Methods for Summarizing and Imputing High-Frequency Digital Biomarkers in the Context of Longitudinal Data Analysis

by

Nicole Irene Wakim

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in the University of Michigan
2023

Doctoral Committee:

Professor Tom Braun, Co-Chair
Professor Zhenke Wu, Co-Chair
Professor Roger Albin
Professor Veerabhadran Baladandayuthapani
Professor Hiroko Dodge

Nicole Irene Wakim

nwakim@umich.edu

ORCID iD: 0000-0003-4540-2946

ACKNOWLEDGMENTS

First, I thank my committee for their guidance. Thank you to my co-chairs Dr. Tom Braun and Dr. Zhenke Wu for meeting with me during the COVID-19 pandemic and helping me form a dissertation out of an analysis question. Thank you Dr. Zhenke Wu for accepting me as a mentee when my dissertation was only an idea. I am grateful for your expertise and guidance. Thank you Dr. Tom Braun for your persistent support through my personal and academic struggles. Your mentoring has been an invaluable source of statistical and personal growth. Thank you to my committee members: Dr. Veera Baladandayuthapani for your gentle advice and expertise in functional data analysis (FDA), Dr. Roger Albin for providing helpful insights to Neurology research, and Dr. Hiroko Dodge for a wonderful research assistantship that inspired this dissertation work.

I would also like to acknowledge all the other teachers, professors, and mentors that have supported my statistical growth. Thank you to Mrs. Julia Anton, who encouraged me to continue in a quantitative field. Thank you to the professors within the University of Michigan's Department of Biostatistics that directly or indirectly mentored me: Dr. Bhramar Mukherjee for promoting women in statistics, Dr. Brisa Sanchez for being an influential professor in the early days of my Master's degree, and Dr. Matt Zawistowski for showing me how to be a great biostatistics teacher.

Thank you to the administrative staff in the Department of Biostatistics office. Your support of me as a student has kept me on track to earning this degree. Thank you in particular to Nicole Fenech who fielded the large breadth of questions that I asked over my tenure.

Thanks to my PhD buddies, Dr. Holly Hartman and Margaret Banker. Without your friendship and help, my PhD career would not have been so enjoyable and successful.

Lastly, my family has supported me endlessly through my academic career. To my older brothers, Alex and Peter Wakim, for inspiring growth and providing sanctum. Anne and Paul Wakim, Mom and Dad, you raised me to believe in my mathematical competency and encouraged me to pursue my education. Thank you for supporting me. It means the world.

TABLE OF CONTENTS

LIST OF FIGURES

FIGURE

LIST OF TABLES

TABLE

LIST OF APPENDICES

LIST OF ACRONYMS

Aβ amyloid-β

ACF auto-correlation function

BLUP best linear unbiased predictor

CDR clinical dementia rating

CSF cerebrospinal fluid

FDA functional data analysis

GEE generalized estimating equation

GLMM generalized linear mixed effects model

HD Huntington disease

IS Impute then Summarize

ISAAC Intelligent Systems for Assessing Aging Change

IQR interquartile range

LDA longitudinal data analysis

LMM linear mixed effects model

MAR missing at random

MCAR missing completely at random

MCI mild cognitive impairment

MICE Multivariate Imputation by Chained Equations

MLR multiple linear regression

MMRM Mixed-Effect Model Repeat Measure models

MNAR missing not at random

MVPA moderate-to-vigorous physical activity

NP neuropsychological

ORCATECH Oregon Center for Aging and Technology

RMSE root mean square error

RNN recurrent neural network

SI Summarize then Impute

SNR signal to noise ratio

VAR vector auto-regression

ABSTRACT

Digital biomarkers have the potential to aid early detection of cognitive decline, resulting in interventions that delay dementia onset. Digital biomarkers are measured at high frequencies, which increases opportunities for assessment of cognitive ability. Summarization of digital biomarkers helps to reduce computational burden or variance among values. However, the consequences of summarization processes are not statistically investigated, especially the potential bias in longitudinal data analysis (LDA). This dissertation will systematically investigate methods for summarization and imputation of high-frequency digital biomarker data. These methods are created in the context of longitudinal analysis of a binary cognitive outcome with digital biomarker predictors.

In Chapter 2, we discuss factors of high-frequency digital biomarker data that need to be considered when faced with a time granularity decision, defined as the frequency at which measurements are observed or summarized. It is important to find a balance between ease of analysis by condensing data and the integrity of the data, which is reflected in a chosen time granularity. Via simulation, we investigate the factors and examine how each affects the ability to detect the true, underlying digital biomarker pattern using RMSE. Then we apply our procedure to the ISAAC study, which involves longitudinal walking speed. The example sheds light on typical problems data present and how we can use the above factors in exploratory analysis to choose an appropriate time granularity.

In Chapter 3, we examine the process of simultaneously imputing and summarizing digital biomarker data. When analyzing the association of digital biomarker data and a lower-frequency outcome, we want to summarize biomarker data to match the outcome's frequency while simultaneously imputing missing biomarker values. We define two methods for the imputation and summarization process: (1) Impute then Summarize (IS), where we impute at the finest time granularity, then summarize, and (2) Summarize then Impute (SI), where we summarize the biomarker then impute at the summary level. Via simulation, we assess two processes involving imputation of biomarkers for longitudinal analysis of a binary outcome. Our results show that accuracy of coefficient estimation depends on percent missing data, length of consecutive missing days, and the rate of trajectory change of the biomarkers. We apply these processes to the ISAAC study, and find that the odds of testing for mild cognitive impairment (MCI) is negatively associated with increased walking speed.

In Chapter 4, we investigate simultaneous imputation of multiple, potentially correlated digital biomarkers, and association analysis of one biomarker with a longitudinal outcome. We build off the two methods in Chapter 3 using an updated regression model that incorporates information from additional biomarkers through random effects. Via simulation, we vary levels of correlation between digital biomarkers and percent missing values of each digital biomarker to examine the effect on imputed digital biomarker values and relative bias of coefficient estimates from longitudinal analysis. the coefficient estimate of the longitudinal analysis. Our results show increased correlation has the biggest effect on reducing relative bias, followed by increased number of digital biomarkers. We also found that at low correlation levels, an increase in percent missing values of other digital biomarkers results in the decreased in magnitude of relative bias. We apply our imputation methods to the ISAAC study with two and four digital biomarkers, and find that imputed values do not improve with additional digital biomarkers.

CHAPTER 1

Introduction

Digital biomarkers are measurements of physical behavior that can signal an underlying biological phenomenon (Vasudevan et al., 2022). These measurements can include heart rate, steps per day, time on a computer, sleep time or quality, and more. Similar to biological biomarkers, changes in digital biomarkers can indicate disease progression. For example, physical symptoms like gait change can indicate progression in Huntington disease. Digital biomarkers by nature are measured using less invasive and expensive procedures than traditional biological biomarkers. Digital biomarkers are also measured at high-frequencies, often using sensors in the home or wearable devices. Thus, digital biomarkers are an unobtrusively measured and potentially rich source of data that can aid in disease diagnosis.

In the context of Alzheimer's disease and dementia, research shows that dementia onset often follows changes in physical behavior that indicate cognitive decline, such as changes in gait, sleep, speech, and motor activity (Kourtis et al., 2019). These physical behavior changes can occur 10-15 years before a diagnosis of Alzheimer's disease (Albers et al., 2015). For example, there is evidence that amyloid-β (Aβ), an established pathway of Alzheimer's disease, starts to build up in the brain up to a decade before clinical diagnosis of Alzheimer's disease (Sperling et al., 2011). This build-up affects brain function and structure that have downstream effects on a person's motor function. This means changes in motor function, like gait or walking variability, may indicate an increase in Aβ in the brain. Thus, the use of digital biomarkers to measure motor function can serve as proxy for Aβ build-up and cognitive decline.

Research aims to use high-frequency digital biomarkers to aid in early diagnosis of neurodegenerative diseases, including Alzheimer's disease (Dorsey et al., 2017). Currently, diagnosis of Alzheimer's disease relies on a neuropsychological (NP) test that is administered every six months or every year (Goldberg et al., 2015). Time between serial assessments prevents learning bias, or practice effect, meaning test takers will not improve their score from familiarity with the test. However, time between tests also means there are limited opportunities to diagnose Alzheimer's disease, including early stages like mild cognitive impairment (MCI). MCI describes cognitive decline that is worse than normal aging levels, but less severe than dementia levels. By using high-frequency digital biomarkers as a proxy, we can increase opportunities for diagnosis. This is especially important for diagnosis of MCI. Early-diagnosis of MCI allows for interventions that reverse declining cognitive ability. Thus, digital biomarkers can aid in prevention of full dementia onset.

In order to demonstrate high-frequency digital biomarkers' potential as a proxy for MCI, we analyze the association between MCI and clinically relevant digital biomarkers. We aim to use traditional longitudinal data analysis (LDA) methods, like generalized linear mixed effects model (GLMM), to analyze the association. However, the outcome (MCI) is measured at a lower frequency than the predictors (digital biomarkers), which presents a problem in LDA. In another field studying longitudinal data, functional data analysis (FDA) considers this discrepancy by modeling the digital biomarkers with a functional form by smoothing the data. Smoothing data minimizes the noise due to measurement error while maintaining the overall pattern of the data (Ullah and Finch, 2013; Wang et al., 2016). Different smoothing techniques are used to handle specific pattern characteristics of the data. For example, splines are often used for non-periodic data by splitting data into knots that can fit different polynomials for different segments of the data (Likhachev, 2017; Newell et al., 2006; Parker and Wen, 2009). The number of knots are chosen to optimize the fit of the splines while minimizing the number of knots. The smoothed data can be used as predictors or outcomes within the FDA framework. However, the advantageous process of smoothing does not translate well to traditional longitudinal analysis.

This dissertation focuses on summarizing digital biomarker predictors for traditional longitudinal analysis. Whether a statistician or clinician are preforming analysis, it is helpful to gauge association between an outcome and predictors before implementing more statistically vigorous analysis, like FDA. However, within traditional longitudinal methods, there is little specification on a summarization process for high-frequency predictors (Keadle et al., 2014; Zhang et al., 2019). We specifically aim to demonstrate methods for analysts using traditionally taught longitudinal analyses.

In Chapter 2, we examine the summarization process for high-frequency digital biomarkers through different time granularities. Time granularities are the frequency at which a digital biomarker is measured or summarized. Time granularities are determined by the summarization window size, which is the number of consecutive time points that a summary is computed over. By summarizing data, we reduce the data to a courser time granularity. Similar to smoothing in FDA, summarization can decrease the variability between digital biomarker values, but it can also misrepresent the true, underlying signal. Using a simulation study, we examine factors that determine an appropriate window size for summaries. We found that increased follow-up time, increased sine period, and increased biomarker standard deviation generally lead to an increased window size. However, it is important to examine a dataset's specific characteristics. Thus, we examine specific examples within the simulation study to demonstrate these nuances. Finally, we apply this process to a case study involving data from the Intelligent Systems for Assessing Aging Change (ISAAC) study. We step through the exploratory analysis needed to determinate a range of appropriate window sizes for summary. This work has been published without the simulation study in *Alzheimer's and Dementia: Translational Research and Clinical Interventions* (Wakim et al., 2020).

For both LDA and FDA, summarization or smoothing is further complicated in the presence of missing data. Studies with follow-up times spanning weeks to years are likely to have missing data. Thus, it is important for research to investigate appropriate imputation methods for high-frequency predictors, such as digital biomarkers.

Within FDA, work has been done to simultaneously smooth and impute data. Wavelet-based functional mixed models have been used to incorporate incomplete data into FDA by using missing wavelet coefficients that characterize the functional data (Morris et al., 2006). This method uses data from the entire follow-up time to impute and smooth. Another method developed by Leroux et al. (2018) uses observed functional data to impute future functional data for dynamic prediction.

Again, in the context of traditional LDA, little research has investigated simultaneously summarizing and imputing predictors. While studies have investigated imputing a high-frequency predictor, they have not discussed how the order of summarization and imputation of predictors affect coefficient estimates in LDA (Di et al., 2022). For example, in a simulation study performing imputation on predictors for an intensive longitudinal study, Ji et al. (2018) found that imputing predictors independently from the outcome, using autoregressive methods, produced the least biased coefficient estimates aside from the full data. However, the authors did not investigate other imputation methods or processes for the predictor. Another study examines imputation on the summary-level using classifications of observed, partially observed, or missing summaries (Tackney et al., 2021). While this study examines summarization and imputation of high-frequency accelerometer data, it does not discuss the order of summarization and imputation, and it uses the accelerometer data as a longitudinal outcome, not a predictor. Thus, there is a need for an investigation of the process of summarizing and imputing predictors in the context of LDA

In Chapter 3, we investigate the effects of simultaneously summarizing and imputing digital biomarkers on the association analysis with a lower-frequency binary outcome. For many clinicians and statisticians, the initial instinct when imputing missing data may be to impute at the finest time granularity, but work within FDA has shown that smoothing data (i.e. a form of summarization) may lead to less variable imputations (Leroux et al., 2018; Morris et al., 2006). For example, Doherty et al. (2017) discuss imputation of accelerometer data during non-wear times on the "one minute granularity" without discussing possible imputation on a courser time granularity. Thus, we examine two processes for imputing and summarizing: (1) imputing on the finest time granularity available then summarizing the observed and imputed values, and (2) summarizing the observed values into missing and observed summaries, then imputing the missing summary values. Using a simulation study with missing completely at random (MCAR) assumptions, we found that the process in which we impute then summarize is prone to attenuation bias of the coefficient estimates within GLMM with a binary outcome. Attenuation bias occurs when there is large stretches of missing data. The imputed digital biomarker values at the finest time granularity are more variable than the true missing values, leading to an increase of variance of our predictor. An increase in the predictor variance results in a decrease in the coefficient estimate towards zero. Again, we apply this process to a case study involving data from the ISAAC study. Our simulation study and case study are limited to a single digital biomarker as a predictor.

Individuals with Alzheimer's disease or dementia have many shared physical symptoms, including gait, deterioration of fine motor skills, heart rate variation, and sleep disruption (Kourtis et al., 2019). While many researchers investigate multiple physical features as predictors of cognitive impairment, they do not thoroughly investigate the relationships between features (Aggarwal et al., 2006; Scarmeas et al., 2005). Physical symptoms are clearly connected to each other through Alzheimer's disease and dementia. Thus, changing patterns of physical features may inform missing values of other physical symptoms. We continue our work on imputations of digital biomarkers by extending our methods to multiple markers.

In Chapter 4, we examine imputation methods to model correlation between digital biomarkers and examine scenarios in which correlated digital biomarkers improve imputations. Once again, we examine the imputations in the context of longitudinal analysis through GLMM. We propose an imputation method that models and imputes all digital biomarkers simultaneously through a linear mixed effects model (LMM) that incorporates a individual-specific random effect at a given time point. We implement a simulation study with varying levels of correlation, missingness, and number of digital biomarkers to examine the imputed values from our method. We found that relative bias was minimized when there was high levels of correlation and increased number of digital biomarkers. Similar to Chapter 3, we found that summarizing then imputing the data was a more robust process when correlation was lower and missingness was higher. We continue the case study with the ISAAC data by incorporating other digital biomarkers into the imputations.

CHAPTER 2

Statistical Approach to Selecting Time Granularity for High-Frequency Digital Biomarkers

2.1 Introduction

Digital biomarker data measure human characteristics that describe a person's behavior or physiology. Common examples include duration of sleep, steps per day, and heart rate. Like other biomarkers, digital biomarkers can be used to potentially identify underlying biological processes, including cognitive function, that may not yet have clear clinical symptoms (Akl et al., 2015; Austin et al., 2017; Buchman et al., 2012; Dodge et al., 2012; Eby et al., 2012; Gorus et al., 2008; Hayes et al., 2014; Kaye et al., 2012, 2014; Kaye et al., 2011; Lyons et al., 2015; Silbert et al., 2016). Digital biomarker data can be collected at high frequencies using readily available commercial devices. For example, Fitbit (Fitbit Inc, San Francisco, California, USA) activity trackers can measure on the order of seconds to produce daily step counts, sleep quality, and heart rate. Other studies use a system of sensors to record digital biomarker data. In one study, Eby et al. (2012) measured driving behavior of early stage dementia patients through in-vehicle technology and sensors that continuously recorded data while subjects were driving. In this study, sensors in the vehicles recorded measurements in interval lengths of no more than 1 second.

Trajectories of high-frequency digital biomarker data are increasingly used in health care research to monitor health status in support of disease prevention, treatment, and management (Dodge and Estrin, 2019). Use of digital biomarker data can be extended to disease progression of dementia, which is a process that is currently monitored with limited assessment frequency. For example, neuropsychological (NP) tests used to assess dementia can be administered at most every 6 months to reduce tasking participants and avoid learning and practice effects. As a result, these tests only provide two opportunities for diagnosis of dementia per year (Dodge et al., 2015). Trajectories of digital biomarker data can potentially help increase the opportunities for early diagnosis because assessments are made more frequently and can improve precision of person-specific trajectories.

However, digital biomarkers are noisy in their raw data format. The raw data typically include a series of time stamps with indicators that require processing before they can be used as intelligible measurements. For example, a computer sensor will have a series of indicators for its mouse movements, but those indicators must be translated into comprehensible variables like computer use start time, end time, or duration (Kaye et al., 2014; Seelye et al., 2018).

