations, we simulate entry age, a_0, from a Normal$(50, 5)$. We set $T_0 = 50$. We set the coefficient for the covariate of interest, X, as $\beta_x = \ln(1.5)$ and the coefficient for baseline age (where appropriate) as $\beta_{a_0} = \ln(1.3)$. We simulate event times from a Weibull distribution parameterized as $f(x) = \frac{\gamma}{\kappa}\left(\frac{x}{\kappa}\right)^{\gamma-1} exp\{-\left(\frac{x}{\kappa}\right)^{\gamma}\}$, with shape parameter of $\gamma = 4$ and scale parameter $\kappa = \lambda^{-1/\gamma}$, where λ is the hazard defined for each time scale. For all simulations, we denote the observed age of the event or censoring as $a = \min(a_E, a_C)$, where a_E is the age of the event and a_C is the age of censoring. Similarly, we denote the observed time of the event as $t_{obs} = \min(a_E - a_0, a_C - a_0)$. The time-invariant covariate of interest, X, we simulate from a Bernoulli(p), where $p \approx 0.5494$. For X associated with age, a, we simulate

$$X \sim \begin{cases} \text{Bernoulli}(0.3) & a_0 < 45 \\ \text{Bernoulli}(0.5) & 45 \le a_0 < 50 \\ \text{Bernoulli}(0.6) & 50 \le a_0 < 55 \\ \text{Bernoulli}(0.75) & a_0 \ge 55. \end{cases}$$

For the age time scale simulations, we simulate event times such that a_E from the Weibull distribution above with $\lambda = \lambda_0 e^{x\beta}$, where $\lambda_0 = 0.000000015$. If $a_E < 50$, then we simulate a_E from an exponential distribution parameterized as $f(x) = \psi e^{-\psi x}$, where $\psi = 0.0096 e^{x\beta_x}$. We simulate the censoring times for approximately 20% event rate such that $C \sim$ Uniform$(10, 16)$. Note that under these settings, $a_C = C + a_0$, $a = \min(a_E, a_C)$, and $t_{obs} = a - a_0$. For any cases where $a_E < a_0$, we remove those subjects as would occur in an epidemiological study with inclusion/exclusion criteria that requires subjects be disease-free upon screening.

For the time-on-study time scale simulations, we simulate the events such that for disease starts at T_0. T_E is simulated from the Weibull distribution (parameterized above) with $\lambda = \lambda_0 e^{x\beta_x + (a_0 - T_0)\beta_{a_0}}$, where $\lambda_0 = 0.000001$. Thus, $a_E = T_E + T_0$ and $t_E = T_E - (a_0 - T_0)$. To obtain approximately a 20% event rate, we simulate $C \sim \text{Uniform}(15, 25)$. Again, subjects are removed if $a_E < a_0$. Otherwise, $a = \min(a_E, a_0 + C)$ and $t_{obs} = \min(t_E, C)$.

Table 5.1 depicts the mean coefficient estimates from 10,000 simulations for models M1, M2, and M3 in the settings with time-invariant X. The top half of Table 5.1 shows the results when X is uncorrelated with a_0. Our results show little bias for all models regardless of the correct time scale. All three models are unbiased when age is the correct time scale. When time-on-study is the correct time scale, M3 has under 2% bias. Models M1 and M2 are both models about 5% biased, which is a deviation of about 0.02 in raw difference from the true coefficient value.

The bottom half of Table 5.1 shows the results when X is correlated with a_0. Our results are similar to Chalise et al (2012) [34], showing that M3 is more robust than M1 and M2. When age is the correct time scale, models M2 and M3 estimate the truth, while M1 is biased by about 4%, which is an unconcerning amount of bias. When time-on-study is the correct time scale, M3 has less than 2% bias while both of the age time scale models, M1 and M2, have 17% bias. The bias in M1 and M2 is likely due to residual confounding because those models do not adjust for a_0 in the estimating procedure.

5.3 Discussion

In this chapter, we assess the performance of the Cox model when fit with three common time scales in settings where those time scales are misspecified. Previous research by Thiebaut

X uncorrelated with a_0					
		Truth	M1	M2	M3
Correct Time Scale	Age	0.4055	0.4049 (0.1%)	0.4047 (0.2%)	0.4051 (0.1%)
	TOS	0.4055	0.3834 (5.4%)	0.3834 (5.4%)	0.4133 (1.9%)

X correlated with a_0					
		Truth	M1	M2	M3
Correct Time Scale	Age	0.4055	0.3886 (4.1%)	0.4045 (0.2%)	0.4053 ($<$ 0.1%)
	TOS	0.4055	0.4755 (17.3%)	0.4756 (17.3%)	0.4126 (1.7%)

Table 5.1: Simulation results from the simulation settings with time-invariant a predictor of interest, X. We present the mean coefficient estimates and percent bias from M1, M2, and M3 for each of the four simulation settings. M1 is the model that assumes age is the time scale with no adjustment for age-at-entry to the study. M2 is the model that assumes age is the time scale and conditioned on age-at-entry in the baseline hazard. M3 is the model that assumes time-on-study is the correct time scale with linear covariate adjustment for age-at-entry. TOS stands for time-on-study being the correct time scale.

and Benichou [35] assesses the performance of models M1 and M3 when the age time scale is correct. They observe that M3 is biased, although we would contend that the bias is minimal in the settings that are most realistic in practice. They also observe that M1, which is correctly-specified in their simulations, provides unbiased estimates of the hazard ratios. They conclude that M1 is the better model when the outcome is a lifetime disease. We contest that their simulations assume age is the correct time scale, but we believe this is rarely the case for late-life conditions. Dementia risk is extremely low for the first several decades of life, but grows quickly after age 65. Thus, the age time scale is not quite correct when studying dementia. We agree that the time-on-study time scale is rarely correct for many late-life conditions where there is not a clear exposure time, however, we do not agree that the age time scale is generally better in these settings.

Similar work by Chalise et al. (2012) [34] considers the robustness of models M1, M2, and M3 when the wrong time scale is selected for modeling. They observe that M3 is more robust to misspecification of the time scale. We consider the setting where the true time scale is based on some age where disease risk starts to grow. Similar to [34], our results suggest that M3 is more robust to this misspecification of the time scale. In the settings where disease-onset is possible over the full lifespan and the predictor of interest is time-invariant, the age time scale models (M1 and M2) estimate approximately the true value. However, when the risk for disease-onset is zero until some age, T_0, models M1 and M2 are about 17% biased in their estimates of the the log-hazard ratios. Conversely, the time-on-study model, M3, is fairly unbiased in all settings.

The choice of time scale is a fundamental aspect of time-to-event analyses. Our results suggest that the time-on-study model with covariate adjustment for age-at-entry is the more robust model when the exact time scale is unknown. This may be most useful in time-to-event analyses using electronic health records data, where the outcome is the onset of some lifetime disease. Our results support previous work by Chalise et al [34, 38] that the time-on-study model may provide more robust estimates of hazard ratios when the time scale is misspecified. This is likely due to the partial likelihood in the time-on-study model adjusting for the distribution of covariates conditioned on age via the age-at-entry covariate. Since the age time scale models do not include an age covariate in the model, age only contributes to ordering the events, but its distribution is not accounted for the partial likelihood.

In this analysis, we only considered a single, binary covariate of interest. It is possible that more complex patterns of relationships between multiple variables may change these results. It is further possible that time-varying covariates may also yield different results. These are areas of future work.

Chapter 6

Conclusion

This dissertation investigates the importance of careful consideration of both the implicit and explicit assumptions made in time-to-event analyses with electronic health records data. Patients of EHR data sets may have multiple options for seeking healthcare. The system migration that occurs when patients seek care from other providers can result in biased estimates of hazard ratios for the Cox model. We propose a multiple imputation method to estimate the probability of system migration among subjects and to adjust for the delayed diagnosis times that occur when patients migrate out of system. This method relies heavily on the ability of a prediction model to both identify system migration and to impute diagnosis times for patients identified as migrating out-of-system. To improve the prediction method, we develop a time-dependent receiver operating characteristic curve in the recurrent event survival setting. We observe that estimates of area under the curve depend on the number of events between subpopulations. To adjust for imbalance in the number of events within a subject between subpopulations, we propose a method for estimating the AUC that equally weights each subject's contribution to the AUC by dividing by their number of events. The methods from Chapters 3 and 4 account for potential bias when estimating associations between risk factors and the time-of-diagnosis. A fundamental aspect of survival analysis is the choice of time scale. Current literature on time scale selection, however, is inconclusive at best and oftentimes contradictory. We assess the robustness of common time scale modeling strategies for the Cox model under time scale misspecification. We observe that the time-on-study model is more robust than the age models when the time scale is neither exactly age nor time-on-study. In this final chapter, we provide some areas of future work to further develop these methods.

6.1 Future Work

Our development of the MA-Cox model assumes that subjects would be diagnosed with the condition of interest upon return to the system for which we have observed data. This may not be realistic in practice. Patients may be diagnosed in another system several visits before they receive a diagnosis in the system-at-hand. Further, even if subjects who do migrate out of system are not diagnosed at other facilities, if there are fundamental differences between subjects who migrate out of system and subjects who do not, the risk sets for the Cox model may not be accurate. Since the Cox model compares the covariate structure for the subject who has an event at time t to all the subjects in the risk set at time t, it is imperative to have accurate risk sets. Patients who are migrating out of system at time t may not be in the risk set at that time, thus they should not be part of the comparison. This may result in model misspecification, which Struthers and Kalbfleisch (1986) [75] show results in an estimating equation that is consistent for a quantity that depends on the censoring distribution. It has further been shown by Xu and O'Quigley (2000) [74], Boyd et al (2012) [76], and Nuño and Gillen (2021) [61] that reweighting the estimating equation by the censoring distributions can remove that dependence and result in robust estimates of the hazard ratios. We hypothesize that the estimating equation for the Cox model may depend on knowing who is migrating out of system. Our proposed solution is to reweight the estimating equation by the inverse probability of an individual migrating out of system. We hypothesize this will allow for robust estimation of the hazard ratios in this setting.

The proposed time-dependent AUC estimator for recurrent events could use future development. An analytical variance estimator and bootstrap variance estimator will be required to perform hypothesis tests comparing the difference between the predictive ability of two models. The method also weights each event within an individual the same. In some settings, it may be reasonable to reweight events in different manners, for example giving more weight to earlier or later events. Or perhaps a weighting structure that considers the autocorrelation of events within a subject. Further, Nuño and Gillen [68] proposed a censoring robust estimator of the AUC for single-event right-censored data. They show that the estimated AUCs depend on the censoring distribution when the model that developed the risk score was misspecified. It may be valuable to extend that estimator to the recurrent event setting.

Bibliography

[1] Kevin T. Ong. Challenges in dementia studies. In Jolanta Dorszewska and Wojciech Kozubski, editors, *Alzheimer's Disease*, chapter 7. IntechOpen, Rijeka, 2017.

[2] Barbara A Israel, Chris M Coombe, Rebecca R Cheezum, Amy J Schulz, Robert J McGranaghan, Richard Lichtenstein, Angela G Reyes, Jaye Clement, and Akosua Burris. Community-based participatory research: a capacity-building approach for policy advocacy aimed at eliminating health disparities. *Am J Public Health*, 100(11):2094–2102, Nov 2010.

[3] Native American Center for Excellence. Steps for conducting research and evaluation in native communities. Technical report, NACE, n.d.

[4] Felicia Schanche Hodge. No meaningful apology for american indian unethical research abuses. *Ethics & Behavior*, 22(6):431–444, 2012.

[5] Eric Whitney. Native americans feel invisible in u.s. health care system, Dec 2017.

[6] B Ashleigh Guadagnolo, Kristin Cina, Petra Helbig, Kevin Molloy, Mary Reiner, E Francis Cook, and Daniel G Petereit. Medical mistrust and less satisfaction with health care among native americans presenting for cancer treatment. *Journal of health care for the poor and underserved*, n/a(n/a):n/a, Feb 2009.

[7] Kathleen Thiede Call, Donna D McAlpine, Pamela Jo Johnson, Timothy J Beebe, James A McRae, and Yunjie Song. Barriers to care among american indians in public health care programs. *Medical Care*, n/a(n/a):n/a, Jun 2006.

[8] Emily A Haozous, Carolyn J Strickland, Janelle F Palacios, and Teshia G Arambula Solomon. Blood politics, ethnic identity, and racial misclassification among american indians and alaska natives. *Journal of environmental and public health*, n/a(n/a):n/a, Feb 2014.

[9] Benjamin A Goldstein, Nrupen A Bhavsar, Matthew Phelan, and Michael J Pencina. Controlling for informed presence bias due to the number of health encounters in an electronic health record. *American Journal of Epidemiology*, 184(11):847–855, Dec 2016.

[10] Benjamin A Goldstein, Ann Marie Navar, Michael J Pencina, and John P A Ioannidis. Opportunities and challenges in developing risk prediction models with electronic

health records data: a systematic review. *Journal of the American Medical Informatics Association*, 24(1):198–208, 2017.

[11] Benjamin S Glicksberg, Kipp W Johnson, and Joel T Dudley. The next generation of precision medicine: observational studies, electronic health records, biobanks and continuous monitoring. *Human Molecular Genetics*, 27(R1):R56–R62, Apr 2018.

[12] Robin A. Cohen, Michael E. Martinez, Amy E. Cha, and Emily P. Terlizzi. Health insurance coverage: Early release of estimates from the national health interview survey, january–june 2021. Technical report, National Center for Health Statistics, 2021.

[13] Indian health service. fact sheet, Indian Health Services, 2021.

[14] Indian Health Service. Eligibility, n.d.

[15] National patient information reporting system. fact sheet, Indian Health Services, n.d.

[16] Joan O'Connell, Soyeon Guh, Judith Ouellet, Jennifer Rockell, Yaqiang Li, Calvin Croy, and Margaret Gutilla. Arra action: Comparative effectiveness of health care delivery systems for american indians and alaska natives using enhanced data infrastructure. Technical report, Agency for Healthcare Research and Quality U.S. Department of Health and Human Services, 2014.

[17] Congressional Research Service. The indian health services (ihs): An overview. Technical report, Congressional Research Service, 2016.

[18] D. R. Cox. Regression Models and Life Tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220, 1972.