In this chapter we only refer to digital biomarkers that have been preprocessed into daily measurements. However, one may choose a different starting frequency for measurements if the raw data with finer measurements are available. For example, we may want to examine average hourly heart rate so that we can evaluate a patient's exercise regimen. The approach described in this chapter can still be applied to data on the hourly scale, or even minute or second scale, but for ease of demonstration, we will assume a daily scale. It is also important to note that the method for processing data may affect aspects of the data, and thus the choice of time granularity.

It is important to remember that as the frequency of measurements increases, the size of the resulting dataset also increases, which then requires a greater amount of computer storage space and computation time for analysis. Thus, it is useful to condense data, such as hourly measurements, to daily, weekly, or monthly summaries, to reduce both data storage demands as well as computation time. At the same time, we would like to maximize the chance of capturing clinically meaningful changes or shifts for each individual. Therefore, it is important to clarify the decision process used to determine the frequency, or time granularity, needed to analyze data.

Determining the appropriate time granularity of data is a complex process, one that has many possible variations. But, regardless of variations, the process to determine time granularity must be explicitly defined before the data are analyzed. It is inappropriate to use statistical significance, ie, P-values, as a metric for confirming the most appropriate time granularity. Doing so will only serve to inflate the rate of false-positive findings and may lead one to incorrectly assume that the model with the most statistically significant result is detecting the true underlying signal. Instead, one should make decisions of time granularity in the exploratory portion of analysis, using summary statistics and trajectory plots to help in the decision.

This chapter is organized in the follow manner. In Section 2.2, we address critical issues to be considered when we decide the level of time granularity to monitor longitudinal digital biomarker trajectories. In Section 2.3, we examine simulation scenarios over varying factors to assess time granularities. In Section 2.5, we examine a case study involving dementia research to demonstrate empirical differences between time granularities.

2.2 Factors of data impacting time granularity decisions

When investigating repeated digital biomarker data, one should first consider the statistical model that will be used. For examining trajectories, one typically uses longitudinal data analysis (LDA) to determine how disease progression relates to changes over time in the biomarkers. In this chapter, we will examine digital biomarkers in the context of linear mixed effects models (LMMs) and generalized linear mixed effects models (GLMMs). We consider digital biomarkers as potential predictors in these LMM and GLMM models that will be matched to a repeatedly measured outcome. For more information on LMMs, GLMMs, and LDA please refer to statistical textbooks (James, 2002; Verbeke and Molenberghs, 2008).

While considering digital biomarkers as predictor in LMM or GLMM, we must examine factors that help improve the fit. These factors are: (i) duration of follow-up, (ii) variables of interest in analysis, (iii) pattern detection, and (iv) signal to noise ratio (SNR), and each term will be defined in greater detail as it is presented in the following sections. It is important to note that the factors we have listed do not include data management or missing data issues. We assume storing and maintaining the raw data is not an issue, but we understand that this may motivate summarizing data as it is stored.

2.2.1 Duration of Follow-up

We recommend that readers initially investigate duration of follow-up, T, which is the length of time that measurements are recorded for each subject. Follow-up time will help determine which time granularity is not acceptable for examining trajectories. Time granularities that are close to the total follow-up time will not provide enough information for analysis because there will not be a sufficient number of data points to accurately model trajectories over time. Thus, these time granularities should not be considered further.

For example, one can measure the weekly estimated fetal weight for pregnant mothers. The maximum follow-up time for each mother is around 40 weeks, which is limited by the human gestation period. Therefore, we record 40 weekly measurements on fetal weight. If one wanted to condense these measurements into monthly averages, then one would have approximately 9 monthly averages. If one wanted to condense these measurements into trimester averages, then one would have three trimester averages. Lastly, one may calculate a single gestation average. Thus, there are 40 weekly measurements, 9 monthly averages, 3 trimester averages, and one gestation average. If trajectory analysis aims to detect fetal weight change over time, then one can immediately eliminate the single gestation average as a feasible time granularity because a single point cannot detect change. The other three time granularities are feasible, in terms of follow-up time alone, but it is necessary to consider the other factors.

In the case of fetal growth, the duration of follow-up is short enough that weekly measurements will not be an overwhelming amount of data for analysis, but reduction to monthly, trimester, or gestation averages may greatly reduce power to detect change over time. There is no exact cut off for follow-up time at which reduction of frequent measurements to courser time granularities becomes appropriate and is context dependent. It is important to remember that we reduce the data by approximately four-fold when we condense weekly data to monthly averages, so there must be some benefits to analyzing monthly averages to offset this reduction in the number of unique data points.

2.2.2 Variables in analysis

Variables of interest may influence selection of data granularity. For example, in Dodge et al. (2012), the authors found that the variability of a subject's walking speed, in addition to walking speed itself, is associated with mild cognitive impairment (MCI). In this example, the data were processed into daily measurements that included average walking speed. The variance was calculated over each week (variability of daily walking speed within each week). Because the authors used a week's worth of data to create the new variance variable, it is no longer viable to model the daily time increments. In this case, using data with weekly time granularity adds an informative variable to analysis that was not available with daily granularity.

Another example is measurement of moderate-to-vigorous physical activity (MVPA) using heart rate monitors. Heart rate monitors take time-stamped measurements of a person's heart rate in beats per minute. These time-stamped heart rates can stand alone as a measurement of someone's physical activity, but researchers have found a more useful summary of this data to analyze someone's physical activity (Fletcher et al., 2013). Using established thresholds, heart rate can be categorized into different levels of physical exercise. Moderate exercise is defined as a heart rate within 50-70% of an individual's maximum heart rate, and vigorous exercise intensity is 70-85% (*Target Heart Rates Chart* 2020). From heart rate data, one can summarize the time spent within MVPA in a single day. Within the field of physical exercise tracking, daily MVPA is an established summary, therefore it is often used to analyze trajectories of physical exercise (Butera et al., 2019; Ehlers et al., 2017; Shi et al., 2020).

We suggest the reader consult with literature in their field to identify evidence for informative variables that may require summaries of their current measurements. These measurements may include peak values, trough values, number of times hitting a threshold, ranges, area under the curve, and other summary statistics. Inclusion of summary variables may elevate the analysis and warrant a reduction in time granularity.

2.2.3 Pattern Detection

Before selecting a time granularity for analysis, one needs to understand the general pattern of outcomes. This should be assisted by clinical and biological knowledge in the field. If an outcome is generally understood to change slowly over time, then we will not lose the signal by condensing the data from daily to seven, 30, 60, or even 90-day summaries. However, if the outcome is generally understood to change rapidly, a coarse time granularity may not capture the pattern. For example, a person's melatonin level, which is associated with sleep, rises during night hours and lowers during day hours (Dollins et al., 1994). If we calculated an individual's average daily melatonin level, then we would lose the hourly pattern associated with sleep.

In the case of Alzheimer's research, it is understood that clinical evidence of cognitive decline does not change rapidly during pre-symptomatic stages (Jack and Holtzman, 2013). Clinical symptoms and changes occur on the order of years. Therefore, if we are examining trajectories of pre-symptomatic subjects or those with MCI, we believe using condensed time granularities, like weekly or monthly data, is an acceptable approach, but it is important to consider all the factors discussed before making a final decision.

2.2.4 Signal to Noise Ratio and Within-Individual Variability

SNR describes the ratio of the underlying signal, or pattern, of the data compared to the noise in the data, which is sample to sample variability. A higher value of SNR indicates stronger belief that the observed signal is real and not random artifact. For longitudinal analysis, SNR for a trajectory is the ratio of (i) the expected pattern over time divided by (ii) the standard error for that pattern. For the exploratory analysis involving LMMs, we keep our examination of digital biomarkers simple. Often, one will plot the digital biomarker over time. To gauge the linear time pattern, we can fit a simple exploratory model, with time as the independent variable and our digital biomarker as the dependent variable. Thus, the expected pattern is solely dictated by the coefficient for time. If one includes more variables into the model then the expected pattern over time will be a linear combination of those coefficients. In Figure 2.1, we present a simple LMM ratio as an equation that we break into smaller components. We note that these equations are used to facilitate explanation, and are not meant to be used for actual calculations.

In theory, the data at different time granularities are attempting to uncover the same, true pattern, so the expected pattern over time should be the same for each time granularity. However, we do not know the true pattern, so we must compare the signal of each dataset drawn from different time granularities to each other. Typically, one plots a non-linear fit for data to visually gauge the signal and make comparisons. In general, we want the non-linear fit of each time granularity to resemble the fit of others. If data quality is good, there should not be major disparities between fits produced by different levels of time granularities. However, data with missing or extreme values are susceptible to biased summaries, leading to different trajectories. Thus, it is important to identify sections of data with missing data or extreme values. These sections should not factor

$$SNR = \frac{Expected\ value\ of\ slope}{Standard\ error\ of\ slope}$$

$$\downarrow$$

$$SNR = \frac{Expected\ value\ of\ slope}{\frac{Total\ variability}{\sqrt{Sample\ size}}}$$

$$\downarrow$$

$$SNR = \frac{Expected\ value\ of\ slope}{\frac{v_{bi} + v_{wi}}{\sqrt{Sample\ size}}}$$

$$v_{bi} = variance\ between\ individuals$$
$$v_{wi} = variance\ within\ individuals$$

Figure 2.1: Three representations of the SNR equation. For LMMs and GLMM, SNR is the ratio of the expected value of slope to the standard error of the slope. The standard error can be broken down further in two steps. Our final equation for SNR includes two important sources of variance (within and between individuals) and the square root of the sample size.

into the comparison between non-linear fits of different time granularities. We will elaborate on data quality in the discussion section. For now, one should assume that there is no missing data within the dataset.

We will examine standard error further to inform our choice between time granularities. The standard error of the slope is inversely proportionate to the SNR, so a higher standard error will result in a lower, or worse, SNR. We will not make any absolute claims about direct comparison of standard error between time granularities. Instead, we aim to establish an understanding of the trade-offs for variables of different time granularities by examining the components of the standard error. In the second equation of Figure 2.1, we see that standard error is the ratio of two components: (i) total variability and (ii) the square root of the sample size.

Within total variability, as we saw in the third equation of Figure 2.1, there are two important sources for repeated measures over time: variability between individuals and variability within an individual. Variability between individuals is how closely the data of one individual resemble that of any other individual. In Figure 2.2a, we can see a group of individuals that have relatively low between-individual variability as compared to that of Figure 2.2b. In Figure 2.2a, individuals

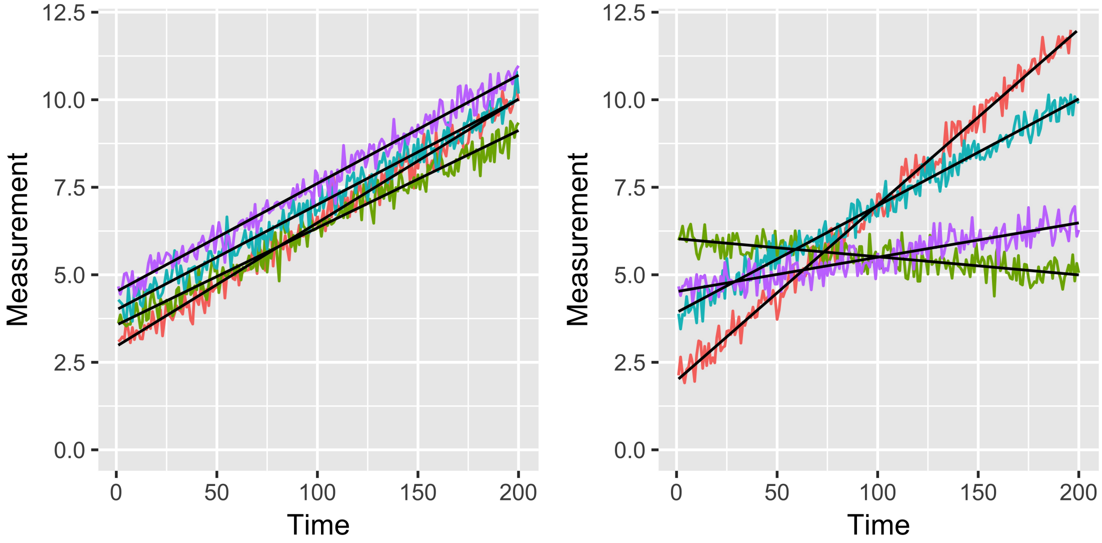


Figure 2.2: Examples of potential longitudinal data for four subjects. Subjects are represented by different colors. The linear model used to generate the data is displayed in black. Data were generated with equal variation within subjects. (a) Case where variability between the subjects' underlying linear model is low. (b) Case where the variation between subjects' linear model is high.

have similar slopes and values of their linear fit, while in Figure 2.2b, the slopes are different for each individual. That is, between-individual variability is much larger in Figure 2.2b than that of Figure 2.2a. Between-individual variability has limited influence on the decision of time granularity because individual trajectories will generally be maintained across time granularities.

Variability within an individual describes the closeness of an individual's data point to any other of the same individual's data points. When one uses summaries from coarse time granularity data, the difference between one individual's digital biomarker values changes, and typically decreases. In Figure 2.3a, we see that the data closely follow the linear fit for each individual. However, in Figure 2.3b, we see a larger spread of data around the same individual fits, indicating larger within-individual variability. For both sources, within and between, if the data are more variable, the support for any estimate replicating the true value is lessened because SNR is smaller.

The second important part of standard error is the sample size. With coarser time granularities, the sample size decreases. For example, the sample size for weekly data is 7 times less than that of daily data. If we consider weekly and daily data, each that have the same total variability, then the standard error for any estimate produced from the weekly data is 2.64 (the square root of 7) times

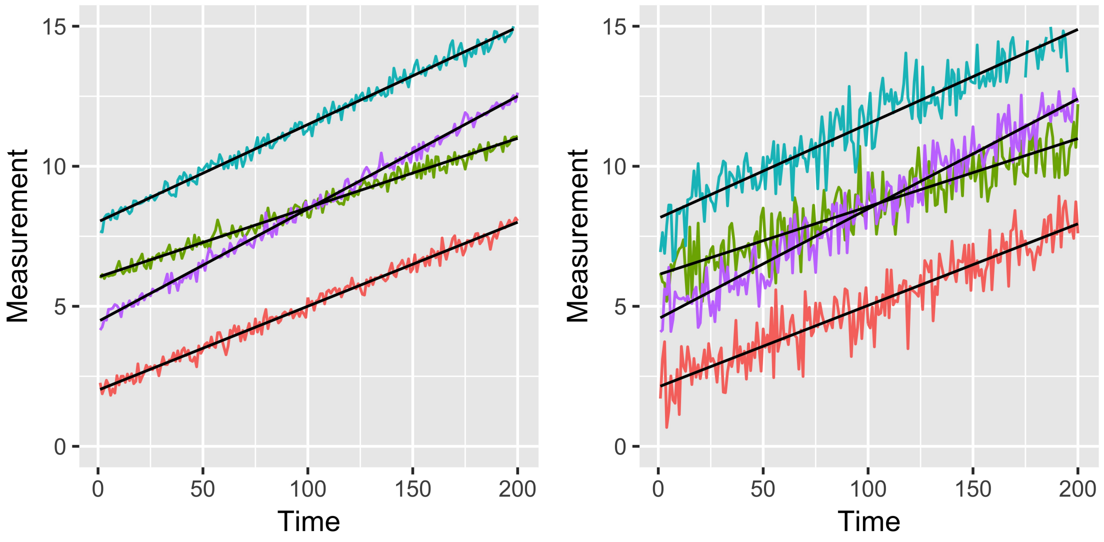


Figure 2.3: Examples of potential longitudinal data for four subjects. Subjects are represented by different colors. The linear model used to generate the data is displayed in black. Data were generated with same linear model. (a) Case where variation within subjects is low. (b) Case where variation within subjects is high.

greater than the corresponding estimate produced from the daily data, which leads to a smaller, or worse, SNR for weekly data. This thought process can be applied to the comparison of any two levels of time granularity.

However, the standard error of estimates produced from daily data will not always be smaller than those produced from weekly data because the standard error comprises the variability in the numerator and the square root of the sample size in the denominator. Often, a coarser time granularity reduces the variability of the data in addition to the sample size. The reduction in variability needs to counteract the resulting reduction of the sample size. For example, when condensing daily data to weekly data, the weekly data must have variability that is more than 2.64 (square root of 7) time smaller than the daily data to make the reduced granularity's SNR comparable.

Although one cannot directly compare the standard errors of each time granularity before fitting the models, one can use visualizations of the data to help determine which has less noise. This process will be similar to the one described for Figure 2.3. This will help direct us towards the more appropriate time granularity, but it is important to keep in mind that more than one time granularity can have low enough noise to strongly detect the signal. It is also important to note that there is no strict cutoff when one compares SNRs, one needs to balance SNR with other factors presented previously.

2.3 Simulation Study

For our simulation, we focus on one variable of interest, a single digital biomarker that is normally distributed without missing data and examine average values for each given time granularity. Any unit of time can be used for the finest granularity, but we will assume time is measured in days. Thus, daily measurements are the finest time granularity available. We have N individuals and each can be followed for a maximum of T days. For individual $i = 1, 2, \dots N$, we let m_{it} denote the value of the biomarker at day $t = 1, 2, \dots T$. We let $\mathbf{m}$ denote the $(N \times T)$ matrix of biomarker values for all N individuals, with row t of $\mathbf{m}$ equal to $\mathbf{m}_{\cdot t}$ and column i of $\mathbf{m}$ equal to $\mathbf{m}_{i\cdot}$.

We denote different time granularities with window sizes, w, in days. Note that time granularity and summary window size will be used interchangeably. For example, daily granularities use $w = 1$, while weekly time granularities use $w = 7$. In our simulations we examine $w = 1, 7, 30, 60, 90$. We compute the averages for each time granularity using the window of time points immediately preceding the given time point. For individual $i = 1, 2, \dots N$, we let $\overline{m}_{ij}^{w} = \sum_{k=1+w(j-1)}^{wj} m_{ik}/w$ denote the average of the biomarker for window $j = 1, 2, \dots T/w$ and window size w. Note that the number of summary biomarker values calculated is dependent on the window size. For example, if we have $T = 120$ days of marker data, then a window size of 1 day will result in 120 biomarker values, but a window size of 30 days will result in 4 biomarker summary values.

Section 2.2 focuses on four different factors: variables of interest, follow-up time, pattern of signal, and SNR. For variables of interest, our simulation study focuses on average value of a single digital biomarker that is normally distributed. For the other three factors, we simulate different scenarios to examine the effect of follow-up time, pattern of signal, and SNR on the average biomarker values at different time granularities.

2.3.1 Generation of data

For individual $i = 1, 2, \dots N$, we let μ_{it} denote the true, systematic value of the biomarker at day $t = 1, 2, \dots T$. We generate μ_{it} as

$$\mu_{it} = a_i + q_i t + b \sin \frac{t}{d} \tag{2.1}$$

in which the subject-specific intercepts $a_1, a_2, \dots a_N$ are independent and normally distributed with mean 0 and standard deviation 1, the subject-specific linear time effects $q_1, q_2, \dots q_N$ are independent and normally distributed with mean 0 and standard deviation 0.01; and d is the period of the sine function divided by 2π. Note, we will refer to μ_{it} as the true marker value.

In Equation 2.1, we vary values of d to change the period ($2\pi d$). Larger values of d leads to changes in the marker that are slower over time. We look at values of $d = 5, 25, 50, 100, 150, 200, ..., 400$. We also vary values of T to examine different follow-up times. We examine follow-up time from 0 to 500 days in increments of 50 days.

For the marker value, we want to include noise. Thus, we generate $m_{it} = \mu_{it} + \epsilon_{it}$, in which $\epsilon_{1t}, \epsilon_{2t}, \dots \epsilon_{Nt}$ are independent and normally distributed with mean 0 and variance σ_m^2. We vary values of σ_m^2 from 0 to 1 in increments of 0.1 in order to examine various amounts of SNR.

In Figure 2.4 we present an example of one individual's simulated data for a 500-day follow-up time. Note, simulated data for shorter follow-up times are a subset of simulated data for larger follow-up times. For example, simulated data for a 300-day follow-up time are the first 300 data points of the simulated data for a 500-day follow-up time. Each column in Figure 2.4 represents one of four periods, 10π, 50π, 200π, and 800π, used to generate the data. Each row in Figure 2.4 corresponds to five values 0, 0.2, 0.4, 0.6, 0.8, and 1 for the standard deviation (σ_m) of the error (ϵ_{1t}) used to generate the marker data, m_{1t}. For each plot of simulated data, the red line represents the mean marker values, μ_{1t}. Thus, the red line is the same for each plot in the same column because only the standard deviation of the error changes within a column.

2.3.2 Analysis

To assess time granularities, we compare the average biomarker value of a given granularity, $\overline{m}_{ij}^{w}$, to the true value of the marker, $\mu_{i,jw}$, at a given time point. Recall that we compute an average biomarker value using the preceding w time points. For example, a 7-day average ($w = 7$) at time point $t = 14$ would correspond to the second window summary ($j = 2$), and the average marker value, $\overline{m}_{i2}^{7}$, would be based on biomarker values from day 8 to day 14. The average marker value,

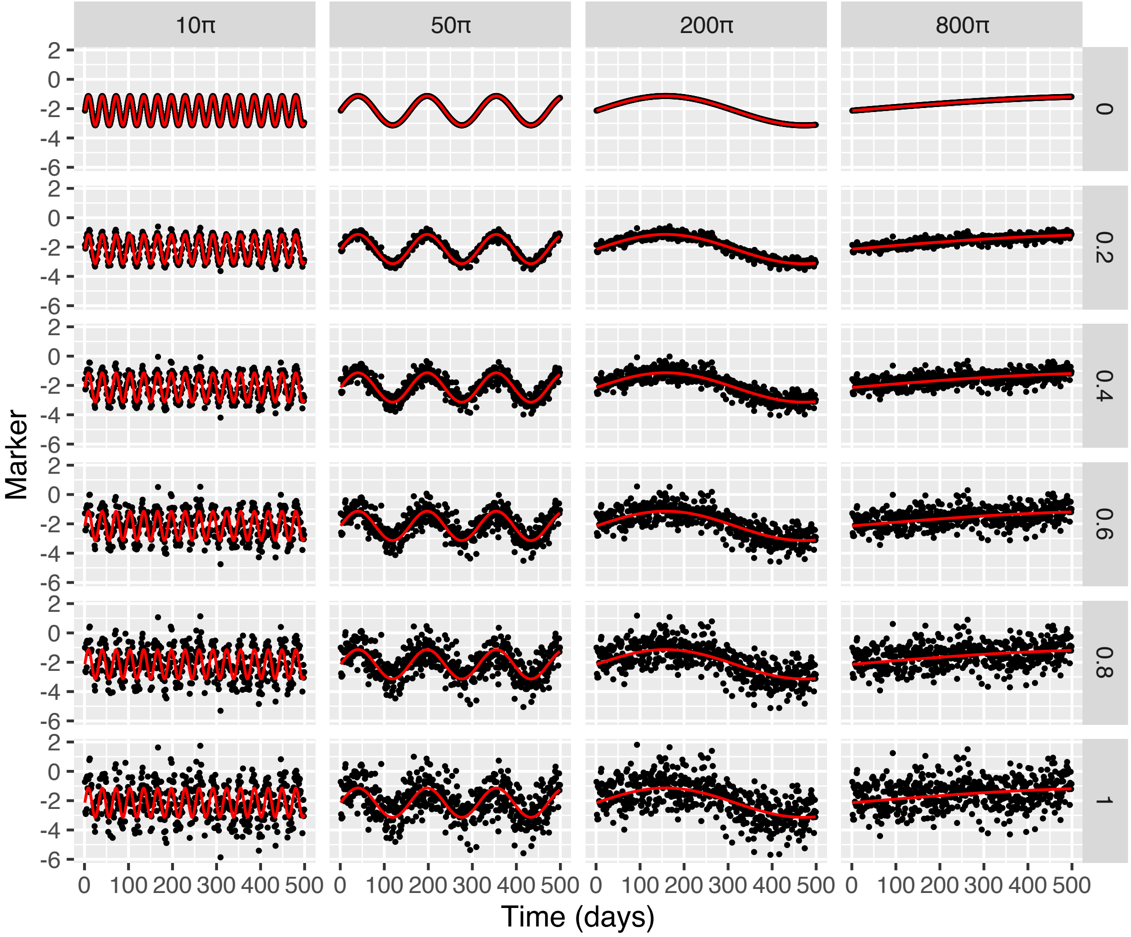


Figure 2.4: Sample of one individual's simulated data for a 500-day follow-up time. Each column represents a different sine period used to generate the data from 10π to 800π. Each row represents the standard deviation of the error added to the true marker value (from 0 to 1). The red line represents the true, underlying marker signal (μ_{it}). The black dots represent the simulated marker value with error (m_{it}).

$\overline{m}^7_{i2}$, will be compared to the corresponding true marker signal at day 14, $\mu_{i,14}$.

We use root mean square error (RMSE) to quantify average overall difference between the average biomarker value of each window and the true value at the end of the window. Specifically, we have

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}\sum_{j=1}^{J}(\overline{m}^w_{ij} - \mu_{i,(jw)})^2}{NJ}} \tag{2.2}$$

which will vary with four quantities: follow-up time, sine period, marker standard deviation, and window size. For a given setting defined by a value for each of the four quantities, we calculated average RMSE and 95% confidence interval of the RMSE based upon 1,000 simulations.

2.3.3 Results

Recall that the simulation study uses RMSE to quantify the impact of three factors: follow-up time, pattern of signal, and SNR. Our aim is to investigate which window values appropriately average daily marker values (as measured by RMSE) when factors change. However, there is an inherent relationship between the factors. Thus, to present these relationships clearly, we present two sets of figures where we hold one factor constant and examine changes in the other factors. We also examine how the window size of our summary average relates to these factors.

In Figure 2.5, we present average RMSE for a follow-up time of 500 days. Each separate plot examines a different sine period (pattern of signal) from 10π to 800π. From Figure 2.5, we see that as the standard deviation of the marker error increases, regardless of sine period, the average RMSE increases. For example, the 30-day average for a sine period of 200π starts with an average RMSE of approximately 0.1 and monotonically increases to approximately 0.23. Additionally, with a one-day average (i.e. the marker value at a single day), the average RMSE mirrors the standard deviation added to the marker.

In Figure 2.5, we also see how averages over multiple days help decrease overall RMSE for larger marker standard deviations and larger sine periods. In the plot for 10π, we see that for lower marker standard deviations (approximately less than 0.4) the seven-day average has higher RMSE than the one-day average. Within the RMSE equation, the average marker values are compared to the true marker value, μ_{ij}, at time point j. Thus, for smaller marker standard deviations, the

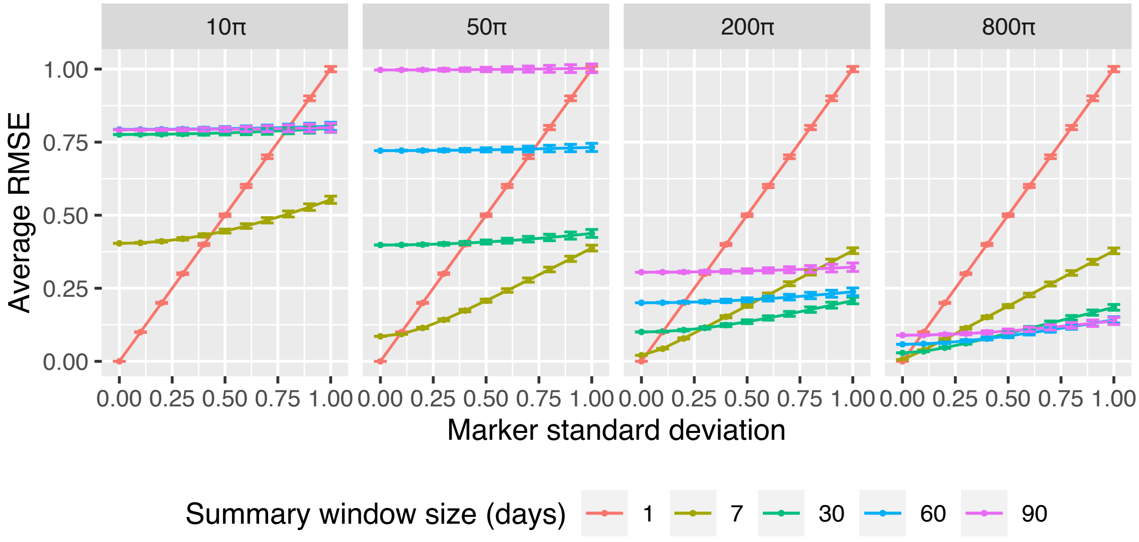


Figure 2.5: Average RMSE for various summary window sizes, holding follow-up time at 500 days. Window sizes are represented by different colors in each plot. Four plots of RMSE are shown for four selected period sizes (10π, 50π, 200π, and 800π). Marker standard deviation is shown on the x-axis. 95% confidence intervals are presented using error bars.

one-day average (at j) is the marker value (m_{ij}), which is close to the *true* marker value (at j). However, the seven-day average contains seven marker values close to their respective true marker value (i.e. the marker value at time point $j - 6$ is close to the true marker value at $j - 6$), but the true marker value is changing over time (with sine period of 10π). Thus, the average of the seven marker values (at time points $j-6$ to j) may not well represent the single, true marker value (at time point j). However, as the standard deviation of the marker increases, the RMSE for a seven-day average increases at a slower rate than the one-day average as seen in Figure 2.5. This is because the standard deviation of an average is inversely proportional to the number of values averaged over, which is reflected in a smaller average RMSE. As discussed in Section 2.2.4, averaging over more time points reduces the variability between averages. Thus, the seven-day average has less variance around the true marker value than the one-day average.

In Figure 2.5, as the sine period increases for each window size greater than one, the respective intercepts of the trends of average RMSE get closer to zero. For the 60-day window, and the 10π period, the average RMSE for zero standard deviation is 0.8, while at the 800π period and marker standard deviation of zero, the 60-day window has a RMSE approximately 0.05. For window sizes greater than one, we are averaging over multiple days. Thus, for a smaller period, with quick

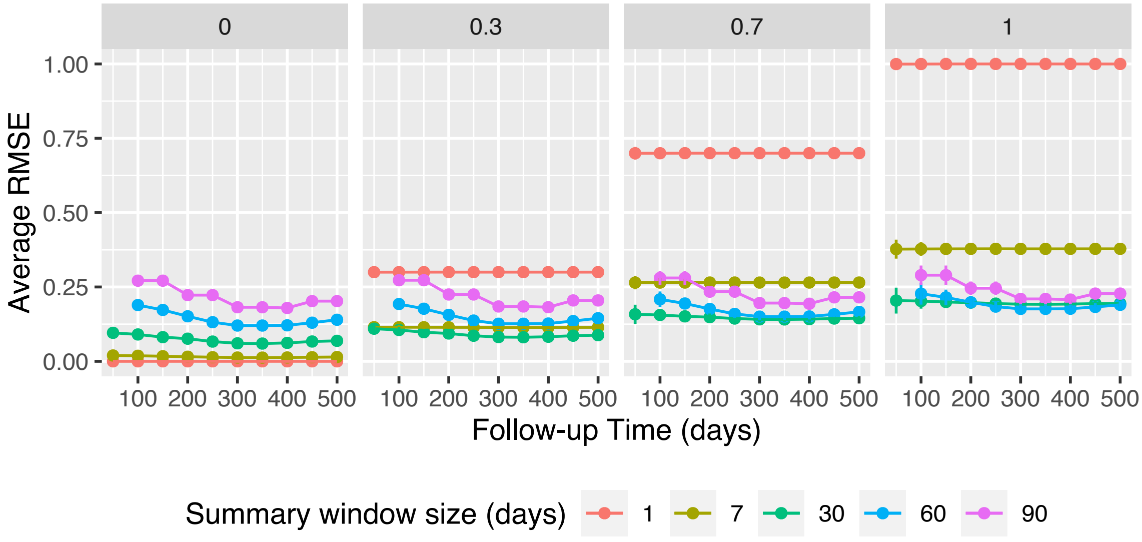


Figure 2.6: RMSE for various summary window sizes, holding sine period at 300π. Window sizes are represented by different colors in each plot. Four plots of RMSE are shown for four selected marker standard deviations (0, 0.3, 0.7, and 1). Follow-up time is shown on the x-axis and measured from 50 to 500. 95% confidence intervals are presented using error bars. If average RMSE values are missing, then follow-up time was less than the needed time points for a summary.

changes in the true marker value, averaging over multiple days means we can summarize changing values into one value that does not necessarily capture the changing marker values. Thus, our average is too granular to detect the true, underlying mean. However, for a larger period, the true marker value is not as variable between immediate, consecutive time points. Thus, an average taken over multiple days can effectively, as measured by the RMSE, detect the underlying, true signal. In this case, we reduce the variability of the averages distributed around the true, underlying marker value while still capturing the signal.

In Figure 2.6, we present average RMSE for a sine period of 300π; each separate plot examines a different marker standard deviation from 0 to 1 (labelled at top of plot). The x-axis represents the follow-up time of the marker. Each window size is represented by different colors. Thus, we can compare how marker standard deviation and follow-up time relate to the average RMSE of summary averages by their window sizes. From Section 2.2, we know that a 50-day follow-up time cannot accommodate a 60-day and 90-day average. Thus, there are no average RMSE for those window sizes.