[19] John P. Klein and Melvin L. Moeschberger. *Survival analysis: techniques for censored andtruncated data.* Springer Science & Business Media, 2005.

[20] Roderick J.A. Little and Donald B. Rubin. *Statistical Analysis with Missing Data.* John Wiley & Sons, 3rd edition, 2019.

[21] Jelena Epping, Siegfried Geyer, and Juliane Tetzlaff. The effects of different lookback periods on the sociodemographic structure of the study population and on the estimation of incidence rates: analyses with german claims data. *BMC Medical Research Methodology*, 20(1):229, 2020.

[22] Yolanda Hagar, David Albers, Rimma Pivovarov, Herbert Chase, Vanja Dukic, and Noémie Elhadad. Survival analysis with electronic health record data: Experiments with chronic kidney disease. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 7(5):385–403, Oct 2014.

[23] David R Cox. Partial likelihood. *Biometrika*, 62(2):269–276, 1975.

[24] John D. Kalbfleisch and Ross L. Prentice. *The Statistical Analysis of Failure Time Data.* John Wiley & Sons, Inc., New Jersey, 2002.

[25] T. N. Herzog and D. B. Rubin. Using multiple imputations to handle nonresponse in sample surveys. *Incomplete Data in Sample Surveys*, 1983.

[26] S.J. Schieber. A comparison of three alternative techniques for allocating unreported social security income on the survey of the low-income aged and disabled. *Proceedings of the Survey Research Methods Section*, 1978.

[27] M.H. David, R.J. Little, M.E. Samuhel, and R.K. Triest. Alternative methods for cps income imputation. *Journal of the American Statistical Association*, 1986.

[28] Donald B. Rubin. *Multiple Imputation for Nonresponse in Surveys.* Wiley, 1987a.

[29] Margaret Pepe, Wendy Leisenring, and Carolyn Rutter. 12 evaluating diagnostic tests in public health. In *Bioenvironmental and Public Health Statistics*, volume 18 of *Handbook of Statistics*, pages 397–422. Elsevier, 2000.

[30] Patrick J. Heagerty, Thomas Lumley, and Margaret S. Pepe. Time-dependent roc curves for censored survival data and a diagnostic marker. *Biometrics*, 56(2):337–344, 2000.

[31] Patrick Heagerty and Yingye Zheng. Survival Model Predictive Accuracy and ROC Curves. *Biometrics*, 61:92–105, 2005.

[32] Lloyd E. Chambless and Guoqing Diao. Estimation of time-dependent area under the roc curve for long-term risk prediction. *Statistics in Medicine*, 25(20):3474–3486, 2006.

[33] NE Breslow and NE Day. The design and analysis of cohort studies. *Statistical methods in cancer research*, 2(82):1–46, 1987.

[34] Prabhakar Chalise, Eric Chicken, and Daniel McGee. Baseline age effect on parameter estimates in cox models. *Journal of Statistical Computation and Simulation*, 82(12):1767–1774, 2012.

[35] Anne C. M. Thiébaut and Jacques Bénichou. Choice of time-scale in cox's model analysis of epidemiologic cohort data: a simulation study. *Statistics in Medicine*, 23(24):3803–3820, 2004.

[36] E L Korn, B I Graubard, and D Midthune. Time-to-event analysis of longitudinal follow-up of a survey: choice of the time-scale. *Am J Epidemiol*, 145(1):72–80, Jan 1997.

[37] Michael J. Pencina, Martin G. Larson, and Ralph B. D'Agostino. Choice of time scale and its effect on significance of predictors in longitudinal studies. *Statistics in Medicine*, 26(6):1343–1359, 2007.

[38] Prabhakar Chalise, Eric Chicken, and Daniel McGee. Performance and prediction for varying survival time scales. *Communications in Statistics - Simulation and Computation*, 42(3):636–649, 2013.

[39] K. Bruce Bayley, Tom Belnap, Lucy Savitz, Andrew L. Masica, Nilay Shah, and Neil S. Fleming. Challenges in using electronic health record data for cer. *Medical Care*, 51(Supplement 8Suppl 3):S80–S86, Aug 2013.

[40] Luohua Jiang, Xiaoyi Niu, Laura Grau, Maria M Corrada, Spero Manson, and Joan O'Connell. Accuracy in estimating prevalence and incidence of dementia using longitudinal electronic health record data from the indian health service. *Alzheimer's & Dementia*, 17(S10):e056279, 2021.

[41] Xiaoyi Niu, Jenny Chang, Maria M. Corrada, Ann Bullock, Spero Manson, Joan O'Connell, and Luohua Jiang. The relationship between all-cause dementia and acute diabetes complications among american indian and alaska native peoples. *Alzheimer's & Dementia*, 18(S11):e064867, 2022.

[42] Joan O'Connell, Jennifer Rockell, Judith C. Ouellet, and Mark LeBeau. Disparities in potentially preventable hospitalizations between american indian and alaska native and non-hispanic white medicare enrollees. *Medical Care*, 55(6), 2017.

[43] Joan O'Connell, Laura Grau, Turner Goins, Marcelo Perraillon, Blythe Winchester, Maria Corrada, Spero M. Manson, and Luohua Jiang. The costs of treating all-cause dementia among american indians and alaska native adults who access services through the indian health service and tribal health programs. *Alzheimer's & Dementia*, 18(11):2055–2066, 2022.

[44] Indian Health Services. The indian health service (ihs): An overview. fact sheet, Indian Health Services, Jan 2016.

[45] C. Gordon Law and Ron Brookmeyer. Effects of mid-point imputation on the analysis of doubly censored data. *Statistics in Medicine*, 11(12):1569–1578, 1992.

[46] Wei Pan. A multiple imputation approach to cox regression with interval-censored data. *Biometrics*, 56(1):199–203, 2000.

[47] D. B. Rubin. *Multiple Imputation for Nonresponse in Surveys*. Wiley, 1987.

[48] Kaifeng Lu and Anastasios A. Tsiatis. Multiple imputation methods for estimating regression coefficients in the competing risks model with missing cause of failure. *Biometrics*, 57(4):1191–1197, 2001.

[49] P. K. Andersen and R. D. Gill. Cox's regression model for counting processes: A large sample study. *The Annals of Statistics*, 10(4):1100 – 1120, 1982.

[50] Anastasios A. Tsiatis. A Large Sample Study of Cox's Regression Model. *The Annals of Statistics*, 9(1):93 – 108, 1981.

[51] R Turner Goins, Blythe Winchester, Luohua Jiang, Laura Grau, Maggie Reid, Maria M Corrada, Spero M Manson, and Joan O'Connell. Cardiometabolic Conditions and All-Cause Dementia Among American Indian and Alaska Native People. *The Journals of Gerontology: Series A*, 77(2):323–330, 04 2021.

[52] Yosef Zenebe, Baye Akele, Mulugeta W/Selassie, and Mogesie Necho. Prevalence and determinants of depression among old age: a systematic review and meta-analysis. *Annals of General Psychiatry*, 20(1):55, 2021.

[53] Center for Disease Control and Prevention. Summary health statistics: National health interview survey, 2018.

[54] Center for Disease Control and Prevention. Therapeutic drug use, 2019.

[55] Sujuan Gao, Hugh C. Hendrie, Kathleen S. Hall, and Siu Hui. The Relationships Between Age, Sex, and the Incidence of Dementia and Alzheimer Disease: A Meta-analysis. *Archives of General Psychiatry*, 55(9):809–815, 09 1998.

[56] T F Hughes, A R Borenstein, E Schofield, Y Wu, and E B Larson. Association between late-life body mass index and dementia: The kame project. *Neurology*, 72(20):1741–1746, May 2009.

[57] I K Wium-Andersen, J Rungby, M B Jørgensen, A Sandbæk, M Osler, and M K Wium-Andersen. Risk of dementia and cognitive dysfunction in individuals with diabetes or elevated blood glucose. *Epidemiol Psychiatr Sci*, 29:e43, Aug 2019.

[58] Liu Yang, Yue-Ting Deng, Yue Leng, Ya-Nan Ou, Yu-Zhu Li, Shi-Dong Chen, Xiao-Yu He, Bang-Sheng Wu, Shu-Yi Huang, Ya-Ru Zhang, Kevin Kuo, Wei Feng, Qiang Dong, Jian-Feng Feng, John Suckling, A. David Smith, Fei Li, Wei Cheng, and Jin-Tai Yu. Depression, depression treatments, and risk of incident dementia: A prospective cohort study of 354,313 participants. *Biological Psychiatry*, 93(9):802–809, 2023.

[59] Indian health services profile: Fact sheets. fact sheet, Indian Health Services, Aug 2020.

[60] Center for Disease Control and Prevention. About adult bmi, 2022.

[61] Michelle M. Nuño and Daniel L. Gillen. Robust estimation in the nested case-control design under a misspecified covariate functional form. *Statistics in Medicine*, 40(2):299–311, 2021.

[62] Gregory Campbell. Advances in statistical methodology for the evaluation of diagnostic and laboratory tests. *Statistics in Medicine*, 13(5-7):499–508, 1994.

[63] National Alzheimer's Coordinating Center. National alzheimer's coordinating center uniform data set researchers data dictionary, 2015.

[64] Mary S Fewtrell, Kathy Kennedy, Atul Singhal, Richard M Martin, Andy Ness, Mijna Hadders-Algra, Berthold Koletzko, and Alan Lucas. How much loss to follow-up is acceptable in long-term randomised trials and prospective studies? *Archives of disease in childhood*, 93(6):458–461, Jun 2008.

[65] Kyle R. Conniff, Joshua D. Grill, and Daniel L. Gillen. Retention of american indian and alaska native participants in the national alzheimer's coordinating center uniform data set. *Alzheimer's & Dementia*, n/a(n/a).

[66] National Alzheimer's Coordinating Center. About nacc data.

[67] R. L. Prentice, B. J. Williams, and A. V. Peterson. On the regression analysis of multivariate failure time data. *Biometrika*, 68(2):373–379, 08 1981.

[68] Michelle M. Nuño and Daniel L. Gillen. Censoring-robust time-dependent receiver operating characteristic curve estimators. *Statistics in Medicine*, 40(30):6885–6899, 2021.

[69] Center for Disease Control and Prevention. About dementia, 2019.

[70] Alzheimer's Association. Down syndrome and alzheimer's disease, n.d.

[71] Prabhakar Chalise, Eric Chicken, and Daniel McGee. Time scales in epidemiological analysis: An empirical comparison. *International Journal of Statistics and Probability*, 2016.

[72] Brian W Whitcomb and Ashley I Naimi. Defining, quantifying, and interpreting "noncollapsibility" in epidemiologic studies of measures of "effect". *Am J Epidemiol*, 190(5):697–700, May 2021.

[73] Sven Ove Samuelsen. Cox regression can be collapsible and aalen regression can be non-collapsible. *Lifetime Data Anal*, 29(2):403–419, Apr 2023.

[74] Ronghui Xu and John O'Quigley. Estimating average regression effect under non-proportional hazards. *Biostatistics*, 1(4):423–439, 12 2000.

[75] C. A. Struthers and J. D. Kalbfleisch. Misspecified proportional hazard models. *Biometrika*, 73(2):363–369, 08 1986.

[76] Adam P. Boyd, John M. Kittelson, and Daniel L. Gillen. Estimation of treatment effect under non-proportional hazards and conditionally independent censoring. *Statistics in Medicine*, 31(28):3504–3515, 2012.

[77] Duane L. Beekly, Erin M. Ramos, William W. Lee, Woodrow D. Deitrich, Mary E. Jacka, Joylee Wu, Janene L. Hubbard, Thomas D. Koepsell, John C. Morris, and Walter A. Kukull. The national alzheimer's coordinating center (nacc) database: The uniform data set. *Alzheimer disease and associated disorders.*, 21(3):249–258, 2007.

[78] National Congress of American Indians. Tribal ownership of health-related data, 2005.

[79] National Congress of American Indians. Support of us indigenous data sovereignty and inclusion of tribes in the development of tribal data governance principles, 2018.

[80] Stephanie Russo Carroll, Ibrahim Garba, Oscar L. Figueroa-Rodríguez, Jarita Holbrook, Raymond Lovett, Simeon Materechera, Mark Parsons, Kay Raseroka, Desi Rodriguez-Lonebear, Robyn Rowe, Rodrigo Sara, Jennifer D. Walker, Jane Anderson, and Maui Hudson. The care principles for indigenous data governance. *Data Science Journal*, 19, 2020.

[81] First Nations Information Governance Centre. Ownership, control, access and possession (ocap™): The path to first nations information governance, 2014.

[82] Administration on Community Living. 2020 profile of american indians and alaska natives age 65 and older. Technical report, U.S. Department of Health and Human Services, 2020.

[83] Centers for Disease Control and Prevention. U.s. burder of alzheimer's disease, related dementias to double by 2060. Report, Centers for Disease Control and Prevention, 2018.

[84] Alzheimer's Association. Native americans and alzheimer's, 2021.

[85] Kevin A. Matthews, Wei Xu, Anne H. Gaglioti, James B. Holt, Janet B. Croft, Dominic Mack, and Lisa C. Mcguire. Racial and ethnic estimates of alzheimer's disease and related dementias in the united states (2015–2060) in adults aged 65 years. *Alzheimer's Dementia*, 15(1):17–24, 2019. OA status: bronze.

[86] Disparities: Fact sheets, Oct 2019.

[87] Elizabeth Rose Mayeda, M.Maria Glymour, Charles P. Quesenberry, and Rachel A. Whitmer. Inequalities in dementia incidence between six racial and ethnic groups over 14 years. *Alzheimer's & Dementia*, 12(3):216–224, 2016.

[88] Deionna Vigil, Ninet Sinaii, and Barbara Karp. American indian and alaska native enrollment in clinical studies in the national institutes of health's intramural research program. *Ethics Human Research*, 43(3):2–9, 2021.

[89] Linda Tuhiwai Smith. *Decolonizing methodologies : research and indigenous peoples / Linda Tuhiwai Smith.* Zed Books, London [England, third edition. edition, 2021. Includes bibliographical references and index.

[90] Andrea L. Gilmore-Bykovskyi, Yuanyuan Jin, Carey Gleason, Susan Flowers-Benton, Laura M. Block, Peggye Dilworth-Anderson, Lisa L. Barnes, Manish N. Shah, and Megan Zuelsdorff. Recruitment and retention of underrepresented populations in alzheimer's disease research: A systematic review. *Alzheimer's Dementia: Translational Research Clinical Interventions*, 5(1):751–770, 2019.

[91] Richard E. Kennedy, Gary R. Cutter, Guoqiao Wang, and Lon S. Schneider. Challenging assumptions about african american participation in alzheimer disease trials. *The American Journal of Geriatric Psychiatry*, 25(10):1150–1159, 2017.

[92] Joshua D. Grill, Jimmy Kwon, Merilee A. Teylan, Aimee Pierce, Eric D. Vidoni, Jeffrey M. Burns, Allison Lindauer, Joseph Quinn, Jeff Kaye, Daniel L. Gillen, and Bin Nan. Retention of alzheimer disease research participants. *Alzheimer disease and associated disorders.*, 33(4):299–306, 2019.

[93] David Nunan, Jeffrey Aronson, and Clare Bankhead. Catalogue of bias: attrition bias. *BMJ Evidence-Based Medicine*, 23(1):21–22, 2018.

[94] Miriam T. Ashford, Joseph Eichenbaum, Tirzah Williams, Monica R. Camacho, Juliet Fockler, Aaron Ulbricht, Derek Flenniken, Diana Truran, R. Scott Mackin, Michael W. Weiner, and Rachel L. Nosheny. Effects of sex, race, ethnicity, and education on online aging research participation. *Alzheimer's Dementia: Translational Research Clinical Interventions*, 6(1), 2020.

[95] Sarah M. Hatcher, Christine Agnew-Brune, Mark Anderson, Laura D. Zambrano, Charles E. Rose, Melissa A. Jim, Amy Baugher, Grace S. Liu, Sadhna V. Patel, Mary E. Evans, Talia Pindyck, Christine L. Dubray, Jeanette J. Rainey, Jessica Chen, Claire Sadowski, Kathryn Winglee, Ana Penman-Aguilar, Amruta Dixit, Eudora Claw, Carolyn Parshall, Ellen Provost, Aurimar Ayala, German Gonzalez, Jamie Ritchey, Jonathan Davis, Victoria Warren-Mears, Sujata Joshi, Thomas Weiser, Abigail Echo-Hawk, Adrian Dominguez, Amy Poel, Christy Duke, Imani Ransby, Andria Apostolou, and Jeffrey Mccollum. Covid-19 among american indian and alaska native persons — 23 states, january 31–july 3, 2020. *MMWR. Morbidity and Mortality Weekly Report*, 69(34):1166–1169, 2020.

[96] Fesani Mahmood, Dev Acharya, Kanta Kumar, and Vibhu Paudyal. Impact of covid-19 pandemic on ethnic minority communities: a qualitative study on the perspectives of ethnic minority community leaders. *BMJ Open*, 11(10):e050584, 2021.

[97] Brad Boserup, Mark Mckenney, and Adel Elkbuli. Disproportionate impact of covid-19 pandemic on racial and ethnic minorities. *The American Surgeon*, 86(12):1615–1622, 2020.

[98] Vicki Kristman, Michael Manno, and Pierre Côté. Loss to follow-up in cohort studies: How much is too much? *European Journal of Epidemiology*, 19(8):751–760, 2003.

[99] Martha Abshire, Victor D. Dinglas, Maan Isabella A. Cajita, Michelle N. Eakin, Dale M. Needham, and Cheryl Dennison Himmelfarb. Participant retention practices in longitudinal clinical research studies with high retention rates. *BMC Medical Research Methodology*, 17(1), 2017.

[100] Paul Hindmarch, Adrian Hawkins, Elaine Mccoll, Mike Hayes, Gosia Majsak-Newman, Joanne Ablewhite, Toity Deave, and Denise Kendrick. Recruitment and retention strategies and the examination of attrition bias in a randomised controlled trial in children's centres serving families in disadvantaged areas of england. *Trials*, 16(1):79, 2015.

[101] Marwan N. Sabbagh, Nancy Thompson, Deborah Tweedy, Suhair Stipho-Majeed, Claudia Kawas, and Donald J. Connor. Recruitment and retention strategies for clinical trials in alzheimer's disease. *Pharmaceutical Development and Regulation*, 1(4):269–276, 2003.

[102] Samantha Teague, George J. Youssef, Jacqui A. Macdonald, Emma Sciberras, Adrian Shatte, Matthew Fuller-Tyszkiewicz, Chris Greenwood, Jennifer Mcintosh, Craig A. Olsson, and Delyse Hutchinson. Retention strategies in longitudinal cohort studies:

a systematic review and meta-analysis. *BMC Medical Research Methodology*, 18(1), 2018.

[103] Antronette K. Yancey, Alexander N. Ortega, and Shiriki K. Kumanyika. Effective recruitment and retention of minority research participants. *Annual Review of Public Health*, 27(1):1–28, 2006.

[104] Cathleen M. Connell, Benjamin A. Shaw, Sara B. Holmes, and Norman L. Foster. Caregivers' attitudes toward their family members' participation in alzheimer disease research: Implications for recruitment and retention. *Alzheimer disease and associated disorders.*, 15(3):137–145, 2001.

[105] Patricia A. Areán, Jennifer Alvidrez, Rowena Nery, Carroll Estes, Karen Linkins, and Washington Dc Gerontological Society. Recruitment and retention of older minorities in mental health services research. *The Gerontologist*, 43(1):36–44, 2003.

[106] Mary Anne Gauthier and Willie Pearl Clarke. Gaining and sustaining minority participation in longitudinal research projects. *Alzheimer Disease Associated Disorders*, 13:S29–33, 1999.

[107] Nancy A. Hessol, Michael Schneider, Ruth M. Greenblatt, Melanie Bacon, Yvonne Barranday, Susan Holman, Esther Robison, Carolyn Williams, Mardge Cohen, and Kathleen Weber. Retention of women enrolled in a prospective study of human immunodeficiency virus infection: Impact of race, unstable housing, and use of human immunodeficiency virus therapy. *American Journal of Epidemiology*, 154(6):563–573, 2001.

[108] Deborah Parra-Medina, Angela D'Antonio, Sharon M. Smith, Sarah Levin, Gregory Kirkner, and Elizabeth Mayer-Davis. Successful recruitment and retention strategies for a randomized weight management trial for people with diabetes living in rural, medically underserved counties of south carolina: the power study. *Journal of the American Dietetic Association*, 104(1):70–75, 2004.

[109] Susan L. Janson, Maria Elena Alioto, and Homer A. Boushey. Attrition and retention of ethnically diverse subjects in a multicenter randomized controlled research trial. *Controlled Clinical Trials*, 22(6, Supplement 1):S236–S243, 2001.

[110] Pavneet Singh, Twyla Ens, K. Alix Hayden, Shane Sinclair, Pam LeBlanc, Moaz Chohan, and Kathryn M. King-Shier. Retention of ethnic participants in longitudinal studies. *Journal of Immigrant and Minority Health*, 20(4):1011–1024, 2018.

[111] Dorothy Burns, April C. M. Soward, Anne H. Skelly, Jennifer Leeman, and John Carlson. Effective recruitment and retention strategies for older members of rural minorities. *The Diabetes Educator*, 34(6):1045–1052, 2008.

[112] Christian R. Salazar, Marina Ritchie, Daniel L. Gillen, and Joshua D. Grill. Strategies associated with retaining participants in the longitudinal national alzheimer's coordinating center uniform data set study. *Journal of Alzheimer's Disease*, 87(4):1557–1566, 2022.

[113] H. Richard Milner. Race, culture, and researcher positionality: Working through dangers seen, unseen, and unforeseen. *Educational Researcher*, 36(7):388–400, 2007.

[114] Peggye Dilworth-Anderson and Sharon Wallace Williams. Recruitment and retention strategies for longitudinal african american caregiving research. *Journal of Aging and Health*, 16(5):137S–156S, 2004.

[115] Diana Redwood, J. Leston, Elvin Asay, Elizabeth Ferucci, Ruth Etzel, and Anne Lanier. Strategies for successful retention of alaska native and american indian study participants. 32:43–52, 2010.

[116] D. Gallagher-Thompson, N. Solano, D. Coon, P. Arean, and Washington Dc Gerontological Society. Recruitment and retention of latino dementia family caregivers in intervention research: Issues to face, lessons to learn. *The Gerontologist*, 43(1):45–51, 2003.

[117] Yaron G. Rabinowitz and Dolores Gallagher-Thompson. Recruitment and retention of ethnic minority elders into clinical research. *Alzheimer Disease Associated Disorders*, 24:S35–S41, 2010.

[118] Barbara A. Israel, Amy J. Schulz, Edith A. Parker, and Adam B. Becker. Review of community-based research: Assessing partnership approaches to improve public health. *Annual Review of Public Health*, 19(1):173–202, 1998.

[119] Shana D. Stites, R. Scott Turner, Jeanine Gill, Anna Gurian, Jason Karlawish, and Joshua D. Grill. Research attitudes questionnaire scores predict alzheimer's disease clinical trial dropout. *Clinical Trials*, 18(2):237–244, 2021.

[120] Meghan Jernigan, Amanda D Boyd, Carolyn Noonan, and Dedra Buchwald. Alzheimer's disease knowledge among american indians and alaska natives. *Alzheimers Dement (N Y)*, 6(1):e12101, 2020.

[121] Cenders for Disease Control and Prevention. Smoking: Know the facts, 2022.

Appendix A

Retention of American Indian and Alaska Native Participants in the National Alzheimer's Coordinating Center Uniform Data Set

A.1 Abstract

INTRODUCTION: The number of American Indian and Alaska Native (AI/AN) Elders is expected to double by 2060. Thus, it is imperative to retain AI/AN participants in longitudinal research studies to identify novel risk factors and potential targets for intervention for Alzheimer's disease and related dementias in these communities. METHODS: The National Alzheimer's Coordinating Center houses uniformly collected longitudinal data from the network of NIA-funded Alzheimer's Disease Research Centers (ADRCs). We used logistic regression to quantify participant retention at 43 ADRCs, comparing self-identified AI/AN

participants to non-Hispanic White (NHW) participants, adjusting for potential confounding factors including baseline diagnosis, age, sex, education, and smoking. RESULTS: The odds of AI/AN participant retention at the first follow-up visit were significantly lower than those for NHW participants ($aOR : 0.599; 95\% : 0.46 - 0.78; p < 0.001$). DISCUSSION: These results suggest the need for improved strategies to retain AI/AN participants, perhaps including improved researcher-community relationships and community engagement and education.

A.2 Acknowledgement of Data Sovereignty

Acknowledgement of Data Sovereignty As sovereign entities, tribal nations have the right to govern the collection, storage, ownership, application, and dissemination of data collected from members of their nation. Tribal nation members should therefore have input on interpretation of data analyses that include American Indian and Alaska Native (AI/AN) participants to ensure respect for individuals and/or their ancestors and that the research benefits the nation. Data collection by Alzheimer's Disease Research Centers (ADRCs) consists of two parts: 1) center-specific collection protocols, and 2) standardized data collection to be contributed to the National Alzheimer's Coordinating Center (NACC) [77], and includes questions about primary, secondary, and tertiary race and allows participants to self-identify their race. There are no mandatory follow-up questions on tribal affiliation. Additionally, there are no limitations to AI/AN data access, nor safeguards to ensure research conducted using AI/AN data is of benefit to AI/AN communities. In the broadly available NACC data, there are no pathways to ensure results involving AI/AN communities are interpreted correctly and disseminated to the appropriate communities. While individual ADRCs may be working with the local AI/AN communities, there is no oversight or guidance from NACC or the National Institute on Aging to ensure and aid ADRCs in acknowledging the inherent right of data sovereignty for AI/AN communities. We acknowledge that the present research in its current form is not directly aiding AI/AN communities. We withheld some sensitivity analyses due to the risk of loss of confidentiality among nations and peoples who have not had the opportunity to review or provide input on the current work. It is our hope that ADRCs are aware of the priorities, wishes, and goals of AI/AN participants and are working directly with the tribal nations to achieve their research goals. For more information, please see: [78, 79, 80, 81].

A.3 Introduction

The U.S. Census Bureau estimates that the American Indian and Alaska Native (AI/AN) population of adults aged 65 and over will nearly triple between 2016 to 2060.[82] Accordingly, estimates from the Centers for Disease Control and Prevention (CDC), Alzheimer's Association, and academic investigators are that the incidence of ADRD will increase four-to-five fold over the same time frame in this group [83, 84, 85]. AI/AN populations already experience disparities in many health conditions [86], including having the second highest dementia incidence rate among six racial and ethnic groups examined in one assessment of a large healthcare system [87]. Despite this, AI/AN patients are rarely recruited into ADRD research and are frequently grouped into an "Other" race/ethnicity category in research analyses. More specifically, AI/AN populations are especially underrepresented in clinical trials across the range of National Institutes of Health funded intervention studies [88].

Barriers to AI/AN representation in research are numerous. Access to academic medical centers is limited by geography and sociopolitical constraints. Historical actions as well as research abuses [4], have created a setting where the field of research and even the word "research" are not viewed favorably in AI/AN communities [89]. Yet, little work has quantified recruitment and retention of AI/AN research participants.

To reduce the disproportionate burden of ADRD among AI/AN individuals, research must be inclusive of AI/AN communities. A systematic review of 22 identified studies that either 1) examined strategies for recruitment and retention of underrepresented populations in ADRD research or 2) reported on underrepresented participants' attitudes toward ADRD research, found that none focused on or included AI/AN populations [90]. Enrollment is, however, only one element of inclusive research. Kennedy et al, for example, analyzed 18 studies including clinical trials and observational studies and found that non-Hispanic Black participants had higher rates of attrition than did non-Hispanic White (NHW) participants [91]. Failing to retain participants throughout the course of a study can lead to decreased precision, questionable validity, and lack of generalizability of results [92, 93]. Disproportionate loss to follow-up in specific groups may lead to low precision or biased results for those groups.

We assessed retention of AI/AN participants in ADRD research by comparing visitation patterns across racial and ethnic groups in the National Alzheimer's Coordinating Center (NACC) Uniform Data Set (UDS). The NACC UDS consists of longitudinal data collected at Alzheimer's Disease Research Centers (ADRCs) across the US. Enrolled participants are expected to have annual follow-up visits in which a battery of clinical and cognitive assessments are used to arrive at a diagnostic status and to track performance over time. A major goal is to identify patterns of and risk for disease progression, making it an ideal data set for analyzing differences in participant retention. We hypothesized that, like other minority racial and ethnic groups [91, 94], AI/AN patients would have lower odds of study retention when compared to NHW participants.