In Figure 2.6, for the plot containing the marker standard deviation of one, the summary win-

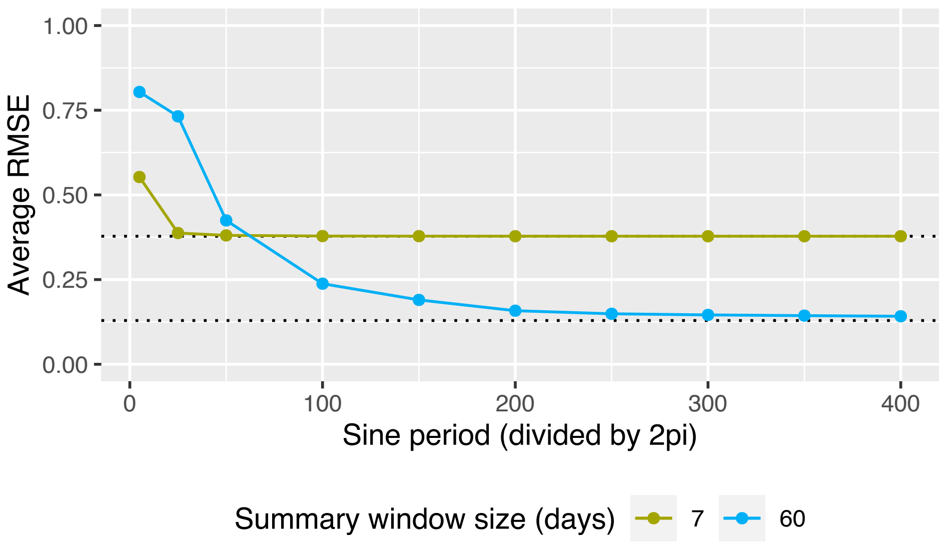


Figure 2.7: RMSE for various sine periods, holding marker standard deviation at 1 and follow-up time at 500 days. Windows sizes are represented by different colors. The black dotted line represents the theoretical standard deviation for a given window size.

dows greater than one have lower average RMSE than the one-day average. This may lead us to incorrectly conclude that a summary of any size is more appropriate than the single marker value. This observation highlights two important considerations. First, we must compare window sizes in the context of all three factors. Figure 2.6 is fixed at a sine period (300π) that is greater than all the summary window sizes; thus, we can capture the pattern. Thus, we can only conclude that for a period of 300π and standard deviation of one, the average marker value of any size is more appropriate than a single day's marker value. Second, the comparison between summary windows is not always the best way to measure how well a summary window is estimating the marker trajectory. Recall in Section 2.2.4, we discuss the advantage of using an average over seven days compared to a single day's value. We used $\sqrt{7}$ (approximately 2.65) to describe the needed decrease in standard deviation between weekly averages to counteract the decrease is sample size. To improve SNR for weekly averages compared to single day values, the average RMSE needs to approach $1/\sqrt{7}$. For each window size, we are interested in how stable the average RMSE is around $1/\sqrt{w}$. This demonstrates that the window average has lower variance between average values while capturing the true, underlying pattern.

For example, in Figure 2.7, we look at a follow-up time of 500 days, marker standard deviation of 1, and a window size of seven or 60 days. We drew a black dotted line to represent $\sigma_m/\sqrt{w}$, which is equal to $1/\sqrt{7}$ and $1/\sqrt{60}$, respectively for the seven day and 60-day windows. For both window sizes, as we approach longer sine periods, the RMSE approaches $1/\sqrt{w}$. However, the seven-day average stabilizes at $1/\sqrt{7}$ at a lower sine period than the 60-day average stabilizes at $1/\sqrt{60}$. Thus, for a given follow-up time, marker standard deviation, and sine period, we want to find the window size with an average RMSE that is stable within an acceptable range of $\sigma_m/\sqrt{w}$ and less than the average RMSE of other window sizes.

2.4 Simulation Examples

In the previous section we investigate various scenarios to identify general patterns of average RMSE across different follow-up times, sine periods, and marker standard deviation. In this section, we take an in-depth look at three specific scenarios within our simulation study. For each example, we demonstrate how to identify a range of appropriate summary window sizes more extensively. Although each example is a specific scenario, we present three scenarios that represent similar relationships between window size and average RMSE for other scenario settings. We discuss how window size balances accuracy and efficiency when averaging marker values over the window. For each scenario, that balance is optimized at different ranges of window sizes. For each example presented, we followed the process for simulating data from Section 2.3, except using window sizes ranging from 1 to 100, with a 5-day incremental increase starting at a 5-day window size. We also limited the number of simulations to 10.

For the first example, the scenario follow-up time is 300 days, the sine period is set to 200π, and the marker standard deviation is set to 0.7. Recall, readers may reference Figure 2.4 for the approximate visualization of each example scenario. In Figure 2.8, we have plotted the average RMSE against window size. For these specific scenario settings, we can now see which window size corresponds to the minimum average RMSE. Typically, a minimized average RMSE indicates that the window size is efficient, meaning that the average marker value calculated in the window has a decreased variance between the window averages. A minimized average RMSE also indicates that the average marker value calculated in the window is accurate, meaning the average marker value is close to the true, underlying mean value at that time point, which we specified in Equation

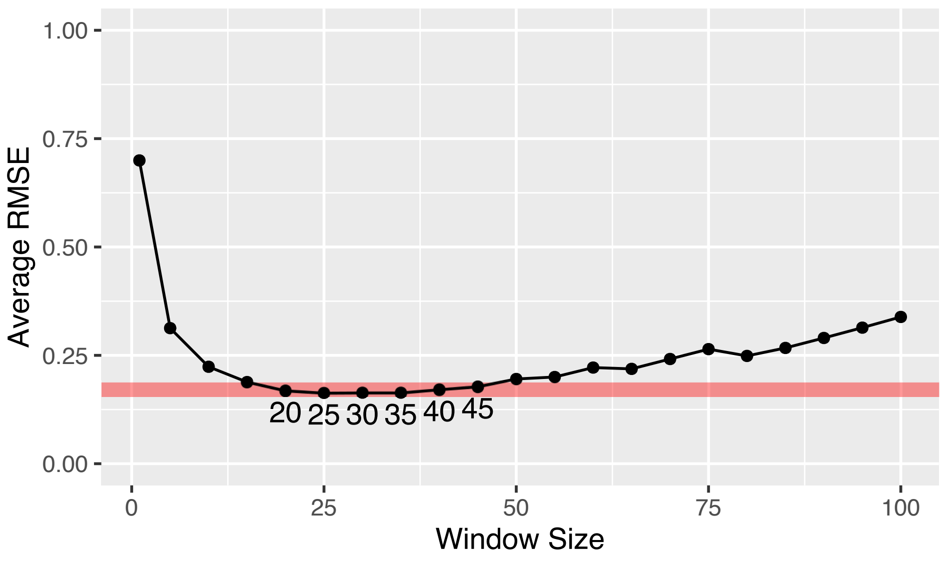


Figure 2.8: Average RMSE versus window size for setting with 300-day follow-up time, 200π period, and 0.7 marker standard deviation. The red shaded region corresponds to the range of average RMSE from the minimum value corresponding to the 25 day window size to a 15% increase in that minimum value. Window sizes that fall within the red shaded region are identified.

2.1. In this example, the window size of 25 days corresponds to the minimum average RMSE. Within Figure 2.8, we have shaded a region that corresponds to the range of average RMSE from the minimum value corresponding to the 25 day window size to a 15% increase in that minimum value. We identify the window sizes from 20 to 45 days that are within that average RMSE region. Outside of window sizes from 20 to 45 days, the average marker value has either lost efficiency or accuracy. For window sizes less than 20 days, there is a loss of efficiency. The variance between average marker values calculated for these windows is large, resulting in increased squared-error between the average and the true, underling mean at a given time point. For window sizes greater than 45 days, there is a gain in efficiency, but a loss in accuracy. The variance between average marker values calculated for these windows is small, but the window size is too coarse to gauge the changes in the true, underlying mean that occur within the window. Thus, window sizes from 20 to 45 days optimize accuracy and efficiency, and are suitable for summarizing markers in this scenario.

For the second example, the scenario settings are the same as the first example, but the sine period is smaller. The scenario follow-up time is 300 days, the sine period is set to 10π, and the marker standard deviation is set to 0.7. In Figure 2.9, we have plotted the average RMSE

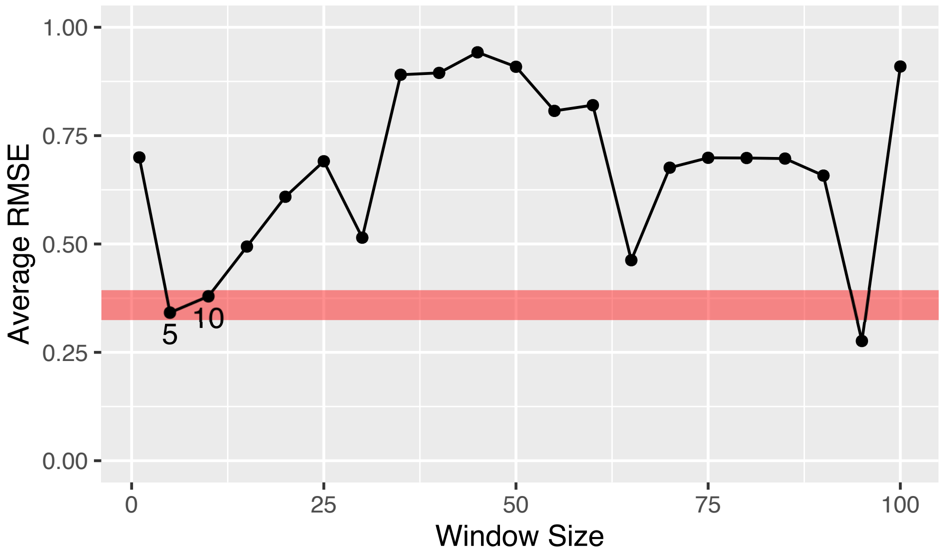


Figure 2.9: Average RMSE versus window size for setting with 300-day follow-up time, 10π period, and 0.7 marker standard deviation. The red shaded region corresponds to the range of average RMSE from the minimum value corresponding to the 5 day window size to a 15% increase in that minimum value. Window sizes that fall within the red shaded region are identified.

against window size. In this setting, the window size with the minimum average RMSE is 95 days. However, in Figure 2.9, we see that the average RMSE of the 90-day and 100-day windows are both more than 15% greater than the average RMSE corresponding to the 95-day window. Additionally, we see see similar decreased average RMSE at the 30-day window and 65-day window. These window sizes correspond to multiples of the period, 10π. In these instances, the windows sizes correspond to lower average RMSE because they do not align with the trough or crest segments of the sine function. These window sizes only include portions of the sine function where there is monotonic increases or decreases in the sine function, and thus, the calculated average summaries have lower RMSE values. Therefore, we identified the 5-day window as the optimal window corresponding to the minimum *stable* average RMSE value. At this window size, there is a clear decrease of average RMSE from the 1-day window indicating that the 5-day window is more efficient. We shaded a region that corresponds to the range of average RMSE from the value corresponding to the 5 day window size to a 15% increase in that minimum value. We identify the window sizes from 5 to 10 days that are within that average RMSE region. We deem this range of window sizes as acceptable for the marker summary.

For the third example, the follow-up time and sine period are increased compared to the first

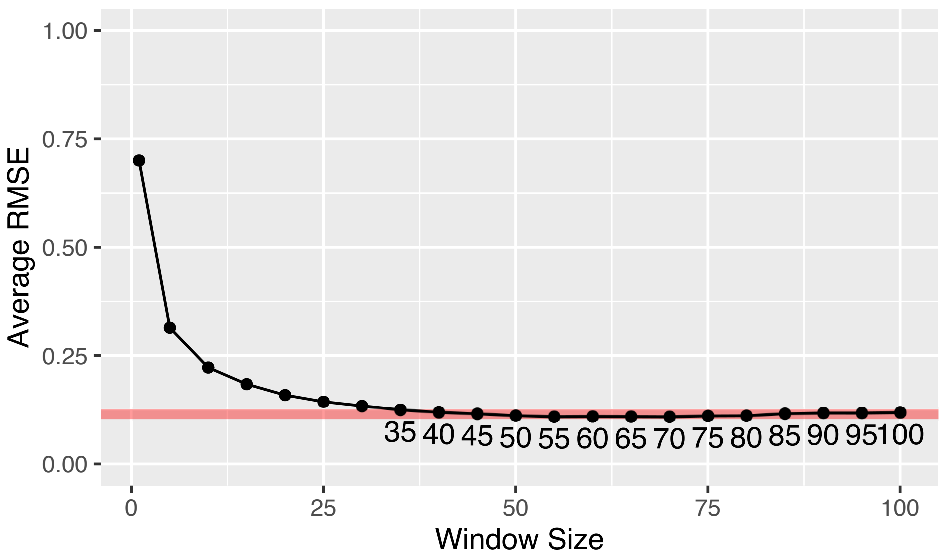


Figure 2.10: Average RMSE versus window size for setting with 500-day follow-up time, 800π period, and 0.7 marker standard deviation. The red shaded region corresponds to the range of average RMSE from the minimum value corresponding to the 65 day window size to a 15% increase in that minimum value. Window sizes that fall within the red shaded region are identified.

and second example. The scenario follow-up time is 500 days, the sine period is set to 800π, and the marker standard deviation is set to 0.7. In Figure 2.10, we have plotted the average RMSE against window size. For these specific scenario settings, we can now see which window size minimizes the average RMSE. The window size with the minimum average RMSE is 65 days. Within Figure 2.10, we have shaded a region that corresponds to the range of average RMSE from the minimum value corresponding to the 65 day window size to a 15% increase in that minimum value. We identify the window sizes from 35 to 100 days that are within that average RMSE region. In this example, the range of window sizes is so large because the sine period is large. Thus, the underlying, true mean marker value slowly increases over time without changing pattern within a window. This means the average values for large window sizes are close to the underlying, true mean, and maintain accuracy while gaining efficiency by increasing the window size. By identifying an above scenario that closely reassembles a clinician's dataset, we can investigate average RMSE versus window size, and thus, identify an acceptable range of window sizes.

2.5 Case Study

We now examine data from a longitudinal cohort study, the Intelligent Systems for Assessing Aging Change (ISAAC) study, which have been collected at the Oregon Center for Aging and Technology (ORCATECH) at the Oregon Health & Science University. More details of this study have been previously published (Gorus et al., 2008; Hayes et al., 2014). In the ISAAC study, walking speed was generated using a series of four motion sensors on the ceiling of a narrow hall that were triggered when a subject passed directly underneath (Dodge et al., 2015). Walking speed measurements were collected on each participant at multiple time points each day (whenever subjects passed under the series of sensors), and these multiple measures were processed into a daily mean walking speed. We use these daily mean walking speeds in our example. To facilitate discussion about time granularity, we have also summarized the daily measurements into weekly ($w = 7$) and 30-day ($w = 30$) values. We follow the summary computation defined in Section 2.3. To simplify our discussion, we have selected the data from four individuals in ISAAC to highlight the factors presented in Section 2. We selected these data simply to illustrate the factors that are important for granularity; it is important to realize that the conclusions in this section are not necessarily generalizable to the remaining data in ISAAC or other studies.

Figure 2.11 contains two columns of three figures each, both columns contain loess fits. The loess fit is a non-linear fit produced from overlapping, moving windows of data. Within each window of data, a weighted average is computed such that data closer to the center point are weighted more heavily than points further from the center. The weighted averages give a rough estimate of the smoothed trajectory, or signal, for each individual. The left column contains the loess fit with the data points, and the right column contains the loess fit with confidence bands. We can see that high frequency data with high within-individual variability (Figure 2.11a) may not help the viewer digest the general pattern over time.

With regard to the factors presented in Section 2.2, we first see in Figure 2.11 that the amount of follow-up time for the four individuals ranges from 3.8 years to 9.2 years. Thus, we could consider reducing the time granularity to seven-day or 30-day summaries because we would still have a sufficient number of summary values for each individual. Second, for our specific analysis, we want to consider an additional variable that needs to be derived from the data. This additional

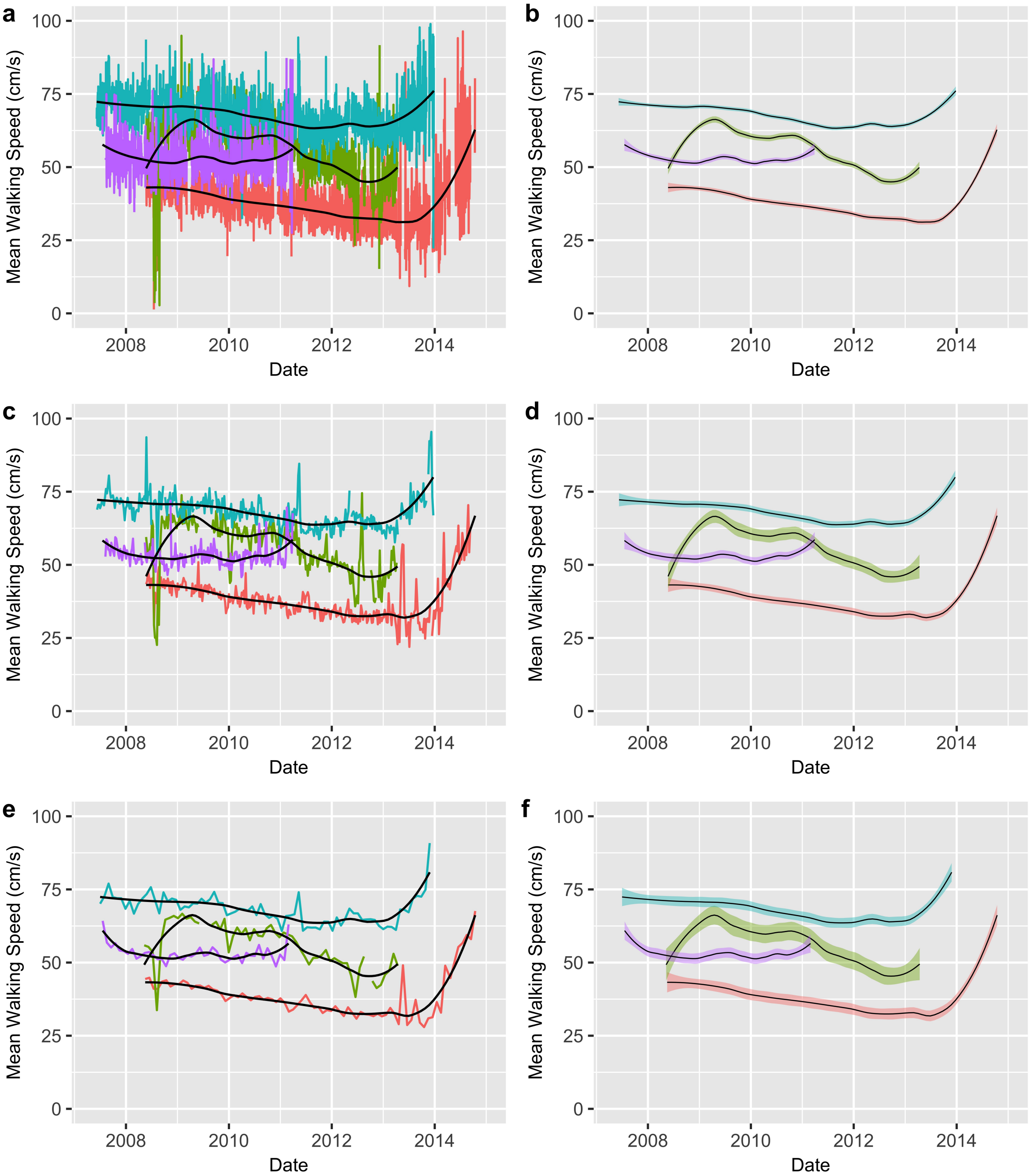


Figure 2.11: Example longitudinal data from four individuals from the ISAAC study. Each individual is marked with a different color for clarity. Mean walking speed is the average speed for an individual for the day (in cm/s). All figures have their loess curve in black. The left-hand column, including (a), (c), and (e), are the plots of the data. (a) is daily data, (c) is the weekly averages of the daily data, and (e) is the monthly average of the daily data. The right-hand column, including (b), (d), and (f), are plots of their respective average's loess fit with the 95% confidence bands in their respective colors.

variable is the amount of variability in walking speed, which has been shown to be associated with cognitive function (Dodge et al., 2012). Because variability, by definition, is based upon several observations of the same quantity, we need to use a level of granularity that would contain several observations in each interval. This again motivates us to reduce the daily walking speeds to seven-day or 30-day averages in order to use seven-day, or 30-day, variance of walking speed in our models. Third, as seen in Figure 2.11, there is no immediate short-term change in walking speeds; changes in walking speed occur instead over months or years, similar to the simulation scenarios with large sine periods. Thus, trends in walking speed can be detected from seven-day or 30-day average values.