A.4 Methods

A.4.1 Study Population:

The NACC UDS consists of longitudinal demographic, clinical, neuropsychological, and diagnostic data on participants enrolled in National Institute on Aging (NIA)-funded ADRCs. Each center recruits participants according to their own protocol. Recruitment methods include clinical, family, and self-referrals, as well as community outreach and active recruitment. Conversely, core information on cognition, demographics, and participant health status is collected uniformly across all centers from participants and study partners directly on UDS forms with standardized evaluation by trained clinicians and clinic personnel. Enrolling with a study partner (e.g., spouse or adult child) is a requirement for participation as they attend visits and complete informant-based assessments.

The current analysis utilized data from 43 ADRCs collected between September 2005 and November 2021. Enrolled participants' initial diagnostic status included normal cognition, impaired but not mild cognitive impairment, mild cognitive impairment (MCI), and dementia. Diagnoses were made by a single expert physician or a clinical team consensus, depending on site-specific ADRC protocols. Annual follow-up appointments generally occurred via in-person office visits. In response to the COVID-19 pandemic, additional options such as telephone and zoom visits were used to collect participant data. As part of the UDS, Milestone Forms [66] were collected to record participant dropout and death, as well as other major life changes (e.g., moving to a nursing home).

We classified participants into racial and ethnic categories based upon self-reported information and the National Institutes of Health (NIH) definitions. Specifically, the NIH defines a person to be AI/AN if that is the only reported race. Selecting any race in addition to AI/AN categorizes that individual as "multiple races." To define our six race and ethnic groups, we first assigned participants based on their reported race with the categories AI/AN, Asian, Black, White, and Other/Multiple Races. We then distinguished Hispanic from non-Hispanic individuals for the White race category to create our final groups: AI/AN, Asian, Black, Hispanic White (HW), non-Hispanic White, and Other/Multiple Races. The "Other/Multiple Race" group consisted of 35 "Native Hawaiian or Pacific Islander" participants, 762 "Unknown or Ambiguous" participants, 1392 "Multiracial" participants, and 128 White participants of "Unknown" Hispanic ethnicities.

Other covariates considered in our analysis were baseline diagnostic status, baseline age, binary sex, years of education, and smoking status. Baseline diagnostic status included "Normal cognition", "Impaired-not-MCI", "MCI", and "Dementia." Baseline age refers to the age of a participant on their initial visit. Binary sex refers to the participant's indication of being either male or female (no other options were available). Education categories were formed by discretizing "years of education" into: "Less than high school" for fewer than 12 years of education, "high school diploma/GED" for 12 years of education, "some college" for 13-15 years of education, "4-year degree" for 16 years of education, and "greater than 4-year degree" for greater than 16 years of education. Smoking status (never, previous, current) was created from participant self-reported answers to four questions: 1) at what age did the participant quit smoking, 2) total number of years the participant smoked, 3) has the participant smoked in the last 30 days, and 4) the average number of packs the participant smoked per day. Participants were assigned smoking status in the order of "never", "previous", "current", and "unknown." Participants were considered to have never smoked if they answered as never smoking or answered all questions as "Not applicable." Participants were considered to have previously smoked if they had a quit age, or an unknown quit age and did not smoke at the time, or an unknown quit age and a non-zero number of years as a smoker. Participants were considered to currently smoke if they had smoked within the last 30 days and had a non-zero number of packs smoked per day, or have an unknown answer to questions 1, 2, and 4. The rest were considered "unknown."

A.4.2 Statistical Methods:

We assessed the retention of AI/AN participants in two ways: (1) the odds of retention at the first scheduled UDS follow-up and (2) the odds of retention at the next scheduled UDS follow-up visit having completed all previous follow-up visits, as defined per protocol. The first of our two analyses sought to determine if the odds of retention among AI/AN participants differed from that of NHW participants. We hypothesized that confounder-adjusted retention among AI/AN participants would be lower than that of NHW participants.

Since the ADRCs encourage annual appointments, the NACC defines participant retention as returning for a visit within 15 months of the previous visit's date. Choosing a more conservative window, we specified the retention vs. dropout cut-off as 18 months. Under this definition in analysis (1), we considered a participant as retained if they attended a second visit within 18 months of their baseline visit. A participant who failed to return or returned for their first follow-up any time after month 18 was considered a dropout. In analysis (2), we counted a participant as retained if they had completed all previous visits within 18 months of the preceding visit. This definition means we considered a participant with a baseline visit and five annual follow-up visits followed by a 19-month gap (or greater) before the sixth follow-up visit to have been retained for five visits. Thus, there are some participants that we considered a dropout per protocol, despite returning for further visits. In both settings, we removed all participants that were not expected to return (e.g., participants enrolled as initial visit only) from the analysis sample. Further, in analysis (1), we only considered participants with at least 18 months of follow-up. In analysis (2), we censored participants at the minimum of time to death or end-of-follow-up.

To estimate the odds of retention at the first follow-up visit, we used logistic regression. To estimate the relative odds of retention at subsequent follow-up visits we used a continuation ratio model [24] with the timescale being number of visits (i.e., annual follow-up appointments attended). This model estimates the relative odds of completing a visit, conditional upon completing all prior visits. In both analyses, we adjusted for a priori hypothesized potential confounding factors. A covariate was a priori hypothesized to be a potential confounder if it was reasonably believed to be related to the probability of retention and related to race and/or ethnicity. Adjustment covariates included in both regression models included baseline diagnostic status, baseline age, binary sex, education categories, and smoking status as shown in Figure A.1. For all analyses we present point estimates along with corresponding 95% Wald-based confidence intervals and p-values for the test of a null association. We assessed influence via Cook's distance. No individual points had outstanding influence compared to the others and hence no data were removed from our analyses. All analyses were performed using R Statistical Software (v4.1.0; R Core Team 2021).

We observed relatively small amounts of missing data. Age, sex, race, and cognitive status were collected completely. Educational status was missing for 332 participants and smoking status was missing for an additional 676 participants for a total of 995 out of the 39,290 participants in the study (2.5%). Further, only 8 of those 995 identified as AI/AN. Due to the small number of missing values, we conducted a complete case analysis for both aims.

We performed four sensitivity analyses to account for potential differential effects of the COVID-19 pandemic and to assess the potential for site-specific effects. To ensure our definition of retention did not influence results, we conducted a sensitivity analysis where we did not censor participants that were seen more than 18 months after their prior scheduled visit (Supplementary Materials Section 1.1 (SM1.1)). To assess if the pandemic differentially impacted follow-up across race and ethnicity groups, we re-fit all models with a study end date of February 2020 (SM1.2). To assess potential site effects of retention, we repeated our analyses with only the sites that had any (¿0) AI/AN participants and with only the sites that had at least 10 AI/AN participants (SM1.3). We assessed potential effect modification over time by splitting the cohort at the midpoint of the total observation period (pre-2013 vs. post-2013).

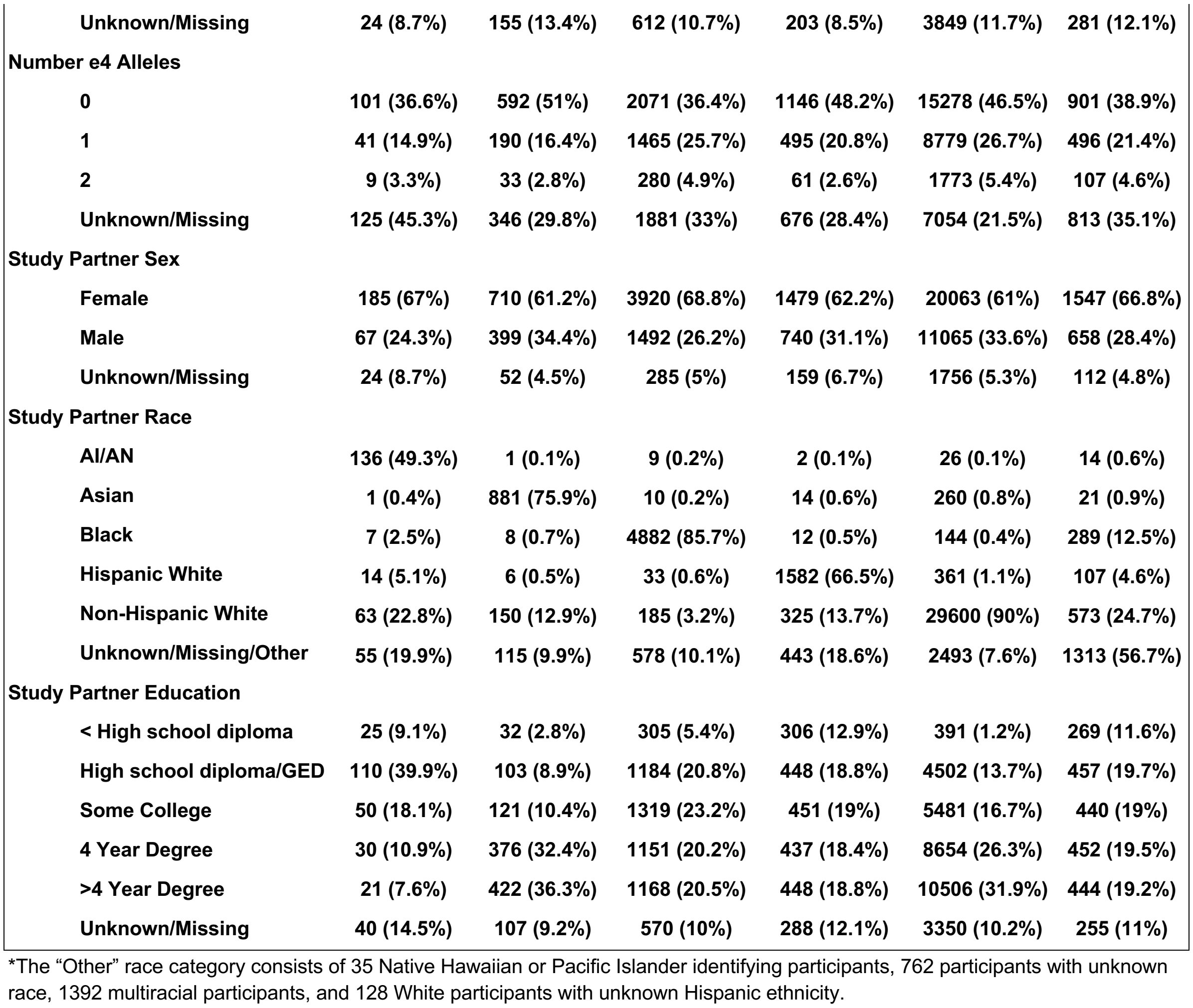
A.5 Results

Figure A.1 describes the study sample. Descriptive statistics revealed some differences among the racial and ethnic groups at their baseline visit. Notably, AI/AN participants accounted for just 0.6% of participants. The other groups ranged from Asian participants (accounting for 2.6% of the sample) to NHW participants (accounting for 73.5% of the sample). AI/AN participants were observed to have the lowest level of formal education, with 60.1% of AI/AN participants self-reporting 12 or fewer years of education. There was a greater proportion of female participants vs. male participants among all race and ethnicity groups. The AI/AN sample was 64.5% female-identifying, a higher proportion than the NHW participants (53.3%) but fewer than Black participants (71.8%). On average, AI/AN participants were the youngest (mean age of 68 years) at baseline. The AI/AN group included similar proportions of participants enrolling with MCI (21%) and dementia (36%) to the NHW group. While over 90% of participants were currently or previously married (i.e., married, divorced, widowed, or separated), AI/AN (53%), Black (40%), and HW (54%) participants had lower current marriage rates when compared to NHW (71%) and Asian (70%) groups. Similar differences were observed for study partner relations. Thirty-nine percent of AI/AN, 30% of Black, and 39% of HW study partners were spouses, partners, or ex-spouses/ex-partners, compared to 61% for NHW and 55% for Asian co-participants. AI/AN participants had a higher observed rate of diabetes (10%) compared to all other groups, as well as double the frequency of obesity compared to NHW participants (38% vs. 19%). AI/AN participants also had the highest rate of individuals who smoke (11%) among all racial and ethnic groups in the NACC UDS.