Last, we need to consider the SNR, and more specifically the within-individual variance similar to the marker standard deviation in our simulations. Nonetheless, we can use the exploratory information provided in Figure 2.11 to roughly compare the relative SNRs resulting from each level of time granularity. We make this comparison through two aspects of each plot: (i) the non-linear loess fit and (ii) the confidence band around the loess fit. Before we further examine the corresponding SNR of each level of granularity, we must address any differences that exist in the loess fits for each time granularity in Figure 2.11. In theory, with complete data, the loess fit of each time granularity should have similar, if not the same, trajectories. In our example, we see that the loess fits for each follow similar patterns. Therefore, each time granularity is detecting roughly the same underlying signal, and we must inspect the confidence bands to determine the relative within-individual variance as part of the SNR.

The confidence bands will help us understand the relative variability of the data for each level of time granularity. In terms of total variability, all three levels of granularity produce data with relatively similar between-individual variation. The within-individual variation decreases as we move from daily granularity to weekly to 30-day averages because the difference in walking speed mean decreases between each individual's data points (Figure 2.11a, 2.11c, 2.11e). However, we should look at the individual variability with respect to the sample size, which is incorporated in the plots in right-hand column through confidence bands (Figure 2.11b, 2.11d, 2.11f). As we move from daily to seven-day to 30-day time granularities, the width of the confidence bands increase, which means we have greater sampling variability with respect to the sample size, which leads to decreased ability to identify significant trends in the data. Most importantly, we would like a summary window size that would allow us to detect each individual's pattern as distinct from the others. Thus, in the right-hand plots of Figure 2.11, we see that both the seven-day and 30-day averages produce data that have SNR values that are large enough to detect the underlying signal in the data, relative to the level of sampling variability. Thus, if we can make a good argument for seven-day average using the other factors we discussed, then we believe their use for analysis would be beneficial.

Summarizing over all our thoughts for the factors presented in Section 2.2 and simulation presented in Section 2.3, we believe a seven-day or 30-day time granularity have sufficient SNR for analysis. Although the daily data have the largest SNR, there are a few drawbacks to using daily granularity. First, visualization of the data does not facilitate an understanding of the data pattern because the noise of one subject's trajectory obscures the trajectory of other subjects. Second, daily granularity requires more computation time than data with lower granularity, and daily granularity does not allow for the use of variability in walking speed that can be produced from data with weekly granularity. Lastly, we believe data with seven-day or 30-day summaries will uncover signals similar to those produced with data with daily granularity. Thus, for this example, reducing the time granularity from daily measurements to weekly or 30-day summaries is appropriate for assessing longitudinal changes in walking speed. Further decisions will depend on the researcher's choices on computational demands.

2.6 Discussion

The decision of time granularity helps balance computational efficiency and the integrity of data. We discussed and examined factors through simulation studies to determine if coarse time granularities maintain the underlying signal of the data. The decision regarding summary window size is a consequence of increased information afforded by digital biomarkers, which allow researchers to assess trajectories more frequently than traditional tests such as pen-and-paper based NP tests or annually collected survey data. While NP tests can be administered at most every six months due to learning bias and participant burden, digital biomarker data are collected with much greater frequency.

Due to the high frequency of digital biomarker data collection, there are issues with data management and storage. As mentioned in Section 1, data collection starts with time stamps that can be taken on the order of seconds. This means that one day for one individual can potentially contain 86,400 data points (one per second). Typically, these data are condensed into longer time intervals that are dictated by data storage or available management. Setting aside these hardware issues, we chose to examine time granularities with respect to statistical analysis. If the reader has more concerns about data management issues, we suggest referencing other publications (Hsu et al., 2017; Li et al., 2017).

Our simulation study was limited to marker values simulated using a sine function. Because sine functions have a recurring pattern, if the summary window size aligns or misaligns with the period, this could result in an average RMSE that reflects this alignment instead of the window summary's ability to capture the true, underlying marker value. In future work, we will investigate scenarios in which the digital biomarker is simulated using regression functions with temporal variability, similar to those used in Ruppert and Carroll's (2000) simulations with spatial variability. Similar to the period of the sine function, we can control the spatial variability, and thus the rate of change over time, using a function parameter.

Our simulation study uses RMSE to measure the difference between our time granularity summaries and the true, underlying marker value. Because the summary window sizes are greater than 1, we decided to compare the window's average to the true marker value at the latest time point within the window. While this is a consistent decision across windows, the comparison may align with specific points of the sine period, which can produce RMSE values that are not comparable across window sizes. In the future, we plan to randomly sample one of the true, underlying marker values within the window.

Additionally, when comparing time granularities, there are issues involving missing data that we did not thoroughly examine in our simulation nor example. In certain cases, missing data can make the data susceptible to biased averages when looking at different time granularities. In the next chapter, we introduce imputation processes for digital biomarkers with missing data.

We also limited our discussion to factors that will strengthen analysis when we fit data using LMMs or GLMMs. Other methods have been used to analyze repeated measures involving dementia, including latent trajectory analysis, path analysis, Mixed-Effect Model Repeat Measure modelss (MMRMs), and functional data analysis (FDA). Readers can approach these forms of analysis using the factors laid out in Section 2, but each method has specific qualities that may require different factors or less emphasis on the ones we mentioned. For example, FDA uses smoothing methods to create continuous functions to represent data (Ramsay, 2006). This means that the decision of time granularity, and time as a discrete measurement, is just a processing step before we represent the data as a function and with continuous time. This may lead us to question whether time granularity should even be considered. However, variables of interest like variance of walking speed, may still be important in our analysis. Thus, we may want to look at a time granularity that supports summary data like variance.

In this paper we addressed factors to be considered when selecting time granularity. We specifically addressed daily data summarized into seven to 90-day averages, but this approach is applicable to data summaries of any time length. We also identified factors of the data that are important to consider in the decision of time granularity. In our exploratory procedure we looked at follow-up time, variables in analysis, pattern detection, and SNR to aid our decision. We showed that these factors of the data are linked to each other, and no single one decides if data reduction is appropriate. We then demonstrated the complex relationship of these factors with time granularity using average RMSE. Lastly, we walked through an example of longitudinal data where weekly time granularity was appropriate for our analysis. In the next chapter, we examine how the summarization process can be expanded to digital biomarkers with missing data using imputation.

CHAPTER 3

Imputing and Summarizing High-Frequency Predictors for Analyzing Lower-Frequency Binary Outcomes

3.1 Introduction

Digital biomarker data measure physical behaviors or physiology, such as steps per hour, heart rate throughout the day, or quality of sleep, and can be measured at high frequencies. Similar to other biomarkers, digital biomarkers can be used to potentially identify underlying biological processes that may be hard to directly measure. While current biomarkers are effective for dementia diagnosis, they often require clinical visits or invasive sampling (Anoop et al., 2010). For example, amyloid β-protein and Tau proteins are highly effective biomarkers for Alzheimer's disease that can be detected in cerebrospinal fluid (CSF) (Blennow et al., 2010). However, acquiring a CSF sample requires a lumbar puncture in patients. By using digital biomarkers instead, studies involving cognitive function can unobtrusively measure cognitive function, avoid recall bias in measurements, and detect asymptomatic disorders (Akl et al., 2015; Austin et al., 2017; Buchman et al., 2012; Dodge et al., 2012; Eby et al., 2012; Gorus et al., 2008; Hayes et al., 2014; Kaye et al., 2012, 2014; Kaye et al., 2011; Lyons et al., 2015; Silbert et al., 2016).

In studies where the clinical outcome cannot be measured frequently, digital biomarkers can help increase opportunities for assessment of the biological process. For example, in dementia studies, neuropsychological (NP) tests are used to assess cognitive function (Dodge et al., 2015). NP tests require a clinical visit and can only be administered at most every 6 months to prevent learning bias. If detected early, cognitive decline can be reversed through intervention and delay dementia (Aisen et al., 2017). However, time between NP test prevents frequent assessment of cognitive ability. Thus, one advantage of digital biomarkers is their high-frequency measurement using unobtrusive devices, like an activity tracker or in-home sensors. Thus, digital biomarkers can help increase the frequency for detection of cognitive decline, and increase opportunities for early dementia diagnosis (Dodge and Estrin, 2019).

In order to leverage information from digital biomarkers, we must first understand the relationship between them and cognitive function. As an initial step towards understanding the relationship, we can use longitudinal data analysis (LDA) methods to identify associations and the relative importance of specific digital biomarkers in the context of cognitive function. Thus, it is helpful to implement traditional statistical approaches, such as linear mixed effects models (LMMs), generalized estimating equations (GEEs), and generalized linear mixed effects models (GLMMs). By applying traditional methods as an initial step, researchers can quickly assess whether there is some association between a cognitive outcome and digital biomarkers before delving into more complicated methods of analysis. However, the difference in measurement frequency between the outcome and digital biomarkers complicates traditional LDA methods. Thus, it is important to summarize digital biomarker data into intervals that correspond that of the outcome.

For many clinicians, the summarization of digital biomarkers is a crucial first step to understanding the relationship between an outcome and the digital biomarker data. However, there is no general summarization process for analysts to follow. This can lead to varying summarization processes across studies, which makes comparisons difficult and can potentially bias results. For example, two studies examine dementia and Alzheimer's disease using accelerometer data over a 7-day follow-up time (Iaboni et al., 2022; Law et al., 2018). Each study summarizes the accelerometer data differently. One study used 1-minute summaries of physical measurements, while the other study used total number of minutes (summed over the 7-day follow-up) in sedentary, light, moderate, or vigorous activity. While both summarization processes may have specific benefits, neither study examined why they summarized the digital biomarkers using their chosen process.

The lack of guidance concerning the summarization process of digital biomarkers is exacerbated in the presence of missing data. Two main questions arise when summarizing digital biomarkers with missing data: (1) how do missing data points translate to missing summary values, and (2) if imputing data, should one impute before or after summarization? There are few publications that investigate the questions concerning summarization, imputation, and different frequencies of predictors and outcomes. There are papers that discuss missing data in the context of repeated measures, but they fail to include measurements that are taken at higher frequencies than the outcome (Tan et al., 2018). Thus, summarization of the predictors is not discussed. There are also papers that discuss missing data of baseline and time-varying predictors, but also fail to discuss higher-frequencies (Enders et al., 2016; Erler et al., 2016). Alternatively, there are studies that compare different summary values, like min, max, mean, and quantiles (Guo et al., 2020). However, these papers do not directly address whether imputation of the raw data or summary level data is more appropriate. Imputation is often performed on the raw data without consideration of potential consequences. Thus, there is no work examining the simultaneous process of summarizing and imputing high-frequency digital biomarker data for fitting lower-frequency outcomes.

Our paper aims to (1) define processes involving simultaneous imputation and summarization of digital biomarkers (high-frequency predictors) for longitudinal analysis of binary outcome, and (2) investigate the performance of imputation and summarization processes in different scenarios involving missing data. In Section 3.2 we describe two processes for imputation and summarization. In Section 3.3 we introduce the simulation study used to evaluate the imputation processes. In Section 3.4 we introduce a case study where we apply both imputation methods to longitudinal data from the Intelligent Systems for Assessing Aging Change (ISAAC) study. Section 3.5 concludes the paper with a discussion of the imputation processes and future work.

3.2 Methods

We now explain two approaches to impute a high-frequency biomarker. Both approaches are intermediate steps that are done prior to analyzing the association between a high-frequency biomarker and an outcome measured at a lower frequency than the biomarker.

For the approaches, any unit of time can be used, but we will assume time is measured in days. We have N individuals and each can be followed for a maximum of T days. For individual $i = 1, 2, \ldots N$, we let m_{it} denote the value of the biomarker at day $t = 1, 2, \ldots T$. We let $\mathbf{m}$ denote the $(N \times T)$ matrix of biomarker values for all N individuals, with row t of $\mathbf{m}$ equal to $\mathbf{m}_{\cdot t}$, the vector of biomarker values for all N individuals at time point t. Column i of $\mathbf{m}$ is equal to $\mathbf{m}_{i\cdot}$, the vector of biomarker values for individual i for all time points. During the T days of maximum follow-up for each individual, the low-frequency outcome is measured every h days. For individual $i = 1, 2, \ldots N$, we let Y_{ij} denote the value of the low-frequency outcome measured at the end of an h-day period $j = 1, 2, \ldots J$. We let $\mathbf{Y}$ denote the $(N \times J)$ matrix of binary outcomes for all N individuals, with row j of $\mathbf{Y}$ equal to $\mathbf{Y}_{\cdot j}$, the vector of outcomes for all N individuals at time point t. Column i of $\mathbf{Y}$ is equal to $\mathbf{Y}_{i\cdot}$, the vector of outcomes for individual i for all time points. For individual i, we let $r_{it} = 1$ if the value of m_{ik} is observed and $r_{it} = 0$ if the value of m_{ik} is missing.

Our goal is to examine the association of the high-frequency biomarker with the low-frequency outcome. Thus, we ultimately want to condense the vector of T values in $\mathbf{m}_{i\cdot}$ such that we have a summary of values in T that correspond to every one of the J values in $\mathbf{Y}_{i\cdot}$. We choose to do this by computing an average of the biomarker values that occur at the same time or precede each outcome measure in a window of $w \leq h$ days. For example, suppose the outcome is measured once every 30 days and we set $w = 7$ days. Thus, corresponding to the first outcome measured at day 30, we would compute an average of the biomarker values measured on days 24, 25, . . . , 30, and corresponding to the second outcome measured at day 60, we would compute a corresponding average of biomarker values measured at days, 54, 55, . . . , 60.

The appropriate ordering of imputation and summarization of the biomarker is uncertain, which motivates us to examine two different approaches to imputation:

1. **Impute then Summarize (IS)**: Impute biomarker at its original granularity (days); then summarize biomarker corresponding to the period of measurement for the outcome.

2. **Summarize then Impute (SI)**: First, summarize the biomarker corresponding to the period of measurement for the outcome. Summaries will be considered as missing if less than a threshold of observed values exist in that window. Then, impute values for the missing summaries.

3.2.1 Impute then Summarize (IS) Process

For the IS process, we first impute values at the original granularity of the biomarker. We then summarize the biomarker with lagged averages that correspond with the outcome. Imputation is performed using the Multivariate Imputation by Chained Equations (MICE) library in R (vanBuuren and Groothuis-Oudshoorn, 2011). We now describe the process for producing a single imputation for all of the missing biomarker values for each of the N individuals.

We impute missing marker values at time point t using multiple linear reg ression. We impute by first regressing a bootstrap sample of the observed values of the biomarker at t on their corresponding observed or imputed biomarker values at times $(t-1),(t-2),\ldots,(t-p)$, where p is a pre-defined number of prior days to use for imputation.

Let $R_t = \sum_{i=1}^{N} r_{it}$ denote the number of observed values of the biomarker at time t. Let $\dot{\mathbf{m}}_{\cdot t} = (\dot{m}_{1t}, \dot{m}_{2t}, \ldots, \dot{m}_{R_t,t})$ denote a bootstrap sample of size R_t from the observed values in $\mathbf{m}_{\cdot t}$. Note that $\dot{\mathbf{m}}_{\cdot t}$ at each time point is comprised of only observed values. However, when we use past time points to impute the current time point t, the past values can be observed or imputed because we sequentially impute. Thus, for individual $i = 1, 2, \ldots N$, we let $\widetilde{m}_{it}$ denote the value of the imputed or observed biomarker at day $t = 1, 2, \ldots T$. At time point t we need to identify the imputed or observed biomarker history at times $(t-1),(t-2),\ldots,(t-p)$ corresponding to the bootstrap sample at time point t, $\dot{\mathbf{m}}_{\cdot t}$. We use $\dot{m}_{k,(t-s)}$, $k = 1, 2, \ldots R_t$ and $s = 1, 2, ..., p$ to denote the biomarker value history, $\widetilde{m}_{i,(t-s)}$, that corresponds to the current time point's bootstrap value $\dot{m}_{kt}$.

We then fit the linear regression model

$$\dot{m}_{kt} = \beta_{t0} + \sum_{\ell=1}^{p} \beta_{t\ell} \dot{m}_{k,(t-\ell)} + \epsilon_{kt} \tag{3.1}$$

in which $\epsilon_{1t}, \epsilon_{2t}, \ldots \epsilon_{R_t,t}$ are independent and normally distributed with mean 0 and variance σ_t^2.

For each individual missing a biomarker value ($r_{it} = 0$), we impute their biomarker value, m_{it}^*, using a single draw from a normal distribution with variance $\widehat{\sigma}_t^2$ and mean $\mu_{it} = \widehat{\beta}_{t0} + \sum_{\ell=1}^{p} \widehat{\beta}_{t\ell} m_{i,(t-\ell)}$, in which $\widehat{\beta}_{tq}$ is the least-squares estimate of $\beta_{tq}, q = 0, 1, \ldots, p$ and $\widehat{\sigma}_t^2$ is the mean-squared error from the fit of the model in Equation 3.1. For individual $i = 1, 2, \ldots N$, we let $\widetilde{m}_{it}$ denote the value of the imputed or observed biomarker at day $t = 1, 2, \ldots T$. Thus, $\widetilde{m}_{it} = m_{it}$

when the marker is observed and $\widetilde{m}_{it} = m^*_{it}$ when the marker is imputed.

This process is recursive, meaning that we start at $t = 1$ and move forward in time. Thus, our model implicitly assumes that there are no missing biomarker values at $t = 1$. We then impute missing biomarker values at $t = 2$ using only the biomarker values from $t = 1$, i.e. our regression parameters are only β_{01} and β_{11}. We then use the entire vector of observed and imputed biomarkers for $t = 2$ in the imputation of missing biomarkers at $t = 3$. Once all missing values are imputed, we have a complete matrix of observed and imputed biomarker values.

Next, we summarize the imputed and observed biomarker values. Specifically, for $j = 1, 2, 3, ..., J$, we compute $\overline{\widetilde{m}}_{ij} = \sum_{t=1+hj-w}^{hj} \widetilde{m}_{it}/w$, with $\overline{\widetilde{\mathbf{m}}}$ denoting the $(N \times J)$ matrix of average biomarker values.

3.2.2 Summarize then Impute (SI) Process

For the SI process, we have three steps: (i) summarize the marker data every w days, (ii) impute summary values that are missing due to insufficient observed data, and (iii) keep the summary values corresponding to the j time points of the outcome and discard all other summary values. Imputation is performed at the summary level once again using the MICE library in R. Below we describe the process for producing a single vector of imputed summary values for all N individuals.

In the IS process, after imputing biomarker values, we only computed summary values for each window most proximate to the each outcome. In the SI process, we summarize marker values at consecutive and non-overlapping intervals of w days, regardless of proximity to outcome. Additionally, in the IS process, biomarker summaries, $\overline{\widetilde{m}}_{ij}$, were computed using observed and imputed values. However, for the SI process, some summary values will be deemed missing if the window contains an insufficient number of observed biomarker values. Specifically, we define p_w as the minimum proportion of observed markers required in a window. We let s_{ig} indicate whether window g meets the threshold of observed biomarker values for individual i. Thus, $s_{ig} = 1$ if $\sum_{t=1+w(g-1)}^{wg} r_{it} \geq p_w w$ and $s_{ig} = 0$ otherwise. For example, suppose we want to average over $w = 30$ days and we set $p_w = 0.6$. For each window g, we compute a summary biomarker value if $\sum_{t=1+w(g-1)}^{wg} r_{it} \geq 18$. In other words, if 18 or more daily marker values are observed in the 30 day window are observed, we compute a summary value for that window. Otherwise, if more than

12 biomarker values are missing, then we need to impute a summary value for that window.

For individual $i = 1, 2, \ldots N$, we let n_{ig} denote the value of the biomarker summary for window $g = 1, 2, \ldots T/w$. We let $\mathbf{n}$ denote the $(N \times G)$ matrix of biomarker values for all N individuals, with row g of $\mathbf{n}$ equal to $\mathbf{n}_{\cdot g}$, the vector of biomarker summary values for all individuals at window g, and column i of $\mathbf{n}$ equal to $\mathbf{n}_{i\cdot}$, the vector of biomarker summary values for individual i for all windows. For $s_{ig} = 1$, we compute the window summary:

$$n_{ig} = \frac{\sum_{t=1+w(g-1)}^{wg} m_{it} r_{it}}{\sum_{t=1+w(g-1)}^{wg} r_{it}} \tag{3.2}$$

If $s_{ig} = 0$ then we need to impute the summary value for the window.

Similar to the IS process, we impute missing marker summary data for window g using multiple linear regression. We impute by regressing a bootstrap sample of the observed summaries of the biomarker for window g on their corresponding computed biomarker summaries for windows $(g-1), (g-2), \ldots, (g-p)$, where p is a pre-defined number of prior summary values to use for imputation.

Let $S_g = \sum_{i=1}^{N} s_{ig}$ denote the number of number of individuals with biomarker summary values for window g. Let $\dot{\mathbf{n}}_{\cdot g} = (\dot{n}_{1g}, \dot{n}_{2g}, \ldots, \dot{n}_{S_g,g})$ denote a bootstrap sample of size S_g from the computed summaries in $\mathbf{n}_{\cdot g}$. Similar to the IS process, note that $\dot{\mathbf{n}}_{\cdot g}$ for each window g is comprised of only computed summary values. However, when we use summary values of past windows to impute the current summary value for window g, the past summary values can be computed or imputed because we sequentially impute. Thus, for individual $i = 1, 2, \ldots N$, we let $\widetilde{n}_{ig}$ denote the value of the imputed or computed biomarker summary value for window $g = 1, 2, \ldots T/w$. For window g we need to identify the imputed or computed biomarker summary history for windows $(g-1), (g-2), \ldots, (g-p)$ corresponding to the bootstrap sample for window g, $\dot{\mathbf{n}}_{\cdot g}$. We use $\dot{n}_{l,(g-s)}$, $l = 1, 2, \ldots S_g$ and $s = 1, 2, ..., p$ to denote the biomarker summary value history, $\widetilde{n}_{i,(g-s)}$, that corresponds to the current time point's bootstrap value $\dot{n}_{lg}$.

We then fit the linear regression model

$$\dot{n}_{lg} = \alpha_{g0} + \sum_{\ell=1}^{p} \alpha_{g\ell} \dot{n}_{l,(g-\ell)} + \gamma_{lg} \tag{3.3}$$

in which $\gamma_{1g}, \gamma_{2g}, \dots \gamma_{S_g,g}$ are independent and normally distributed with mean 0 and variance τ_g^2.

For each individual for whom $s_{ig} = 0$, we impute their biomarker summary value, n_{ig}^*, using a single draw from a normal distribution with variance $\widehat{\tau}_g^2$ and mean $\nu_{ig} = \widehat{\alpha}_{g0} + \sum_{\ell=1}^{p} \widehat{\alpha}_{g\ell} n_{i,(g-\ell)}$, in which $\widehat{\alpha}_{gq}$ is the least-squares estimate of $\alpha_{gq}, q = 0, 1, \dots, p$ and $\widehat{\tau}_g^2$ is the mean-squared error from the fit of the model in Equation 3.3. For individual $i = 1, 2, \dots N$, we let $\widetilde{n}_{ig}$ denote the value of the imputed or computed biomarker summary at window $g = 1, 2, \dots T/w$. Thus, $\widetilde{n}_{ig} = n_{ig}$ when the marker summary is computed and $\widetilde{n}_{ig} = n_{ig}^*$ when the marker summary is imputed.

Like IS, this process is recursive, meaning that we start at $g = 1$ and move forward in time. Again, we implicitly assume that all biomarker summaries are computed at $g = 1$. We then impute missing biomarker summary values at $g = 2$ using only the biomarker summary values from $g = 1$, i.e. our regression parameters are only α_{01} and α_{11}. Once all missing summaries are imputed, we have a complete matrix of computed and imputed biomarker summaries.

At this point, we have an $N \times G$ matrix of computed and imputed summary biomarker values. We now need to reduce the number of columns of $\widetilde{\mathbf{n}}$ from G to J to match the dimensions of Y. We do this by comparing h, the frequency of the outcome, to w, the frequency of each summary. We keep marker summary values for window g if g is a multiple of h/w. For example, suppose the outcome is measured once every $h = 60$ days and we want to average over $w = 30$ days. In this case, $h/w = 2$, thus we keep marker summaries for windows $g = 2, 4, 6, ..., J$. We let $\overline{\mathbf{n}}$ denote this reduced set of imputed and computed biomarker summaries for all N individuals over all J time points.

3.2.3 Visual Example

In Figure 3.1 and 3.2, we present an example of IS and SI processes for a biomarker measured on the first individual. For both Figure 3.1 and 3.2, we start with the individual's data with missing values. The individual's biomarker, m_{1t}, is measured at $t = 1, 2, ..., 480$. In this example, the individual's outcome, Y_{1j}, with $j = 1, 2, 3, 4$, is measured every $h = 120$ days over the $T = 480$ day follow-up time. Thus, the Forty percent of the biomarker values are missing, specified by the grey data points. We will impute and summarize the biomarker so that we have $J = 4$ biomarker summary values that correspond to the $J = 4$ outcome measurements for the individual.

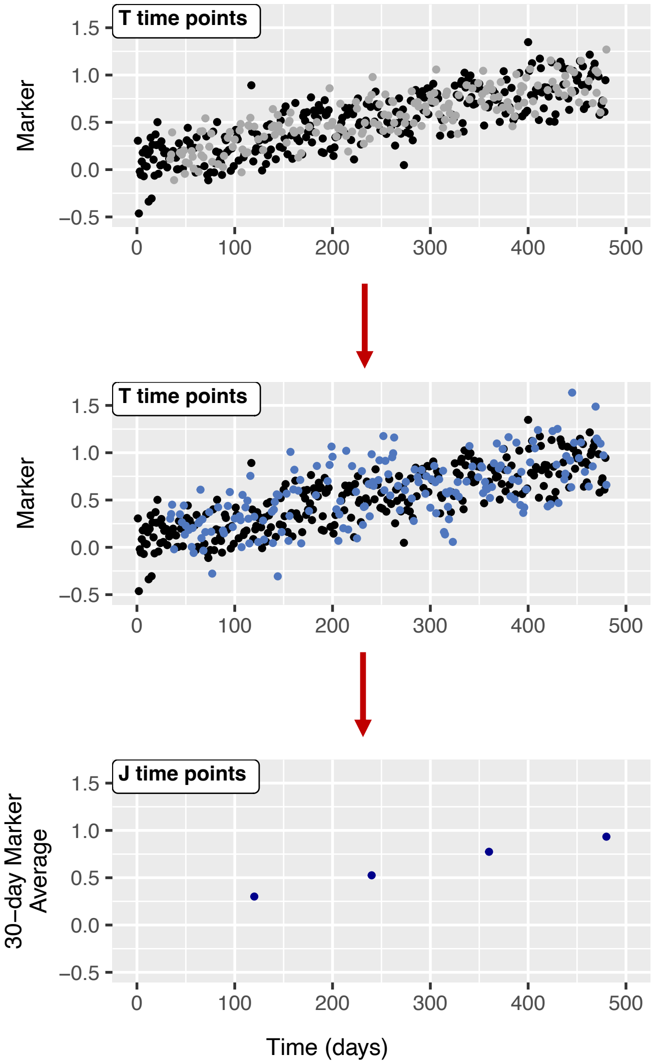


Figure 3.1: Process for IS. Note the number of time points in the top left annotation of each plot. For an outcome with $J = 4$ measurements, the $T = 480$ measurements of the marker will be condensed into 4 summary values. Black data points represent the observed marker data. Grey data points represent the missing marker data. Blue data points represent the imputed marker data.

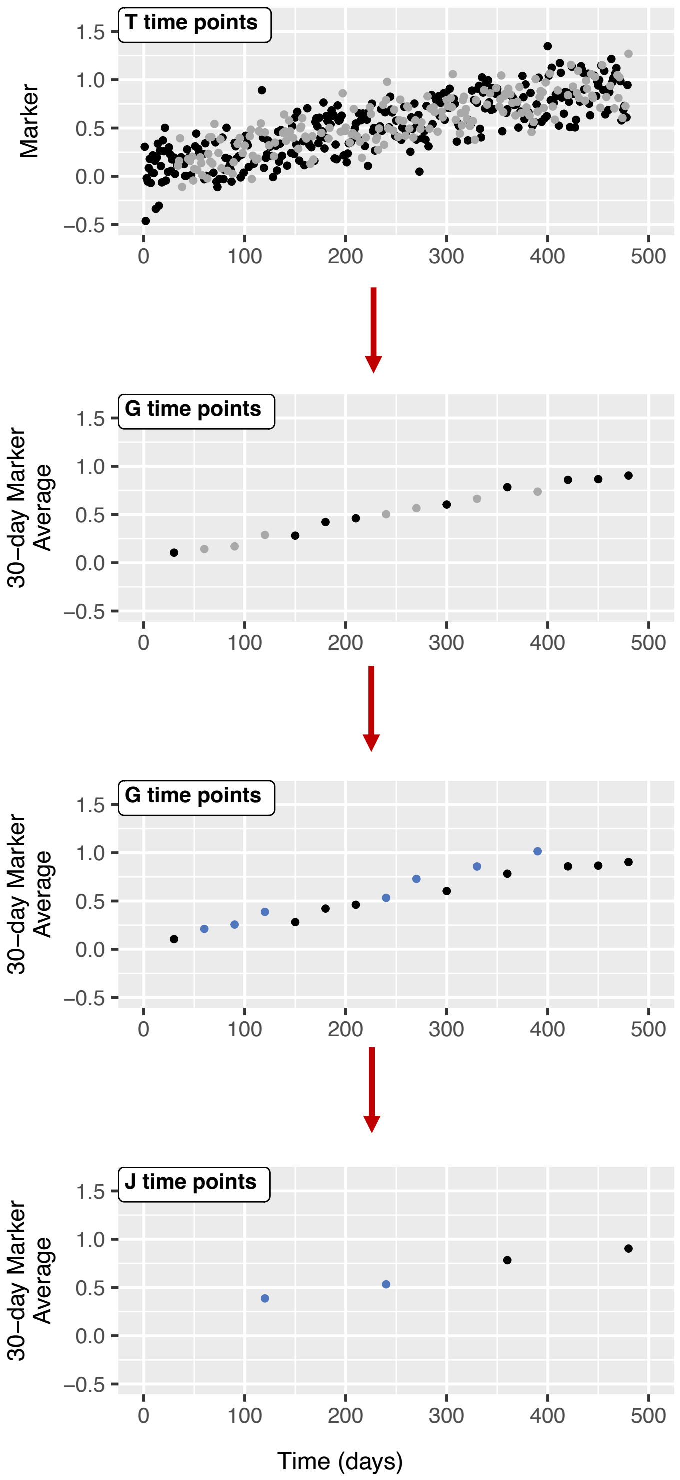


Figure 3.2: Process for SI. Note the number of time points in the top left annotation of each plot. For an outcome with $J = 4$ measurements, the $T = 480$ measurements of the marker will be condensed into 4 summary values. Intermediate steps include $G = 16$ summary measurements at which the markers are imputed. Black data points represent the observed marker data. Grey data points represent the missing marker data. Blue data points represent the imputed marker data.

For the IS process, in Figure 3.1, the first step, signified by the first downward red arrow, is to impute the missing daily biomarker values using the methods described in Section 3.2.1. At the end of this step, we have $T = 480$ imputed or observed daily biomarker values, which constitutes $\widetilde{m}_{1t}$. Following the next downward red arrow, we summarize the markers every $h = 120$ days using a $w = 30$ day window. Thus, we have 30-day marker averages, $\overline{\widetilde{m}}_{1j}$, with $j = 1, 2, 3, 4$, that correspond to the outcome measurements, Y_{1j}.

For the SI process, in Figure 3.2, the first step, shown by the first red arrow, is to summarize the marker data into $w = 30$ day consecutive averages over $G = 16$ windows. Averages are considered missing if they do not meet the $p_w = 0.6$ proportion of observed data. Following the next downward red arrow, we then impute the missing summary biomarker values using the methods described in Section 3.2.2. With the imputed and observed biomarker summary values, $\widetilde{n}_{1g}$, with $g = 1, 2, ...16$, we then keep biomarker summary values every $h = 120$ days. Thus, we have 30-day marker averages, $\overline{n}_{1j}$, with $j = 1, 2, 3, 4$, that correspond to the outcome measurements, Y_{1j}.

The IS and SI processes produce different 30-day summary values. In the final plot in Figure 3.1 and 3.2, we can visually see the differences in the 30-day marker averages. For example, the first marker average is larger for the IS process than the SI process. This is because the first 30-day average corresponding to the IS process consists of 30 daily marker values that are imputed or observed. This is not equal to the 30-day marker average from the SI process, in which the first average is imputed at that summary-level.