Table 1: Descriptive Statistics for NIH Definition of Race/Ethnicity

Characteristics	AI/AN	Asian	Black	Hispanic White	non-Hispanic White	Other*
	N = 276 (0.6%)	N = 1,161 (2.6%)	N = 5,697 (12.7%)	N = 2,378 (5.3%)	N = 32,884 (73.5%)	N = 2,317 (5.2%)
Baseline Age	68.01 (10.63)	70.37 (10.29)	71.88 (8.99)	71.02 (10.11)	71.69 (10.68)	69.98 (10.53)
Binary Sex						
Female	178 (64.5%)	678 (58.4%)	4088 (71.8%)	1552 (65.3%)	17511 (53.3%)	1546 (66.7%)
Male	98 (35.5%)	483 (41.6%)	1609 (28.2%)	826 (34.7%)	15373 (46.7%)	771 (33.3%)
Patient Education						
< High school diploma	61 (22.1%)	88 (7.6%)	829 (14.6%)	792 (33.3%)	970 (2.9%)	649 (28%)
High school diploma/GED	105 (38%)	140 (12.1%)	1395 (24.5%)	474 (19.9%)	5540 (16.8%)	425 (18.3%)
Some College	54 (19.6%)	141 (12.1%)	1352 (23.7%)	389 (16.4%)	5673 (17.3%)	408 (17.6%)
4 Year Degree	28 (10.1%)	317 (27.3%)	851 (14.9%)	302 (12.7%)	8377 (25.5%)	360 (15.5%)
>4 Year Degree	25 (9.1%)	456 (39.3%)	1234 (21.7%)	402 (16.9%)	12082 (36.7%)	437 (18.9%)
Unknown/Missing	3 (1.1%)	19 (1.6%)	36 (0.6%)	19 (0.8%)	242 (0.7%)	38 (1.6%)
Marriage Status						
Married/Partnered	145 (52.5%)	812 (69.9%)	2259 (39.7%)	1279 (53.8%)	23224 (70.6%)	1162 (50.2%)
Previously Married	114 (41.3%)	267 (23%)	2876 (50.5%)	941 (39.6%)	8016 (24.4%)	941 (40.6%)
Never Married	15 (5.4%)	61 (5.3%)	488 (8.6%)	146 (6.1%)	1447 (4.4%)	172 (7.4%)
Other/Unknown	2 (0.7%)	21 (1.8%)	74 (1.3%)	12 (0.5%)	197 (0.6%)	42 (1.8%)
Residence Type						
Private Residence	264 (95.7%)	1063 (91.6%)	5215 (91.5%)	2196 (92.3%)	29403 (89.4%)	2143 (92.5%)
Independent Community	2 (0.7%)	43 (3.7%)	256 (4.5%)	73 (3.1%)	1755 (5.3%)	71 (3.1%)
Assisted Living	0 (0%)	26 (2.2%)	43 (0.8%)	24 (1%)	669 (2%)	28 (1.2%)
Nursing Home	2 (0.7%)	5 (0.4%)	38 (0.7%)	26 (1.1%)	440 (1.3%)	16 (0.7%)
Other/Unknown	8 (2.9%)	24 (2.1%)	145 (2.5%)	59 (2.5%)	617 (1.9%)	59 (2.5%)
Number of Visits	2.39 (1.75)	3.36 (2.97)	3.38 (2.91)	3.24 (2.7)	3.82 (3.14)	3.11 (2.72)
Categorized Number of Visits						
<3	178 (64.5%)	621 (53.5%)	3033 (53.2%)	1250 (52.6%)	15020 (45.7%)	1338 (57.7%)
≥3	98 (35.5%)	540 (46.5%)	2664 (46.8%)	1128 (47.4%)	17864 (54.3%)	979 (42.3%)
Retention to Follow-up 1 (within 18 months of initial visit)						
No	102 (44.5%)	349 (35.4%)	1945 (38.6%)	761 (36.5%)	8533 (29.4%)	742 (38.1%)
Yes	127 (55.5%)	636 (64.6%)	3098 (61.4%)	1322 (63.5%)	20471 (70.6%)	1204 (61.9%)
Retention to Follow-up 2 (within 18 months of F1)						
No	45 (38.5%)	160 (28.2%)	803 (29.1%)	311 (26.9%)	4166 (23.5%)	339 (32.3%)
Yes	72 (61.5%)	407 (71.8%)	1956 (70.9%)	846 (73.1%)	13570 (76.5%)	709 (67.7%)

Baseline Health Status	AI/AN	Asian	Black	Hispanic White	Non-Hispanic White	Other*
Baseline Diagnostic Status						
Normal Cognition	104 (37.7%)	487 (41.9%)	2533 (44.5%)	830 (34.9%)	12839 (39%)	812 (35%)
Impaired-not MCI	14 (5.1%)	51 (4.4%)	363 (6.4%)	133 (5.6%)	1260 (3.8%)	168 (7.3%)
MCI	57 (20.7%)	304 (26.2%)	1313 (23%)	608 (25.6%)	6977 (21.2%)	497 (21.5%)
Dementia	101 (36.6%)	319 (27.5%)	1488 (26.1%)	807 (33.9%)	11808 (35.9%)	840 (36.3%)
Baseline Primary Etiology						
Cognitively Normal	104 (37.7%)	487 (41.9%)	2533 (44.5%)	830 (34.9%)	12839 (39%)	812 (35%)
Alzheimer's	97 (35.1%)	368 (31.7%)	1812 (31.8%)	901 (37.9%)	11678 (35.5%)	866 (37.4%)
Lewy Body	11 (4%)	21 (1.8%)	75 (1.3%)	47 (2%)	1299 (4%)	55 (2.4%)
Frontotemporal	1 (0.4%)	50 (4.3%)	52 (0.9%)	50 (2.1%)	1991 (6.1%)	72 (3.1%)
Vascular	7 (2.5%)	40 (3.4%)	226 (4%)	57 (2.4%)	511 (1.6%)	67 (2.9%)
Other Reason	30 (10.9%)	118 (10.2%)	444 (7.8%)	266 (11.2%)	2069 (6.3%)	255 (11%)
Missing/Unknown	26 (9.4%)	77 (6.6%)	555 (9.7%)	227 (9.5%)	2497 (7.6%)	190 (8.2%)
Family History of ADRD						
No	104 (37.7%)	499 (43%)	2245 (39.4%)	881 (37%)	11900 (36.2%)	860 (37.1%)
Yes	111 (40.2%)	506 (43.6%)	2462 (43.2%)	1171 (49.2%)	17634 (53.6%)	1100 (47.5%)
Unknown/Missing	61 (22.1%)	156 (13.4%)	990 (17.4%)	326 (13.7%)	3350 (10.2%)	357 (15.4%)
Patient Independence						
Completely independent	173 (62.7%)	797 (68.6%)	4272 (75%)	1557 (65.5%)	21031 (64%)	1457 (62.9%)
Some assistance needed	71 (25.7%)	216 (18.6%)	835 (14.7%)	432 (18.2%)	7416 (22.6%)	477 (20.6%)
A lot of assistance needed	23 (8.3%)	97 (8.4%)	423 (7.4%)	220 (9.3%)	2987 (9.1%)	243 (10.5%)
Completely dependent	6 (2.2%)	34 (2.9%)	152 (2.7%)	155 (6.5%)	1250 (3.8%)	93 (4%)
Unknown/Missing	3 (1.1%)	17 (1.5%)	15 (0.3%)	14 (0.6%)	200 (0.6%)	47 (2%)
Diabetes Status						
No	65 (23.6%)	318 (27.4%)	1138 (20%)	543 (22.8%)	7713 (23.5%)	459 (19.8%)
Yes	27 (9.8%)	80 (6.9%)	436 (7.7%)	180 (7.6%)	749 (2.3%)	134 (5.8%)
Missing	184 (66.7%)	763 (65.7%)	4123 (72.4%)	1655 (69.6%)	24422 (74.3%)	1724 (74.4%)
Smoking Status						
Never Smoker	126 (45.7%)	888 (76.5%)	3109 (54.6%)	1510 (63.5%)	18155 (55.2%)	1335 (57.6%)
Previous Smoker	112 (40.6%)	233 (20.1%)	2032 (35.7%)	737 (31%)	12948 (39.4%)	799 (34.5%)
Current smoker	31 (11.2%)	27 (2.3%)	447 (7.8%)	94 (4%)	1267 (3.9%)	139 (6%)
Unknown/Missing	7 (2.5%)	13 (1.1%)	109 (1.9%)	37 (1.6%)	514 (1.6%)	44 (1.9%)
BMI Categories						
Underweight	1 (0.4%)	53 (4.6%)	58 (1%)	30 (1.3%)	449 (1.4%)	22 (0.9%)
Normal	63 (22.8%)	601 (51.8%)	1201 (21.1%)	589 (24.8%)	11291 (34.3%)	577 (24.9%)
Overweight	82 (29.7%)	298 (25.7%)	1763 (30.9%)	904 (38%)	11157 (33.9%)	759 (32.8%)
Obese	106 (38.4%)	54 (4.7%)	2063 (36.2%)	652 (27.4%)	6138 (18.7%)	678 (29.3%)



Unknown/Missing	24 (8.7%)	155 (13.4%)	612 (10.7%)	203 (8.5%)	3849 (11.7%)	281 (12.1%)
Number e4 Alleles						
0	101 (36.6%)	592 (51%)	2071 (36.4%)	1146 (48.2%)	15278 (46.5%)	901 (38.9%)
1	41 (14.9%)	190 (16.4%)	1465 (25.7%)	495 (20.8%)	8779 (26.7%)	496 (21.4%)
2	9 (3.3%)	33 (2.8%)	280 (4.9%)	61 (2.6%)	1773 (5.4%)	107 (4.6%)
Unknown/Missing	125 (45.3%)	346 (29.8%)	1881 (33%)	676 (28.4%)	7054 (21.5%)	813 (35.1%)
Study Partner Sex						
Female	185 (67%)	710 (61.2%)	3920 (68.8%)	1479 (62.2%)	20063 (61%)	1547 (66.8%)
Male	67 (24.3%)	399 (34.4%)	1492 (26.2%)	740 (31.1%)	11065 (33.6%)	658 (28.4%)
Unknown/Missing	24 (8.7%)	52 (4.5%)	285 (5%)	159 (6.7%)	1756 (5.3%)	112 (4.8%)
Study Partner Race						
AI/AN	136 (49.3%)	1 (0.1%)	9 (0.2%)	2 (0.1%)	26 (0.1%)	14 (0.6%)
Asian	1 (0.4%)	881 (75.9%)	10 (0.2%)	14 (0.6%)	260 (0.8%)	21 (0.9%)
Black	7 (2.5%)	8 (0.7%)	4882 (85.7%)	12 (0.5%)	144 (0.4%)	289 (12.5%)
Hispanic White	14 (5.1%)	6 (0.5%)	33 (0.6%)	1582 (66.5%)	361 (1.1%)	107 (4.6%)
Non-Hispanic White	63 (22.8%)	150 (12.9%)	185 (3.2%)	325 (13.7%)	29600 (90%)	573 (24.7%)
Unknown/Missing/Other	55 (19.9%)	115 (9.9%)	578 (10.1%)	443 (18.6%)	2493 (7.6%)	1313 (56.7%)
Study Partner Education						
< High school diploma	25 (9.1%)	32 (2.8%)	305 (5.4%)	306 (12.9%)	391 (1.2%)	269 (11.6%)
High school diploma/GED	110 (39.9%)	103 (8.9%)	1184 (20.8%)	448 (18.8%)	4502 (13.7%)	457 (19.7%)
Some College	50 (18.1%)	121 (10.4%)	1319 (23.2%)	451 (19%)	5481 (16.7%)	440 (19%)
4 Year Degree	30 (10.9%)	376 (32.4%)	1151 (20.2%)	437 (18.4%)	8654 (26.3%)	452 (19.5%)
>4 Year Degree	21 (7.6%)	422 (36.3%)	1168 (20.5%)	448 (18.8%)	10506 (31.9%)	444 (19.2%)
Unknown/Missing	40 (14.5%)	107 (9.2%)	570 (10%)	288 (12.1%)	3350 (10.2%)	255 (11%)

*The “Other” race category consists of 35 Native Hawaiian or Pacific Islander identifying participants, 762 participants with unknown race, 1392 multiracial participants, and 128 White participants with unknown Hispanic ethnicity.

Figure A.1: Descriptive statistics for the research participants in the NACC UDS.

A.5.1 Assessment of the Odds of Retention at the First Follow-up Visit:

Among 38,409 participants, 26,346 (68.6%) attended an expected follow-up within 18 months of their initial visit. We observed that AI/AN participants were retained to first follow-up at the lowest rate (122 out of 221 AI/AN (55.2%)) among all racial and ethnic groups (70.8% for NHW, 64.8% for Asian, 63.6% for HW, and 61.4% for Black participants). We used logistic regression to estimate the odds of retention at the first follow-up visit across racial and ethnic groups. Figure A.2 depicts the relative odds (and 95% Wald-based confidence intervals) of retention for the race and ethnic groups after adjusting for potential confounding factors. We estimated that AI/AN participants had 40.1% lower relative odds of being retained to their first follow-up visit compared to NHW participants, after adjustment for baseline diagnostic status, age, sex, education level, and smoking status (aOR: 0.599; 95% CI: 0.46-0.78; p ¡ 0.001). There was no evidence of differential relative odds of retention for participants enrolled after 2013 compared to participants enrolled before 2013 (i.e., no cohort effect).

A.5.2 Assessment of the Odds of Retention at All Follow-Up Visits:

Approximately one-fifth (19.7%) of the participants in the NACC data set were retained annually from their time of enrollment until either death or November 2021. In the four-fifths of participants who were not retained, we observed that the majority of those missed visits occurred soon after enrollment (within the first three follow-up appointments). Of the 243 AI/AN participants included in our analysis, 205 (84.4%) experienced a missed visit at some point during the course of follow-up. Of the 38 AI/AN participants who did not miss an expected follow-up appointment, 11 (29%) were censored because they either died within 18 months of their previous assessment or attended a previous assessment within 18 months

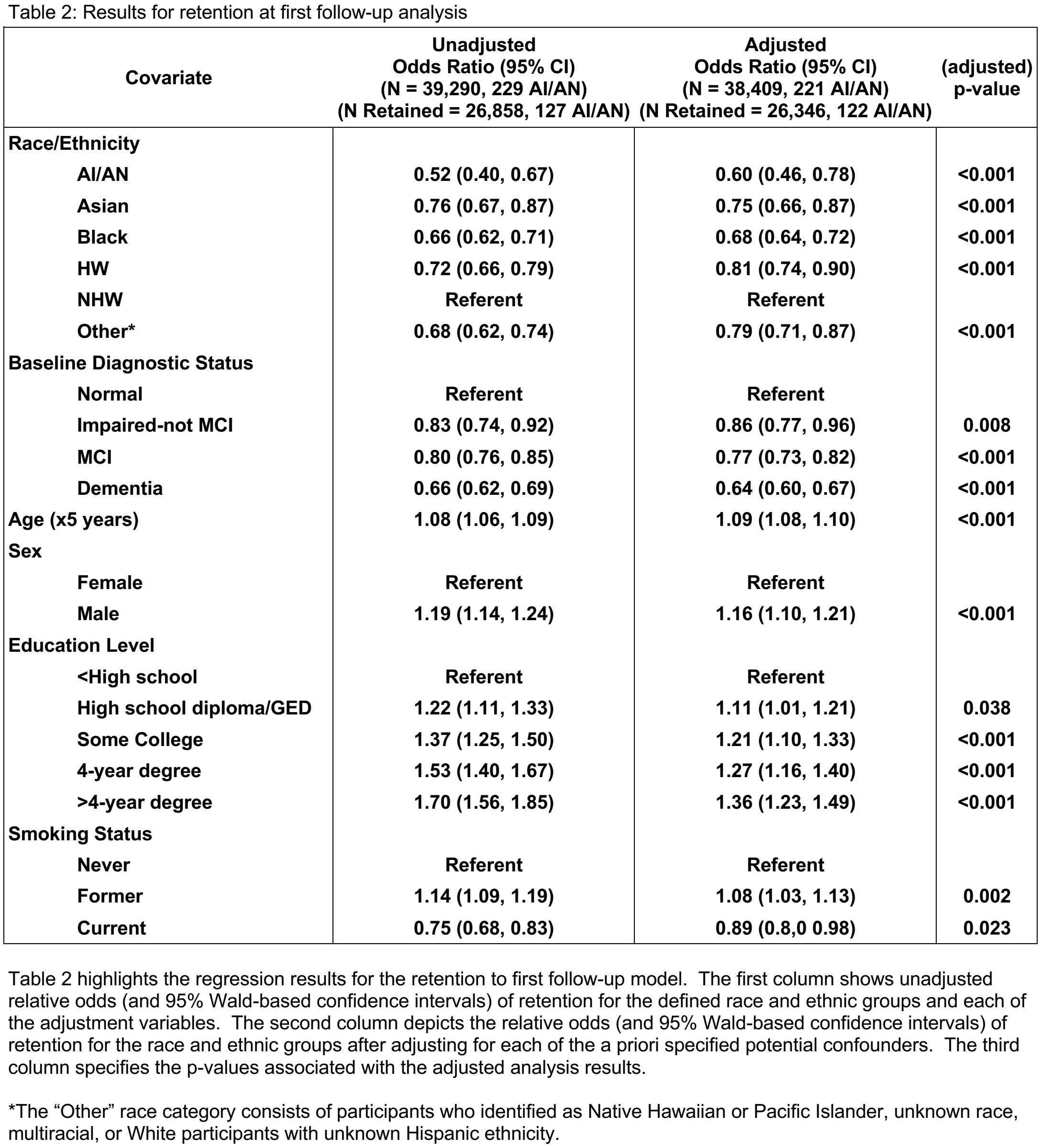


Table 2: Results for retention at first follow-up analysis

Covariate	Unadjusted Odds Ratio (95% CI) (N = 39,290, 229 AI/AN) (N Retained = 26,858, 127 AI/AN)	Adjusted Odds Ratio (95% CI) (N = 38,409, 221 AI/AN) (N Retained = 26,346, 122 AI/AN)	(adjusted) p-value
Race/Ethnicity			
AI/AN	0.52 (0.40, 0.67)	0.60 (0.46, 0.78)	<0.001
Asian	0.76 (0.67, 0.87)	0.75 (0.66, 0.87)	<0.001
Black	0.66 (0.62, 0.71)	0.68 (0.64, 0.72)	<0.001
HW	0.72 (0.66, 0.79)	0.81 (0.74, 0.90)	<0.001
NHW	Referent	Referent	
Other*	0.68 (0.62, 0.74)	0.79 (0.71, 0.87)	<0.001
Baseline Diagnostic Status			
Normal	Referent	Referent	
Impaired-not MCI	0.83 (0.74, 0.92)	0.86 (0.77, 0.96)	0.008
MCI	0.80 (0.76, 0.85)	0.77 (0.73, 0.82)	<0.001
Dementia	0.66 (0.62, 0.69)	0.64 (0.60, 0.67)	<0.001
Age (x5 years)	1.08 (1.06, 1.09)	1.09 (1.08, 1.10)	<0.001
Sex			
Female	Referent	Referent	
Male	1.19 (1.14, 1.24)	1.16 (1.10, 1.21)	<0.001
Education Level			
<High school	Referent	Referent	
High school diploma/GED	1.22 (1.11, 1.33)	1.11 (1.01, 1.21)	0.038
Some College	1.37 (1.25, 1.50)	1.21 (1.10, 1.33)	<0.001
4-year degree	1.53 (1.40, 1.67)	1.27 (1.16, 1.40)	<0.001
>4-year degree	1.70 (1.56, 1.85)	1.36 (1.23, 1.49)	<0.001
Smoking Status			
Never	Referent	Referent	
Former	1.14 (1.09, 1.19)	1.08 (1.03, 1.13)	0.002
Current	0.75 (0.68, 0.83)	0.89 (0.8,0 0.98)	0.023

Table 2 highlights the regression results for the retention to first follow-up model. The first column shows unadjusted relative odds (and 95% Wald-based confidence intervals) of retention for the defined race and ethnic groups and each of the adjustment variables. The second column depicts the relative odds (and 95% Wald-based confidence intervals) of retention for the race and ethnic groups after adjusting for each of the a priori specified potential confounders. The third column specifies the p-values associated with the adjusted analysis results.

*The "Other" race category consists of participants who identified as Native Hawaiian or Pacific Islander, unknown race, multiracial, or White participants with unknown Hispanic ethnicity.

Figure A.2: Results for retention to first follow-up.

of the end of study follow-up. Of the 205 AI/AN participants who were not retained at an expected follow-up appointment, 103 participants were not retained at the first follow-up appointment, 47 were not retained at the second follow-up after having completed the first, and 27 were not retained at the third follow-up after having completed the first two visits. This accounts for 86.3% of the 205 AI/AN participants who missed a visit. In contrast, 71.5% of NHW participants missed the first, second, or third follow-up appointments. Based on the results of a continuation ratio model considering the relative odds of attending an annual visit within 18 months of the preceding visit, we estimated that AI/AN participants had a 47% lower odds of being retained to their next visit, conditional on attending all previous visits, compared to NHW participants (aOR: 0.53; 95% CI: 0.44-0.64). This comparison was adjusted for baseline diagnostic status, age, sex, education level, and smoking status. See Figure A.3 for results.

A.5.3 Sensitivity Analyses:

To assess the impact of the COVID-19 pandemic on retention, we re-ran all analyses after adjusting the end-of-follow-up date to be February 2020. The overall retention rate for pre-COVID data was numerically higher (70.2%) than the main analysis proportion (68.6%). Of the 18 AI/AN participants who enrolled in the study within 18 months of February 2020, only seven returned for their expected follow-up visit. We observed a retention rate of 56.2% for AI/AN participants who enrolled at least 18 months before COVID hit the US, which is slightly higher than the previously observed 55.2% for the analysis that includes participants who enrolled within 18 months of the start of COVID. Results of the model for the outcome of retention to first follow-up remained nearly the same, with AI/AN participants having 40.1% lower odds of returning for the first follow-up appointment compared to NHW participants.

A second sensitivity analysis revealed that COVID impacted retention across all participants

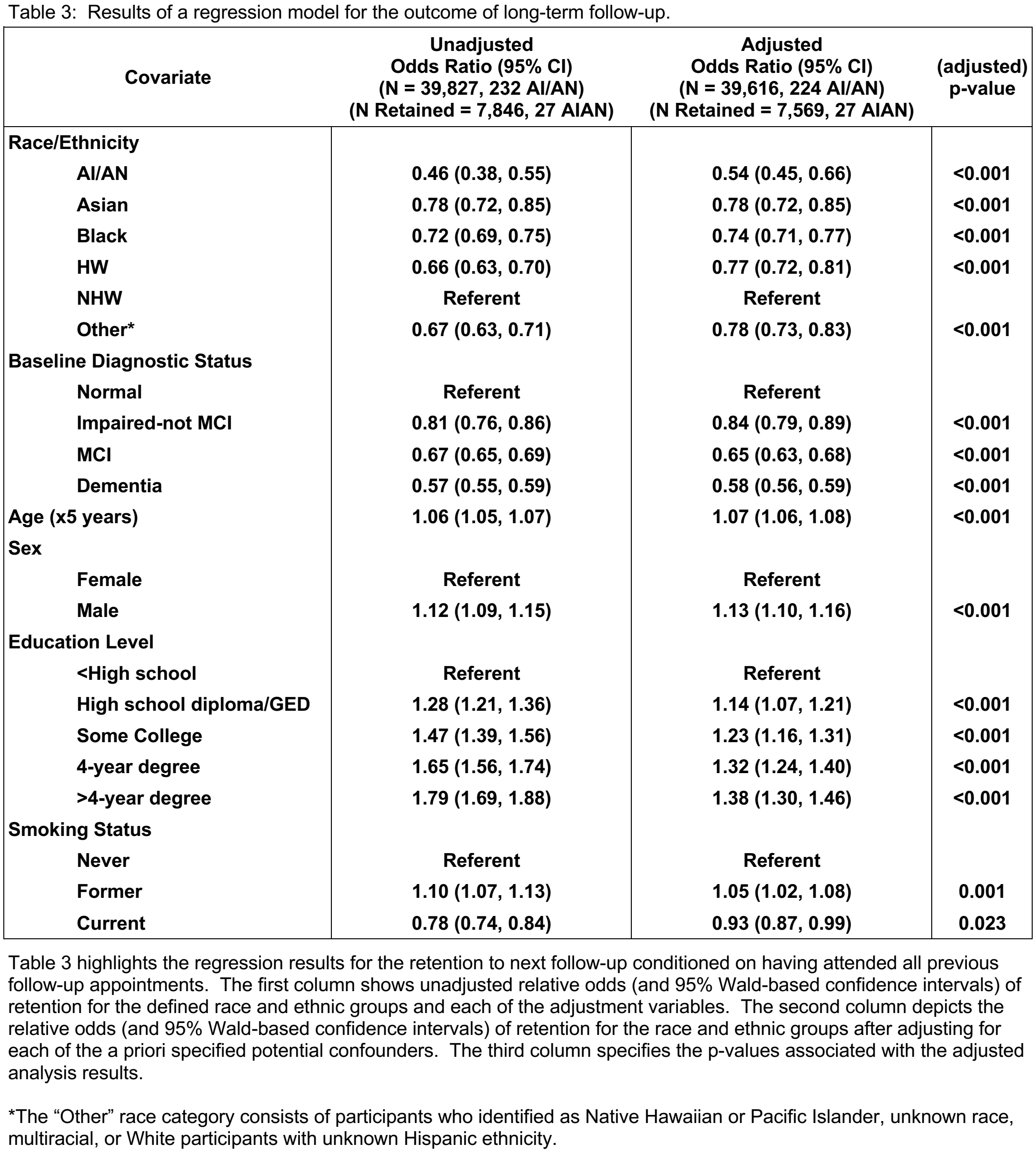


Table 3: Results of a regression model for the outcome of long-term follow-up.

Covariate	Unadjusted Odds Ratio (95% CI) (N = 39,827, 232 AI/AN) (N Retained = 7,846, 27 AIAN)	Adjusted Odds Ratio (95% CI) (N = 39,616, 224 AI/AN) (N Retained = 7,569, 27 AIAN)	(adjusted) p-value
Race/Ethnicity			
AI/AN	0.46 (0.38, 0.55)	0.54 (0.45, 0.66)	<0.001
Asian	0.78 (0.72, 0.85)	0.78 (0.72, 0.85)	<0.001
Black	0.72 (0.69, 0.75)	0.74 (0.71, 0.77)	<0.001
HW	0.66 (0.63, 0.70)	0.77 (0.72, 0.81)	<0.001
NHW	Referent	Referent	
Other*	0.67 (0.63, 0.71)	0.78 (0.73, 0.83)	<0.001
Baseline Diagnostic Status			
Normal	Referent	Referent	
Impaired-not MCI	0.81 (0.76, 0.86)	0.84 (0.79, 0.89)	<0.001
MCI	0.67 (0.65, 0.69)	0.65 (0.63, 0.68)	<0.001
Dementia	0.57 (0.55, 0.59)	0.58 (0.56, 0.59)	<0.001
Age (x5 years)	1.06 (1.05, 1.07)	1.07 (1.06, 1.08)	<0.001
Sex			
Female	Referent	Referent	
Male	1.12 (1.09, 1.15)	1.13 (1.10, 1.16)	<0.001
Education Level			
<High school	Referent	Referent	
High school diploma/GED	1.28 (1.21, 1.36)	1.14 (1.07, 1.21)	<0.001
Some College	1.47 (1.39, 1.56)	1.23 (1.16, 1.31)	<0.001
4-year degree	1.65 (1.56, 1.74)	1.32 (1.24, 1.40)	<0.001
>4-year degree	1.79 (1.69, 1.88)	1.38 (1.30, 1.46)	<0.001
Smoking Status			
Never	Referent	Referent	
Former	1.10 (1.07, 1.13)	1.05 (1.02, 1.08)	0.001
Current	0.78 (0.74, 0.84)	0.93 (0.87, 0.99)	0.023

Table 3 highlights the regression results for the retention to next follow-up conditioned on having attended all previous follow-up appointments. The first column shows unadjusted relative odds (and 95% Wald-based confidence intervals) of retention for the defined race and ethnic groups and each of the adjustment variables. The second column depicts the relative odds (and 95% Wald-based confidence intervals) of retention for the race and ethnic groups after adjusting for each of the a priori specified potential confounders. The third column specifies the p-values associated with the adjusted analysis results.

*The "Other" race category consists of participants who identified as Native Hawaiian or Pacific Islander, unknown race, multiracial, or White participants with unknown Hispanic ethnicity.

Figure A.3: Results of regression of long-term follow-up.

but did not exacerbate the differences between groups. In total, 4,808 participants missed an expected follow-up after COVID (24 of whom were AI/AN participants), for an overall retention rate of 32.0% (22.1% for AI/AN participants). In this analysis, 92, 42, and 22 AI/AN participants were not retained to the first, second, and third follow-up visits, accounting for 90.2% of AI/AN participants who were not retained. Compared to NHW participants, AI/AN participants were retained for the long-term at a lower rate (33.9% NHW vs 22.1% AI/AN). Also similar to the full analysis, AI/AN participants had an estimated 47% lower odds of being retained at an expected follow-up visit than NHW participants (aOR: 0.534; 95% CI: (0.44, 0.65)). See Supplementary Materials section 1.2 for details.

To assess if potential effect-modification by site could account for lower retention among AI/AN participants, we restricted our analyses to two subsets of ADRCs. Figure A.4 shows the odds ratios for retention to the first follow-up appointment across racial and ethnic groups compared to NHW participants from the analyses with all sites, the 24 sites with at least one AI/AN participant, and the five sites with at least ten AI/AN participants. While retention rates were not the same across sites, the overall results of the differences in retention of AI/AN participants relative to NHW remained consistent. Point estimates of retention rates compared to NHW were lower for AI/AN participants (as well as Black and HW participants) in the group of five sites with at least ten AI/AN participants. See Supplementary Materials section 1.3 for details.