3.3 Simulation Study

We present a simulation study to compare the IS and SI processes for four clinically meaningful data scenarios. The four scenario settings differ based on period of the sine function ($2\pi d$) generating the marker and consecutive number of missing days (q). Figure 3.3 shows an example of one individual's biomarker data for each scenario with 40% missing data. Scenario 1 and 3 (Figure 3.3a, 3.3c) have the same period (960π), and scenario 2 and 4 (Figure 3.3b, 3.3d) have the same period (20π). Scenario 1 and 2 (Figure 3.3a, 3.3b) have the same missing section size ($q = 1$), and scenario 3 and 4 (Figure 3.3c, 3.3d) have the same missing section size ($q = 60$). Note that

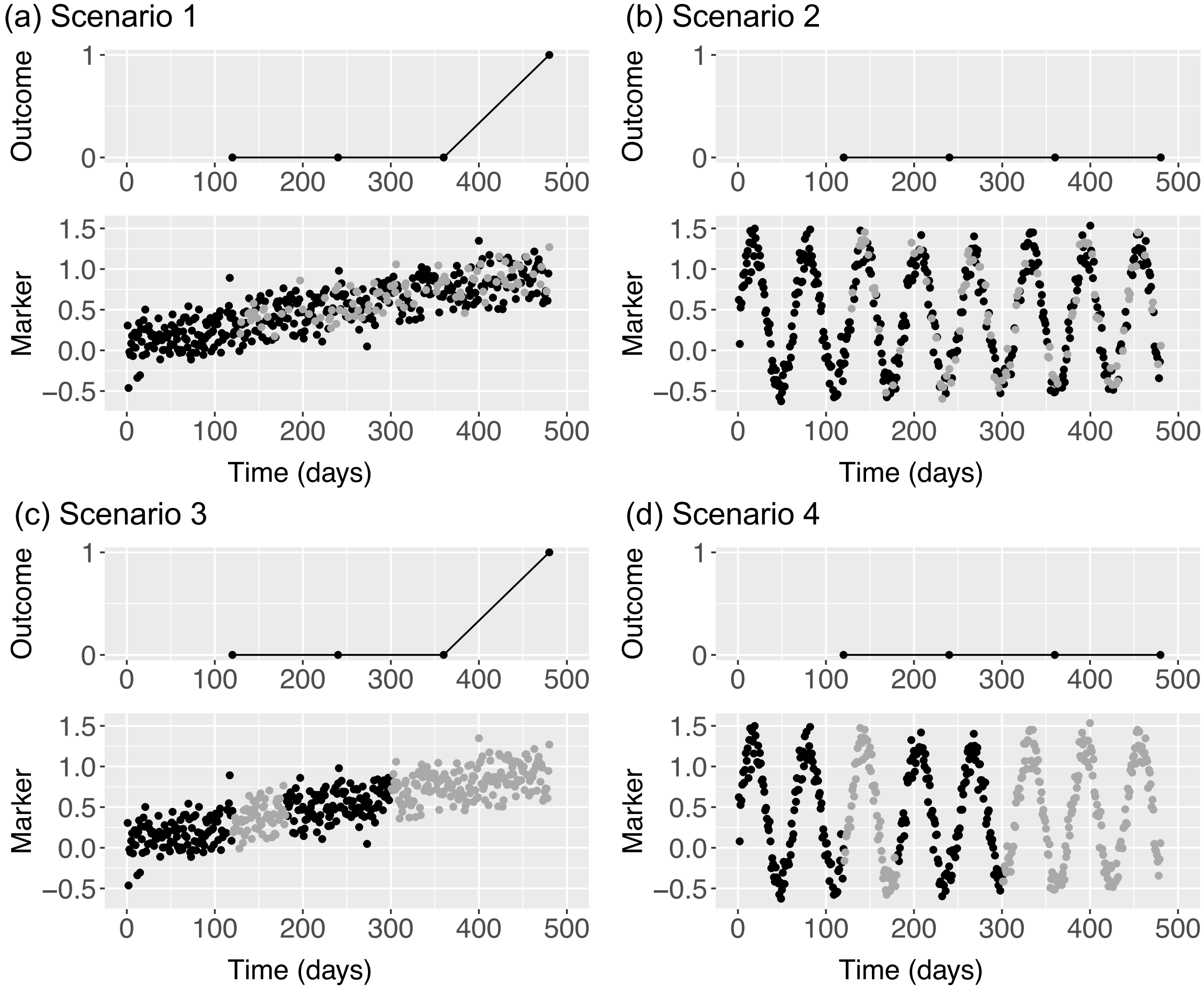


Figure 3.3: Four generated scenarios for one individual with a follow-up time of 480 days. Each scenario contains an binary outcome and marker data. Observed marker data are represented with black data points, and missing marker data are represented with grey data points. There is a 40% probability for missing marker data in each scenario.

in Figure 3.3, Scenario 1 and 3 have the same outcome, which differ from Scenario 2 and 4. The outcome is simulated based on the digital biomarker data with 0% missing digital biomarker values. Thus, scenarios with the same simulated biomarker data, but different missing indicators, will have the same outcome.

For each scenario, 1,000 simulations are performed on a range of percent missing data (from 0 to 40%), totalling 6,000 simulations per scenario. Below we outline the general framework used to simulate and estimate the association of the biomarker average with the outcome for each scenario at a specified missing percent. In following sections, we will further explain each step.

Step 1: Simulate complete biomarker data

Step 2: Generate binary outcome using generalized linear mixed effects model (GLMM) with complete biomarker data (from Step 1) as time-varying predictors with prescribed fixed and random effects

Step 3: Determine missing marker data using missing completely at random (MCAR) at specified missing data percent

Step 4: Impute missing data using above methods:

- Impute then Summarize (IS)
- Summarize then Impute (SI)

Step 5: Fit GLMM for below datasets:

- Complete simulated data (no missing marker data)
- Available simulated data (missing data, no values imputed)
- Imputed data using IS process
- Imputed data using SI process

Step 6: Compare fixed effects estimates to prescribed fixed effects in step 2

The following sections will explain how complete data are simulated (steps 1 and 2), missing indicators are sampled (step 3), missing data are handled (step 4), datasets are analyzed (step 5), and the results of our simulation study (step 6).

3.3.1 Simulating Complete Digital Biomarker and Outcome Data

For each complete simulation, we generate data for high-frequency markers (m_{it}), and lower-frequency binary outcomes (Y_{ij}). We continue to use notation from the methods. We simulate $N = 40$ individuals with $T = 480$ follow-up time.

We generate m_{it} using the following model:

$$m_{it} = a_i + q_i t + 300 \sin \frac{t}{d} + \omega_{it} \tag{3.4}$$

in which $\omega_{1t}, \omega_{2t}, \dots \omega_{iT}$ are independent and normally distributed with mean 0 and standard deviation of 10; $a_1, a_2, \dots a_N$ are independent and normally distributed with mean 0 and standard deviation of 300; and $q_1, q_2, \dots q_N$ are independent and normally distributed with mean 0 and standard deviation of 0.01. For simulation scenarios 1 and 3, $d = 480$, and for scenarios 2 and 4, $d = 10$.

We also generate binary outcomes that are measured at a lower frequency than the biomarker. Recall, the outcome is measured every h days, while the digital biomarker is measured every day. We set $h = 120$, so that J, which is the total number of outcome measurements, is $T/h = 480/120 = 4$, and the window size is $w = 30$. We simulate the binary outcome, Y_{ij}, from a Bernoulli distribution with probability π_{ij} where

$$logit(\pi_{ij}) = \eta_0 + \eta_1 \overline{m}_{i1} + \eta_2 \Delta \overline{m}_{ij} + b_i \tag{3.5}$$

in which $\overline{m}_{ij} = \sum_{t=1-w+jh}^{jh} m_{it}/w$, $\Delta\overline{m}_{ij} = \overline{m}_{ij} - \overline{m}_{i1}$, and the random intercepts, $b_1, b_2, \dots b_N$, are independent and normally distributed with mean 0 and standard deviation of 0.7. We set $\eta_0 = 0.05$, $\eta_1 = 0.45$, and $\eta_2 = 0.9$. Our goal is to estimate the three coefficients in Equation 3.5. For individuals $i = 1, 2, ...40$ and corresponding outcome measurements $j = 1, 2, 3, 4$, the complete data will consist of $\overline{m}_{ij}$ and Y_{ij}.

3.3.2 Simulating Missing Biomarker Data

We create missingness in the biomarker data as follows. After the initial w days, we generate an indicator of missingness for each biomarker value. We exclude the first w days because our imputations require past marker data to inform the current imputed value, and if the first marker value (daily or summary level) were missing, we could not perform imputation.

The missing indicators are generated independently from a Bernoulli distribution with a probability $1-p_{miss}$, so that we have data missing completely at random (MCAR). We also will examine varying levels of missingness by having $p_{miss} \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5\}$. Recall, the scenarios differ based on the consecutive number of days missing. We specified two cases with one-day missing ($q = 1$) and 60 consecutive days missing ($q = 60$). Thus, for $q = 1$, missing indicators are generated independently for marker values each day. For $q = 60$, missing indicators are generated independently for sections of 60 consecutive marker values. Regardless of q, for individual $i = 1, 2, \ldots N$, we let r_{it} denote the missing indicator at day $t = 1, 2, \ldots T$. We let $r_{it} = 1$ if the value of m_{ik} is observed and $r_{it} = 0$ if the value of m_{ik} is missing. However, the processes for assigning missingness differs for q values. In the following we describe the process for generating missing indicators for each.

For $q = 1$, which is the case for scenario 1 and 2, we generate missing indicators for each day independently. Outside of the initial w time points, there is a probability, p_{miss}, of missingness. Thus, each r_{it}, given $t > w$, was generated independently from a Bernoulli distribution with a probability $1 - p_{miss}$.

For $q = 60$, which is the case for scenario 3 and 4, missing data will span q consecutive days. Thus we divided the total follow-up time into sections of size $q = 60$ with sections $z = 1, 2, ...Z$. For each individual $i = 1, 2, ..., N$ and each section $z = 1, 2, ..., Z$, we denote a missing indicator, v_{iz}. Similar to the daily missing indicator, each v_{iz}, given $60z > w$, was generated independently from a Bernoulli distribution with a probability $1 - p_{miss}$. We use the missing section indicator to construct the missing daily indicator. For $q = 60$, we let $r_{it} = v_{iz}$ for $1 + 60(z - 1) < t \leq 60z$.

3.3.3 Summarizing Observed Biomarker Data

We now describe the process for summarizing the observed marker values to create the available dataset. We summarize the observed daily marker values according to the process outlined in Section 3.2.2. We use a threshold, $p_w = 0.6$, to determine whether the summary biomarker value at j is computed. Note that for available data, we only need to compute summary values that correspond to the outcomes. At each outcome j, if the proportion of observed biomarker values within the window is greater than $p_w = 0.6$, then we compute the the window summary value, $n'_{ij} = \sum_{t=1+hj-w}^{hj} m_{it}r_{it} / \sum_{t=1+hj-w}^{hj} r_{it}$. Otherwise, the window summary is considered missing.

3.3.4 Imputing Missing Biomarker Data

We now describe the the implementations of the imputation and summarization processes to create the IS and SI-imputed datasets. For the IS process, we impute missing marker values at time point t using the methods described in Section 3.2.1. We set $p = 5$, and thus use the previous observed or imputed values at $(t-1), (t-2), ..., (t-5)$ to impute the missing digital biomarker value at time point t. We summarize the imputed and observed biomarker values. For $j = 1, 2, 3, 4$, we compute individual averages over a $w = 30$ window.

For the SI process, we summarize the marker over consecutive $w = 30$ day windows as described in Section 3.2.2. We set $p_w = 0.6$ to determine whether the summary biomarker value for a given window is computed. We set $p = 5$, and thus use the previous observed or imputed summary values at window $(g-1), (g-2), ..., (g-5)$ to impute the missing digital biomarker summary value at window g. Then we keep the computed and imputed biomarker summary values at windows $g = 4, 8, 12, 16$ that correspond to the $h = 120$ day frequency of the outcome.

3.3.5 Statistical Analysis

To evaluate the performance of each dataset, we fit the binary outcome using GLMM with the same covariates used to generate the outcome in Equation 3.5 in Section 3.3.1. For example, for the complete data, we fit Equation 3.5. For available and imputed datasets, we replace values of $\overline{m}_{ij}$ with their respective biomarker summary values.

For each simulation, we focus on the estimates for the fixed effects: $\hat{\eta}_0$, $\hat{\eta}_1$, and $\hat{\eta}_2$. We are most interested in the fixed effect for the changes from baseline, which is estimated by $\hat{\eta}_2$.

Recall, we prescribed specific, "true" values for each coefficient in Section 3.3.1. Coefficient estimates for the fixed effects from analyses using each of the four datasets (complete, available, IS-imputed, and SI-imputed data) were compared to one another and the prescribed coefficient parameter value ($\eta_2 = 0.9$). Within each simulation, for IS and SI-imputed data, we performed five multiple imputations and used Rubin's rules (Rubin, 1987) to pool results of all five imputations.

For each of the four datasets, we calculated three summary values concerning $\hat{\eta_2}$, the estimated fixed effect of changes from baseline average biomarker values, across 1,000 simulations: (1) the relative bias of $\hat{\eta_2}$, (2) the coverage rate of the 95% confidence interval coverage for $\hat{\eta_2}$, and (3) the sampling variance of $\hat{\eta_2}$ across simulations.

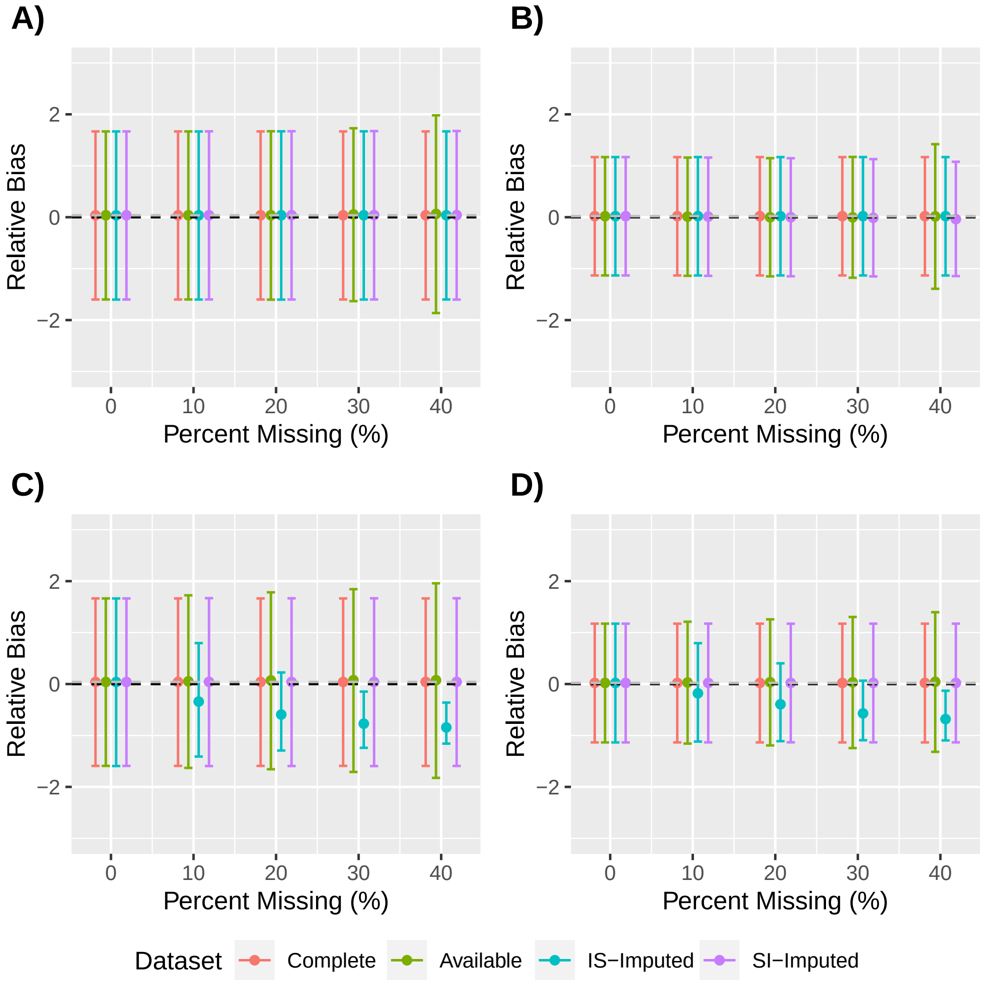
3.3.6 Results

Recall the four simulation scenarios presented in Figure 3.3: (1) one-day missing section size, large period, (2) one-day missing section size, small period, (3) 60-day missing section size, large period, and (4) 60-day missing section size, small period. In the following, we present the three above summary values across simulations concerning $\hat{\eta_2}$.

In Figure 3.4, we present the relative bias using the mean and 95% confidence interval across simulations for each scenario. The relative bias in each scenario is presented in separate plots with percent missing data on the x-axis. Each color in the figure represents a different dataset: (1) complete simulated data (no missing marker data), (2) available simulated data (missing data, no values imputed), (3) imputed data using IS process, and (4) imputed data using SI process. There are two horizontal, dotted lines on each plot. The black one marks the relative bias at zero and the grey one marks the mean relative bias of the complete dataset. As an example, in Figure 3.4A, in which the digital biomarker is simulated with a large period and daily missing data, we see that the mean relative bias of the complete dataset is close to 0 and remains the same regardless of the percent missing data. Recall, the complete dataset does not have missing data and serves as a reference for the other datasets. Thus, the complete dataset remains the same across the percent of missing data. The mean relative bias for the other datasets (available, IS-imputed, and SI-imputed) remain close to that of the complete dataset. However, the confidence interval of the available dataset, in which the marker has missing data without imputation, increases as the percent missing data increases. In Figure 3.4B-D, we see this pattern for the available dataset for all other scenarios.