A.6 Discussion

We examined the rates of retention of AI/AN participants in the NACC UDS. We considered two definitions of participant retention: 1) the odds of retention to the first follow-up visit and 2) the relative odds of completing any follow-up visit having attended all previous follow-up visits. In both cases, AI/AN participants were retained at significantly lower rates when

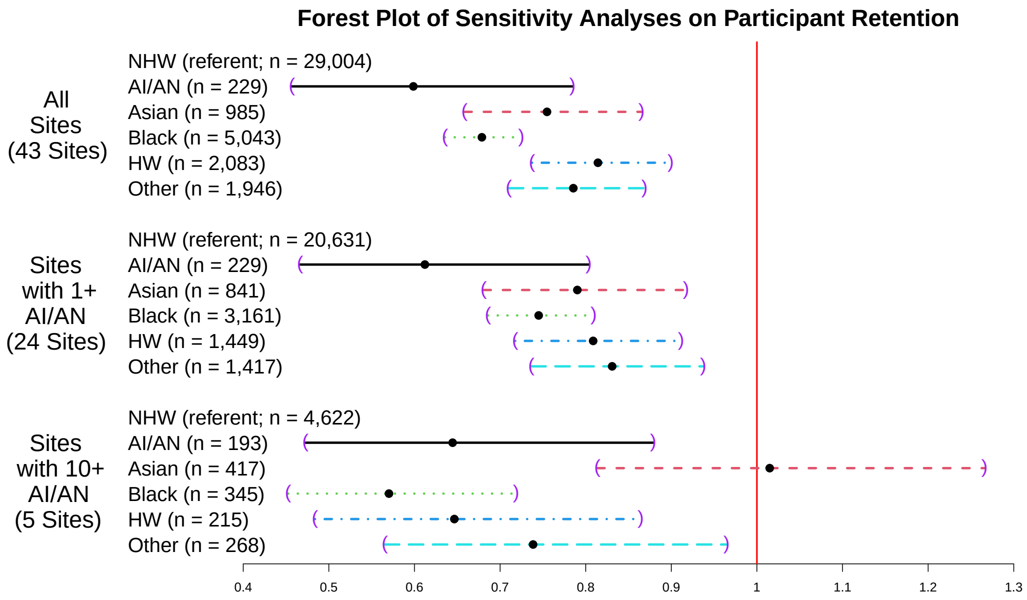


Figure 1: Forest plot of the odds ratios for site−specific sensitivity analyses on retention to first follow−up visit for the race and ethnicity groups compared to non−Hispanic White participants.

Figure A.4: Forest plot of the odds ratios for site-specific sensitivity analyses on retention to first follow-up visit for the race and ethnicity groups compared to non-hispanic White participants.

compared to NHW participants. Both sets of analyses had similar results regardless of the study end date, the exclusion of sites with limited numbers of AI/AN participants, or the adjustment for potential confounders.

In sensitivity analyses, we considered that COVID may have changed retention patterns. We observed that COVID impacted retention rates, but not enough to change the results of these analyses. It was noted early in the pandemic that COVID disproportionately affected AI/AN populations [95]. However, our results do not suggest a differential effect of COVID on retention between AI/AN and NHW participants. In light of how COVID impacted minoritized communities, these results may be surprising [96, 97]. However, few AI/AN participants enrolled within the year and a half of COVID reaching the US, potentially limiting the impact of the pandemic on our primary analyses.

Participant retention in longitudinal studies is critical to maintaining study power, reducing bias, and ensuring generalizability [64]. In longitudinal analyses, statistical methods require a minimum of three visits per individual to estimate patient trends over time. Thus, participants with fewer than three visits are commonly excluded from analyses assessing longitudinal trends. Further, if participants who are lost to follow-up differ from participants who remain in study, then selection bias is likely to result. Specifically, if participant attrition is due to a reason related to an uncollected characteristic, or for a reason related to the outcome of interest, then results from a statistical analysis may contain significant bias, even in settings with little loss to follow-up [98]. Even in the absence of bias, however, the loss in power may increase the probability of type two error. In addition, prospective observational settings (e.g., NACC) may not generalize to the broader population from the start because the sample of participants may not be representative of the overall population. Thus, differential attrition in addition to potential analytical biases may make it difficult to draw meaningful conclusions that generalize to any population [64].

The results of our work imply that efforts to retain AI/AN study participants should be prioritized. Retention of minority participants may be improved through tactics such as including community members in study design and data collection; providing detailed information on the research goals and how participant data will be used, as well as sharing benefits of individual participation; and limiting barriers to entry and continued participation [99, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111]. A recent secondary analysis by Salazar et al. found that retention strategies focused on “study personnel” and “study description” were associated with greater participant retention in NACC UDS data [112]. They did not observe effect modification by race and ethnicity. The “study personnel” strategy included tactics such as diverse staff, regular retention trainings for staff, and continuity between specific staff members and specific participants over time [112].

Within the NACC UDS, we previously observed a positive association between the number of retention tactics employed by a site and the rate of participant retention [92]. Under the possibility of site-specific retention practices being different between sites with AI/AN participants and those without AI/AN participants, we repeated our primary analyses on two subsets of the sites: 1) all sites with at least one AI/AN participant, and 2) all sites with at least ten AI/AN participants. In both settings, our analyses resulted in nearly identical odds ratios for retention between AI/AN and NHW participants. The results of Salazar et al. also suggested that any differences in retention strategies between the sites did not differentially impact the relative odds of attrition between AI/AN participants and NHW participants [112]. However, we observed lower retention rates for Black and HW participants in sites with at least ten AI/AN participants, compared to the analysis with all sites. This suggests that there may be differences in retention among some racial and ethnic groups, but those differences are not observed between AI/AN and NHW participants. Our data do not inform whether these differences could be related to differences in site-specific retention practices or if similar tactics may have differentially affected race and ethnicity groups.

When conducting research with participants from underrepresented communities, it is important to take precautions to ensure all research methods, results, and interpretations are appropriate. For example, it is necessary to be aware of history of research in those communities [113, 114]. Work by the Native American Center for Excellence (NACE) describes how the trust of AI/AN communities has been betrayed by researchers whose work was unethical, drew inappropriate conclusions, and was culturally inconsiderate [3]. To gain trust in Native communities, NACE recommended that researchers work with communities as partners [3]. For example, previous work has suggested developing a sense of community/research partnerships by regularly providing research feedback to participants (and the community), offering small personal items with the study logo, and hiring community members as part of the team [102, 106, 108, 110, 115, 116, 117]. The latter can help ensure culturally appropriate methods, reduce cultural misunderstandings, and provide participants with a trusted point of contact [3, 118]. Redwood et al. also described persistence and detailed tracking of contact attempts as vital to their retention of AI/AN participants [115].

It is worth noting that most of the studies that have considered retention in AI/AN populations are over a decade old. The results of the present work suggest that retention gaps have persisted. It will likely be essential for researchers to actively invest time into building personal relationships in the AI/AN community to improve these outcomes. This will necessitate building enduring relationships, beyond conventional funding cycles, incorporating research strategies with long-term relationships in mind, and developing educational programs to familiarize communities with goals and procedures. Education and relationship-building may improve community attitudes towards research over time, which directly affect willingness to participate and even study retention [119]. Investigators, not just members of the research team but lead PIs, can and should work directly with communities to introduce research topics, explain results, and generally educate on research practice and findings. Specific to ADRD, Jernigan et al. noted the need for education on risk factors and caretaking skills within AI/AN communities [120]. Regular presentations on ADRD topics and hosting small gatherings in which investigators invest time and energy to learn the names of local leaders and cultural traditions of the community all can help to build a relationship beyond the traditional researcher-participant interaction.

There are several limitations to our study. We aimed to assess retention of AI/AN participants in ADRCs across the United States. These results may not generalize to the overall US population, to the general AI/AN population, or to other studies. For example, we observed a smoking rate of 11% among AI/AN participants, which is lower than the 27% estimated by the CDC report on tobacco use among AI/AN adults [121]. Further, we were unable to account for differences in retention tactics employed at centers more likely to have participants of a certain race or ethnicity. We were also limited to the definitions of some covariates. For example, we could not explore the intersection between sex and gender with race and ethnicity due to the wording of the data collection instruments. Nevertheless, our work highlights an important area of deficit for ADRCs with respect to the retention of AI/AN participants. Future work can examine the time-to-return for follow-up, as well as attempt to understand how different retention tactics work with the AI/AN participants specifically, in NACC and other studies.

In conclusion, to reduce bias and improve validity of results, it is essential to retain participants in studies focused on longitudinal outcomes. For generalizability, identification of subgroup-specific risk factors, and ensuring health equity, it is especially important to retain participants from underrepresented populations. Participants who identify as AI/AN are vastly understudied and underrepresented, despite a disproportionate burden of disease. Our analyses of retention show that AI/AN participants were not retained at similar rates as NHW participants. To effectively learn more about ADRD in AI/AN communities, concerted efforts will be needed to increase retention of these participants.

A.7 Supplementary Materials

A.7.1 Sensitivity Analyses

Results from Full Data Analysis

To ensure our definition of retention did not influence results, we conducted a sensitivity analysis where we did not censor participants that were seen more than 18 months after their prior scheduled visit. In this analysis all participants were followed until their last visit. Participants who did not have a subsequent visit were censored if their last observed visit was 18 months or longer from the data freeze date (November, 2021). The results of this sensitivity analysis are qualitatively the same as the two analyses presented in the primary manuscript. We estimated that AI/AN participants had a 56% lower odds of being retained to their next visit, conditional on attending all previous visits, compared to NHW participants (aOR: 0.44; 95% CI: 0.36-0.54). Supplementary Table 1 provides full results.

Supplementary Table 1: Long-term overall retention results using modified censoring definition for retention.

Covariate	Unadjusted Odds Ratio (95% CI) (N =41,403, 236 AI/AN) (N Retained = 13,972, 50 AI/AN)	Adjusted Odds Ratio (95% CI) (N = 40,412, 228 AI/AN) (N Retained = 13,670, 50 AI/AN)	(adjusted) p-value
Race/Ethnicity			
AI/AN	0.35 (0.29, 0.42)	0.44 (0.36, 0.54)	<0.001
Asian	0.77 (0.70, 0.84)	0.77 (0.70, 0.84)	<0.001
Black	0.69 (0.67, 0.72)	0.72 (0.69, 0.75)	<0.001
HW	0.64 (0.61, 0.68)	0.78 (0.73, 0.83)	<0.001
NHW	Referent	Referent	
Other*	0.63 (0.59, 0.67)	0.77 (0.72, 0.82)	<0.001
Baseline Diagnostic Status			
Normal	Referent	Referent	
Impaired-not MCI	0.7 (0.66, 0.76)	0.74 (0.69, 0.8)	<0.001
MCI	0.55 (0.53, 0.57)	0.55 (0.53, 0.57)	<0.001
Dementia	0.46 (0.44, 0.47)	0.47 (0.46, 0.49)	<0.001
Age (x5 years)	1.04 (1.03, 1.05)	1.06 (1.05, 1.06)	<0.001
Sex			
Female	Referent	Referent	
Male	1.08 (1.05, 1.11)	1.11 (1.07, 1.14)	<0.001
Education (x4 years)	1.24 (1.22, 1.26)	1.13 (1.11, 1.15)	<0.001
<High school	Referent	Referent	
High school diploma/GED	1.3 (1.22, 1.38)	1.12 (1.05, 1.19)	<0.001
Some College	1.62 (1.53, 1.72)	1.29 (1.21, 1.37)	<0.001
4-year degree	1.83 (1.73, 1.94)	1.39 (1.3, 1.48)	<0.001
>4-year degree	2.12 (2, 2.24)	1.54 (1.44, 1.64)	<0.001
Smoking Status			
Never	Referent	Referent	
Former	1.08 (1.04, 1.11)	1.03 (1, 1.06)	0.067
Current	0.78 (0.73, 0.83)	0.93 (0.87, 1)	0.061

Supplementary Table 1 highlights the regression results for the retention to next follow-up. The first column shows unadjusted relative odds (and 95% Wald-based confidence intervals) of retention for the defined race and ethnic groups and each of the adjustment variables. The second column depicts the relative odds (and 95% Wald-based confidence intervals) of retention for the race and ethnic groups after adjusting for each of the a priori specified potential confounders. The third column specifies the p-values associated with the adjusted analysis results.

*The “Other” race category consists of participants who identified as Native Hawaiian or Pacific Islander, unknown race, multiracial, or White participants with unknown Hispanic ethnicity.

Supplementary Table 2: Results for retention at first follow-up analysis (Pre-COVID timeframe).

Covariate	Unadjusted Odds Ratio (95% CI) (N = 35,670, 211 AI/AN)	Adjusted Odds Ratio (95% CI) (N = 34,840, 203 AI/AN)	(adjusted) p-value
Race/Ethnicity			
AI/AN	0.50 (0.38, 0.66)	0.60 (0.45, 0.80)	<0.001
Asian	0.76 (0.66, 0.88)	0.74 (0.64, 0.86)	<0.001
Black	0.69 (0.65, 0.74)	0.72 (0.67, 0.77)	<0.001
HW	0.73 (0.66, 0.80)	0.84 (0.76, 0.93)	0.001
NHW	Referent	Referent	
Other*	0.67 (0.61, 0.74)	0.79 (0.71, 0.88)	<0.001
Baseline Diagnostic Status			
Normal	Referent	Referent	
Impaired-not MCI	0.85 (0.76, 0.96)	0.89 (0.79, 1.00)	0.052
MCI	0.81 (0.76, 0.86)	0.78 (0.73, 0.83)	<0.001
Dementia	0.63 (0.60, 0.67)	0.62 (0.59, 0.66)	<0.001
Age (x5 years)	1.07 (1.06, 1.08)	1.08 (1.07, 1.09)	<0.001
Sex			
Female	Referent	Referent	
Male	1.18 (1.13, 1.24)	1.15 (1.10, 1.21)	<0.001
Education (x4 years)	1.22 (1.19, 1.25)	1.13 (1.10, 1.17)	<0.001
<High school	Referent	Referent	
High school diploma/GED	1.23 (1.12, 1.35)	1.12 (1.01, 1.24)	0.026
Some College	1.40 (1.28, 1.54)	1.24 (1.12, 1.37)	<0.001
4-year degree	1.60 (1.46, 1.75)	1.34 (1.21, 1.48)	<0.001
>4-year degree	1.76 (1.61, 1.92)	1.41 (1.28, 1.55)	<0.001
Smoking Status			
Never	Referent	Referent	
Former	1.13 (1.07, 1.18)	1.07 (1.01, 1.12)	0.011
Current	0.73 (0.66, 0.82)	0.87 (0.78, 0.97)	0.011

Supplementary Table 2 highlights the regression results for the retention to first follow-up model. This analysis only considers participants who enrolled before August, 2018 (18 months prior to COVID reaching the US). The first column shows unadjusted relative odds (and 95% Wald-based confidence intervals) of retention for the defined race and ethnic groups and each of the adjustment variables. The second column depicts the relative odds (and 95% Wald-based confidence intervals) of retention for the race and ethnic groups after adjusting for each of the a priori specified potential confounders. The third column specifies the p-values associated with the adjusted analysis results.

*The “Other” race category consists of participants who identified as Native Hawaiian or Pacific Islander, unknown race, multiracial, or White participants with unknown Hispanic ethnicity.

Results from Sensitivity Analyses Assessing Impact of COVID

To assess if COVID impacted retention results, we ran the same analyses with the data freeze date set as February, 2020 instead of the actual freeze date of November 2021. All other details of the two studies were the same. For retention to first follow-up, we estimated that AI/AN participants had 40% lower odds of being retained compared to NHW participants (aOR: 0.60; 95% CI: 0.45-0.80). These results are similar to the full analysis (aOR: 0.599; 95% CI: 0.46-0.78). More details are in Supplementary Table 2.

Based on the results of a continuation ratio model considering the relative odds of attending an annual visit within 18 months of the preceding visit, we estimated that AI/AN participants had a 48% lower odds of being retained to their next visit, conditional on attending all previous visits, compared to NHW participants (aOR: 0.52; 95% CI: 0.43-0.63). These results are similar to the full analysis (aOR: 0.54; 95% CI: 0.45-0.66). More details are in Supplementary Table 3.

1.3 Results from Site-Specific Sensitivity Analyses

We considered potential site-specific effects by restricting analyses to sites with AI/AN participants. To assess if ADRCs with AI/AN participants have different retention rates than other ADRCs, we considered two analyses: 1) ADRCs with at least one AI/AN participant, and 2) ADRCs with at least ten AI/AN participants. We found that AI/AN participants had an estimated 38.2% (aOR: 0.612; 95% CI: 0.47-0.80) lower odds of being retained to first follow-up than NHW participants at sites with at least one AI/AN participants. Similarly, we found that AI/AN participants had a 35.5% (aOR: 0.645; 95% CI: 0.47-0.88) lower odds of being retained to first follow-up than NHW participants at sites with at least ten AI/AN participants. These results are consistent with the results from the analysis including all ADRCs (aOR: 0.599; 95% CI: 0.46-0.78). Full results from the models are presented in

Supplementary Table 3: Results for number of visits until the first missed annual follow-up appointment analysis (pre-COVID timeframe).

Covariate	Unadjusted Odds Ratio (95% CI) (N = 36,142, 214 AI/AN) (N Retained = 9,221, 34 AI/AN)	Adjusted Odds Ratio (95% CI) (N = 34,995, 206 AI/AN) (N Retained = 9,019, 33 AI/AN)	(adjusted) p-value
Race/Ethnicity			
AI/AN	0.42 (0.35, 0.51)	0.52 (0.43, 0.63)	<0.001
Asian	0.81 (0.74, 0.89)	0.79 (0.72, 0.87)	<0.001
Black	0.71 (0.68, 0.74)	0.73 (0.7, 0.77)	<0.001
HW	0.65 (0.61, 0.69)	0.78 (0.73, 0.83)	<0.001
NHW	Referent	Referent	
Other*	0.65 (0.61, 0.69)	0.78 (0.73, 0.83)	<0.001
Baseline Cognitive Status			
Normal	Referent	Referent	
Impaired-not MCI	0.82 (0.77, 0.88)	0.86 (0.81, 0.92)	<0.001
MCI	0.65 (0.63, 0.68)	0.64 (0.62, 0.67)	<0.001
Dementia	0.52 (0.51, 0.54)	0.54 (0.52, 0.55)	<0.001
Age (x5 years)	1.06 (1.05, 1.06)	1.07 (1.06, 1.07)	<0.001
Sex			
Female	Referent	Referent	
Male	1.12 (1.09, 1.15)	1.13 (1.1, 1.17)	<0.001
Education (x4 years)	1.25 (1.21, 1.23)	1.14 (1.12, 1.16)	<0.001
<High school	Referent	Referent	
High school diploma/GED	1.3 (1.22, 1.38)	1.15 (1.08, 1.22)	<0.001
Some College	1.53 (1.44, 1.63)	1.26 (1.19, 1.35)	<0.001
4 year degree	1.75 (1.65, 1.85)	1.38 (1.29, 1.47)	<0.001
>4-year degree	1.94 (1.84, 2.05)	1.47 (1.38, 1.56)	<0.001
Smoking Status			
Never	Referent	Referent	
Former	1.08 (1.05, 1.11)	1.03 (1, 1.06)	0.051
Current	0.74 (0.7, 0.8)	0.89 (0.83, 0.95)	0.001

Supplementary Table 3 highlights the regression results for the retention to next follow-up conditioned on having attended all previous follow-up appointments. This analysis only considers participants who enrolled before August, 2018 (18 months prior to COVID reaching the US). The first column shows unadjusted relative odds (and 95% Wald-based confidence intervals) of retention for the defined race and ethnic groups and each of the adjustment variables. The second column depicts the relative odds (and 95% Wald-based confidence intervals) of retention for the race and ethnic groups after adjusting for each of the a priori specified potential confounders. The third column specifies the p-values associated with the adjusted analysis results.

*The “Other” race category consists of participants who identified as Native Hawaiian or Pacific Islander, unknown race, multiracial, or White participants with unknown Hispanic ethnicity.

Supplementary Table 4: Results of sensitivity analysis on retention at first follow-up for sites with at least one AI/AN participant.

Covariate	Unadjusted Odds Ratio (95% CI) (N = 27,728, 229 AI/AN) (N Retained = 18,982, 127 AI/AN)	Adjusted Odds Ratio (95% CI) (N = 27,106, 221 AI/AN) (N Retained = 18,624, 122 AI/AN)	(adjusted) p-value
Race/Ethnicity			
AI/AN	0.52 (0.40, 0.68)	0.61 (0.47, 0.80)	<0.001
Asian	0.78 (0.68, 0.90)	0.79 (0.68, 0.92)	0.002
Black	0.71 (0.66, 0.77)	0.75 (0.69, 0.81)	<0.001
HW	0.72 (0.64, 0.80)	0.81 (0.72, 0.91)	0.001
NHW	Referent	Referent	
Other*	0.69 (0.61, 0.77)	0.83 (0.74, 0.94)	0.003
Baseline Diagnostic Status			
Normal	Referent	Referent	
Impaired-not MCI	0.84 (0.74, 0.95)	0.85 (0.75, 0.97)	0.014
MCI	0.84 (0.78, 0.89)	0.78 (0.73, 0.84)	<0.001
Dementia	0.68 (0.64, 0.72)	0.65 (0.61, 0.69)	<0.001
Age (x5 years)	1.10 (1.09, 1.11)	1.11 (1.10, 1.13)	<0.001
Sex			
Female	Referent	Referent	
Male	1.21 (1.15, 1.27)	1.19 (1.13, 1.26)	<0.001
Education (x4 years)	1.18 (1.15, 1.22)	1.11 (1.07, 1.15)	<0.001
<High school	Referent	Referent	
High school diploma/GED	1.18 (1.07, 1.31)	1.10 (0.98, 1.23)	0.115
Some College	1.39 (1.25, 1.54)	1.26 (1.13, 1.42)	<0.001
4-year degree	1.47 (1.33, 1.63)	1.27 (1.14, 1.43)	<0.001
>4-year degree	1.62 (1.47, 1.79)	1.33 (1.19, 1.49)	<0.001
Smoking Status			
Never	Referent	Referent	
Former	1.17 (1.11, 1.24)	1.10 (1.04, 1.16)	0.001
Current	0.72 (0.63, 0.82)	0.87 (0.76, 0.99)	0.032

Supplementary Table 4 highlights the regression results for the retention to first follow-up model restricted to the 24 ADRCs with at least one AI/AN participant. The first column shows unadjusted relative odds (and 95% Wald-based confidence intervals) of retention for the defined race and ethnic groups and each of the adjustment variables. The second column depicts the relative odds (and 95% Wald-based confidence intervals) of retention for the race and ethnic groups after adjusting for each of the a priori specified potential confounders. The third column specifies the p-values associated with the adjusted analysis results.

*The "Other" race category consists of participants who identified as Native Hawaiian or Pacific Islander, unknown race, multiracial, or White participants with unknown Hispanic ethnicity.

Supplementary Table 5: Results of sensitivity analysis on retention at first follow-up for sites with at least ten AI/AN participants.

	Unadjusted	Adjusted	
Covariate	Odds Ratio (95% CI)	Odds Ratio (95% CI)	(adjusted)
	(N = 6,060, 193 AI/AN)	(N = 5,863 , 186 AI/AN)	p-value
	(N Retained = 3,852, 106 AI/AN)	(N Retained = 3,746, 101 AI/AN)	
Race/Ethnicity			
AI/AN	0.64 (0.48, 0.85)	0.64 (0.47, 0.88)	0.006
Asian	0.99 (0.81, 1.23)	1.01 (0.81, 1.27)	0.895
Black	0.54 (0.43, 0.67)	0.57 (0.45, 0.72)	<0.001
HW	0.68 (0.51, 0.89)	0.65 (0.48, 0.86)	0.003
NHW	Referent	Referent	
Other*	0.62 (0.48, 0.79)	0.74 (0.57, 0.96)	0.026
Baseline Diagnostic Status			
Normal	Referent	Referent	
Impaired-not MCI	1 (0.74, 1.35)	1.07 (0.78, 1.46)	0.674
MCI	1.17 (1.02, 1.34)	1.03 (0.89, 1.19)	0.711
Dementia	0.88 (0.78, 1)	0.82 (0.71, 0.93)	0.003
Age (x5 years)	1.17 (1.14, 1.2)	1.18 (1.14, 1.21)	<0.001
Sex			
Female	Referent	Referent	
Male	1.05 (0.95, 1.17)	1 (0.89, 1.12)	0.965
Education (x4 years)	1.07 (1, 1.15)	1.02 (0.95, 1.1)	0.591
<High school	Referent	Referent	
High school diploma/GED	1.13 (0.86, 1.49)	1.07 (0.79, 1.43)	0.668
Some College	1.23 (0.94, 1.62)	1.18 (0.88, 1.59)	0.269
4-year degree	1.33 (1.02, 1.74)	1.22 (0.92, 1.64)	0.172
>4-year degree	1.29 (1, 1.67)	1.12 (0.84, 1.49)	0.450
Smoking Status			
Never	Referent	Referent	
Former	1.24 (1.11, 1.39)	1.14 (1.01, 1.28)	0.029
Current	0.64 (0.49, 0.83)	0.83 (0.63, 1.09)	0.177

Supplementary Table 5 highlights the regression results for the retention to first follow-up model restricted to the five ADRCs with at least ten AI/AN participants. The first column shows unadjusted relative odds (and 95% Wald-based confidence intervals) of retention for the defined race and ethnic groups and each of the adjustment variables. The second column depicts the relative odds (and 95% Wald-based confidence intervals) of retention for the race and ethnic groups after adjusting for each of the a priori specified potential confounders. The third column specifies the p-values associated with the adjusted analysis results.

*The “Other” race category consists of participants who identified as Native Hawaiian or Pacific Islander, unknown race, multiracial, or White participants with unknown Hispanic ethnicity.

Supplementary Tables X and Y. All adjustment variables are the same. Note that results are not the same across all the race and ethnicity groups. Black and Asian participants have different results in these two sub-analyses.

A.8 Acknowledgement:

We thank all of the participants and co-participants for participating in ADRC research and NACC data.

The NACC database is funded by NIA/NIH Grant U24 AG072122. NACC data are contributed by the NIA-funded ADRCs: P30 AG062429 (PI James Brewer, MD, PhD), P30 AG066468 (PI Oscar Lopez, MD), P30 AG062421 (PI Bradley Hyman, MD, PhD), P30 AG066509 (PI Thomas Grabowski, MD), P30 AG066514 (PI Mary Sano, PhD), P30 AG066530 (PI Helena Chui, MD), P30 AG066507 (PI Marilyn Albert, PhD), P30 AG066444 (PI John Morris, MD), P30 AG066518 (PI Jeffrey Kaye, MD), P30 AG066512 (PI Thomas Wisniewski, MD), P30 AG066462 (PI Scott Small, MD), P30 AG072979 (PI David Wolk, MD), P30 AG072972 (PI Charles DeCarli, MD), P30 AG072976 (PI Andrew Saykin, PsyD), P30 AG072975 (PI David Bennett, MD), P30 AG072978 (PI Neil Kowall, MD), P30 AG072977 (PI Robert Vassar, PhD), P30 AG066519 (PI Frank LaFerla, PhD), P30 AG062677 (PI Ronald Petersen, MD, PhD), P30 AG079280 (PI Eric Reiman, MD), P30 AG062422 (PI Gil Rabinovici, MD), P30 AG066511 (PI Allan Levey, MD, PhD), P30 AG072946 (PI Linda Van Eldik, PhD), P30 AG062715 (PI Sanjay Asthana, MD, FRCP), P30 AG072973 (PI Russell Swerdlow, MD), P30 AG066506 (PI Todd Golde, MD, PhD), P30 AG066508 (PI Stephen Strittmatter, MD, PhD), P30 AG066515 (PI Victor Henderson, MD, MS), P30 AG072947 (PI Suzanne Craft, PhD), P30 AG072931 (PI Henry Paulson, MD, PhD), P30 AG066546 (PI Sudha Seshadri, MD), P20 AG068024 (PI Erik Roberson, MD, PhD), P20 AG068053 (PI Justin Miller, PhD), P20 AG068077 (PI Gary Rosenberg, MD), P20 AG068082 (PI Angela Jefferson, PhD), P30 AG072958 (PI Heather Whitson, MD), P30 AG072959 (PI James Leverenz, MD).

The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.