In Figure 3.4C-D, the mean relative bias of the IS-imputed dataset decreases starting at 10% missing data. With each increase in 10% missing data, up to 40%, the relative bias approaches -1. The 95% confidence intervals also decrease in range. Thus, as the percent missing increases, the

Figure 3.4: Relative bias for simulation scenarios: (A) daily-level missing data, large period, (B) daily-level missing data, small period, (C) 60-day section missing data, large period, and (D) 60-day section missing data, small period. The dashed grey line marks the relative bias for complete data. The dashed black line marks zero relative bias. Pink represents complete data, green available data, blue daily-level imputations, and purple summary-level imputations. Points are staggered at each percent missing so that comparisons can be made at each.



estimate of η_2 for the IS-imputed data becomes less variable and more biased.

For all scenarios, it is important to note a phenomena where the increased variance of a predictor, markers in this case, can attenuate the coefficient estimate to zero. By imputing the marker, we can increase the variance of the markers. In Figure 3.4, we can see that each scenario where bias increases gradually approaches a relative bias close to -1. This number is important because the *true* coefficient parameter value that was prescribed in the data generation is 0.9. Since relative bias is defined by the difference between the estimated coefficient and the true coefficient, as the coefficient estimate approaches 0, the relative bias approaches -1.

In Figure 3.5, we present the coverage rate across simulations of the 95% confidence interval for $\hat{\eta_2}$. The coverage rate in each scenario is presented in separate plots with percent missing data on the x-axis. Each color in the figure represents a different dataset: (1) complete simulated data (no missing marker data), (2) available simulated data (missing data, no values imputed), (3) imputed data using IS process, and (4) imputed data using SI process. The dotted black line marks the 95% coverage rate. For an example, in Figure 3.5A, in which the digital biomarker is simulated with a large period and daily missing data, the coverage rate for each dataset and each percent missing is between 90% and 95%.

In Figure 3.5C-D, the coverage rate of the IS-imputed dataset increases to 99% at 10% missing data, then decreases from 10% to 40% missing data. This reflects the pattern we saw in 3.4C-D. At 20-40% missing data, $\hat{\eta_2}$ for the IS-imputed data becomes less variable and more biased, thus, the coverage of $\hat{\eta_2}$ decreases. Lastly, in Figure 3.5, the coverage rate of the available dataset generally stays around 95%, if not increasing as percent missing data increases. This means $\hat{\eta_2}$ for the available dataset remains fairly unbiased, as we saw in Figure 3.4, but the standard deviation of $\hat{\eta_2}$ increases because we have less data. Thus, an increase in coverage rate for the available dataset reflects a less precise estimate of η_2.

In Figure 3.6, we present the sampling standard deviation of $\hat{\eta_2}$ across simulations. The sampling standard deviation in each scenario is presented in separate plots with percent missing data on the x-axis. Again, each color in the figure represents a different dataset: (1) complete simulated data (no missing marker data), (2) available simulated data (missing data, no values imputed), (3) imputed data using IS process, and (4) imputed data using SI process.

For an example, in Figure 3.6A, in which the digital biomarker is simulated with a large period

Figure 3.5: Coverage rate for simulation scenarios: (A) daily-level missing data, large period, (B) daily-level missing data, small period, (C) 60-day section missing data, large period, and (D) 60-day section missing data, small period. Dashed black line marks 95% coverage rate. Pink represents complete data, green available data, blue daily-level imputations, and purple summary-level imputations. Points are staggered at each percent missing so that comparisons can be made at each.

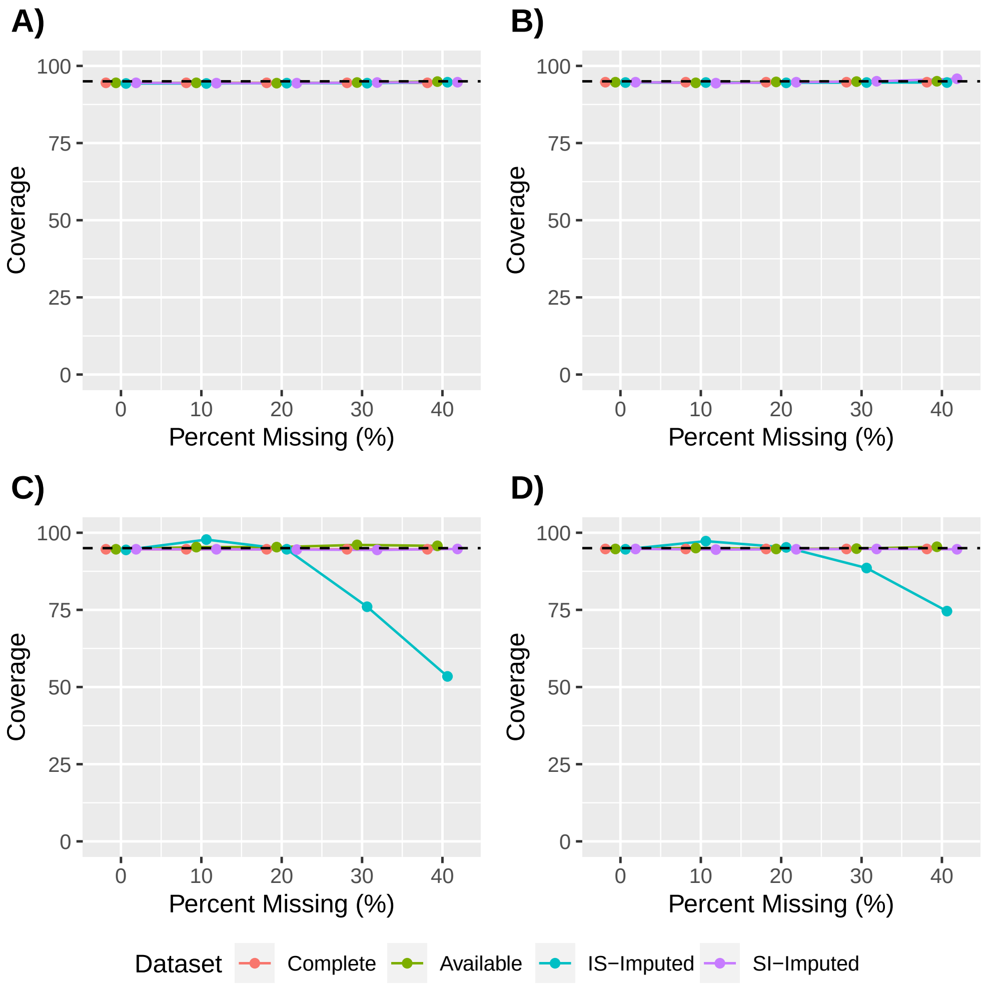
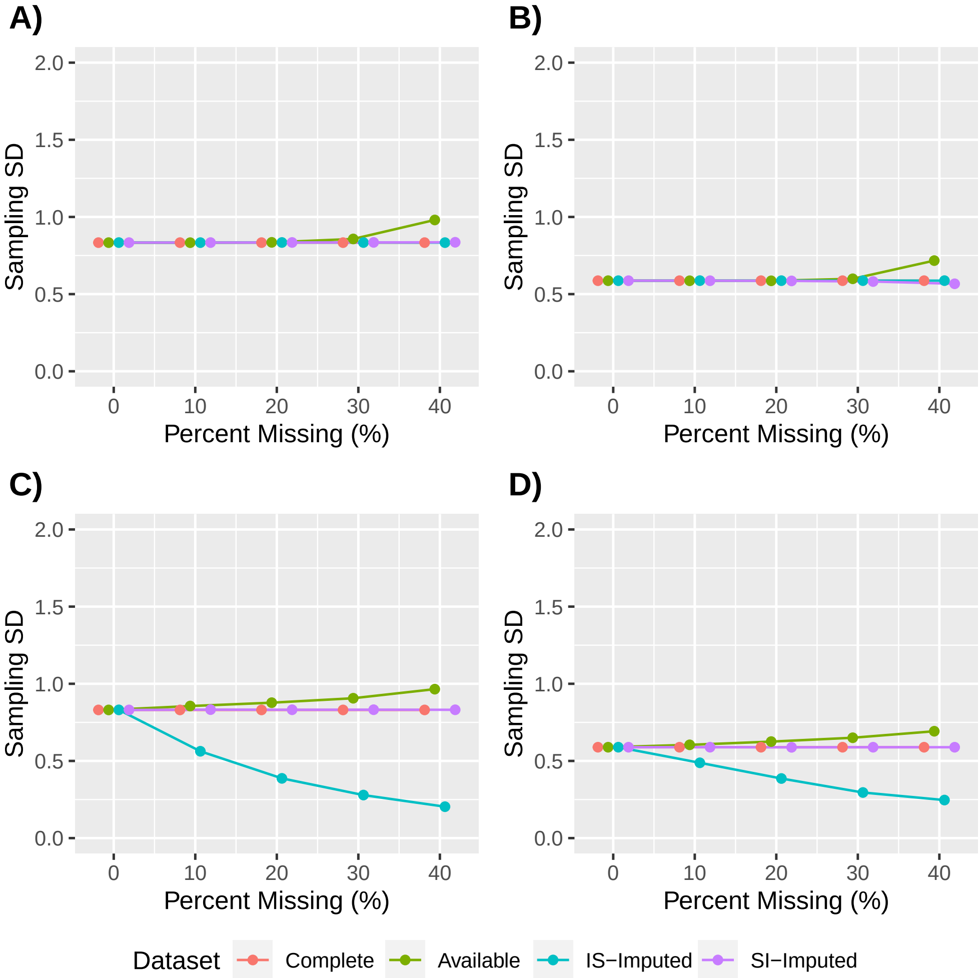


Figure 3.6: Sample standard deviation (SD) for the coefficient estimate for simulation scenarios: (A) daily-level missing data, large period, (B) daily-level missing data, small period, (C) 60-day section missing data, large period, and (D) 60-day section missing data, small period. Pink represents complete data, green available data, blue daily-level imputations, and purple summary-level imputations. Points are staggered at each percent missing so that comparisons can be made at each.



and daily missing data, the mean sampling standard deviation of the complete dataset remains constant at 0.87 regardless of the percent missing data. Recall, the complete dataset does not have missing data and serves as a reference for the other datasets. The sampling standard deviation for the IS-imputed and SI-imputed datasets remain close to that of the complete dataset. However, the sampling standard deviation of the available dataset, in which the marker has missing data without imputation, increases as the percent missing data increases because there is less data. In Figure 3.6B-D, we see this pattern for the available dataset for all other scenarios.

In Figure 3.6C-D, the sampling standard deviation of the IS-imputed dataset decreases towards 0 starting at 10% missing data. This reflects the pattern in Figure 3.4, as the percent missing increases, the estimate of η_2 for the IS-imputed data becomes less variable and more biased.

3.4 Case Study

To demonstrate the empirical differences between the two imputation processes, we examine data from the ISAAC study, which is a longitudinal cohort study collected through the Oregon Center for Aging and Technology (ORCATECH) at the Oregon Health and Science University. The study consists of 158 individuals with measured digital biomarkers using a series of sensors and cognitive function with NP tests administered by clinicians. More details of this study have been previously published (Dodge et al., 2015; Hayes et al., 2014).

We identified two variables to demonstrate the imputation processes: (1) a digital biomarker measurement, specifically daily mean walking speed and (2) mild cognitive impairment (MCI) indicators that are measured biannually or annually. In the study, walking speed was generated using a series of four motion sensors on the ceiling of a narrow hall that were triggered when a subject passed directly underneath (Gorus et al., 2008). Walking speed measurements were collected on each participant at multiple time points each day (whenever subjects passed under the series of sensors), and these multiple measures were processed into a daily mean walking speed. MCI is a binary indicator of mild cognitive impairment, meaning an individual received a clinical dementia rating (CDR) of 0.5 or higher indicating questionable to certain dementia. CDR and MCI are measured through a clinically administered NP test. We aim to determine the association between mean walking speed and MCI using LDA methods, specifically GLMM.

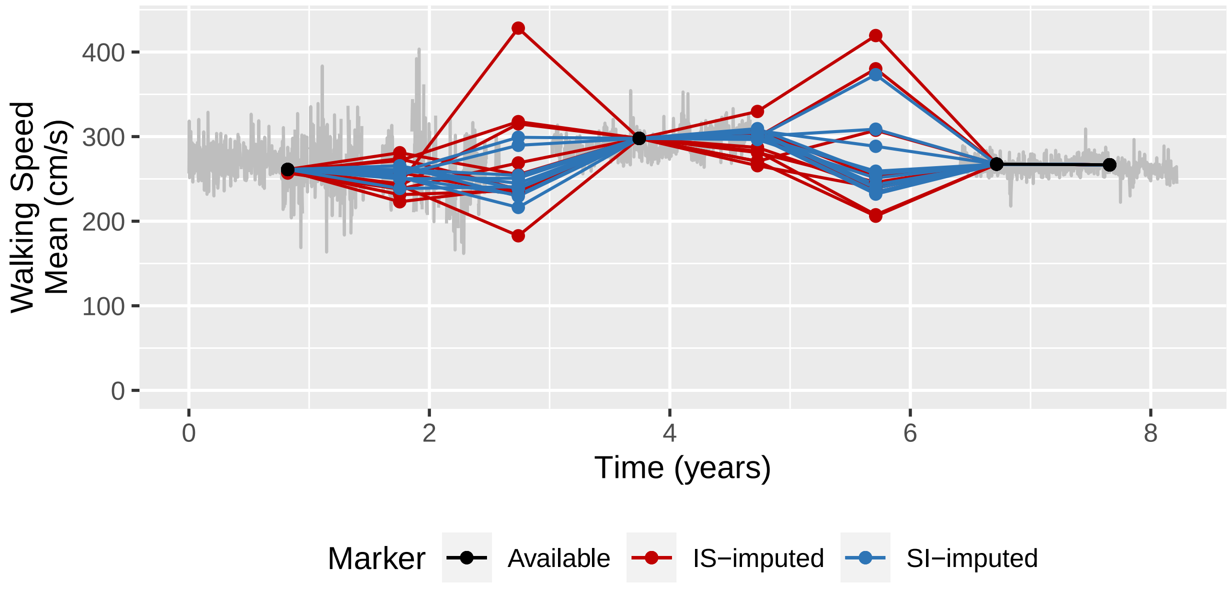
We define each individual's first day of follow-up as the first walking speed mean measurement within the first 30-day summary window that is computed. For example, if an individual's earliest walking speed measurement is within a 30-day window with less than 18 observed daily walking speed measurements, then this is not the first day of follow-up. The first day of follow-up time must be within a window with 18 or more observed measurements, and thus meet the threshold to compute a window summary. The last day of follow-up is 150 days after an individual's last walking speed measurement or a set maximum of 3,000 days. Thus, each individual has different follow-up times. With this restriction, our analysis consists of 145 individuals with an average follow-up time was 1,459 days (approximately 4 years), ranging from 82 days (approximately 0.25 years) to 2,922 days (approximately 8 years). There is a total of 614 MCI measurements with an average of 4.23 MCI measurements per individual.

There is 32.4% missing daily mean walking speeds and 31.4% missing marker summaries. For each missing value or stretch of consecutive missing values, we calculated the number of consecutive days with missing data. The mean stretch of consecutive missing days is 10 days. Of the missing data, 62% are 2 days or less and 10% are 10 days or greater. The maximum stretch is 2,054 days.

We then construct three datasets consisting of summarized walking speeds over 30-day windows and MCI: (1) available data, (2) IS-imputed data, and (3) SI-imputed data. For the available dataset, we summarize walking speed using the process described in Section 3.3.3. The available data consist of MCI and partially missing 30-day walking speed averages. For IS-imputed and SI-imputed data, we implement the imputation processes described in Section 3.2 on the log-transformed walking speed mean. For both imputation processes, the final summarized walking speed consists 30-day averages corresponding to windows preceding MCI measurements. Thus, for the imputed data, each MCI measurement will have a corresponding 30-day walking speed average.

In Figure 3.7, we present an example of one individual's imputed digital biomarker data. We present the original, daily walking speed mean values using a grey line. Summary values that correspond to an outcome are shown. Black data points represent the available summary values. Out of the eight time points corresponding to an MCI measurement, only four instances have enough observed data within the 30-day window to compute a walking speed average value. We

Figure 3.7: An example of one individual's digital biomarker data. The light grey line represents the observed, daily walking speed mean. The black dots indicate the available summary data, red dots are the IS-imputed summary values, and the blue dots are SI-imputed summary values. Summary values that correspond to the outcome are displayed.



also present the imputed summary values for ten multiple imputations that were computed using the IS and SI-imputed data. This individual had MCI measurements approximately every year.

We analyze MCI similar to the process described in Section 3.3.5 for the available data, IS-imputed data, and SI-imputed data. We fit MCI using a GLMM with similar covariates seen in Equation 3.5 in Section 3.3.1. For individual $i = 1, 2, \ldots 145$, we let MCI_{ij} denote the MCI at the individual's jth measurement and $\pi_{ij} = P(\text{MCI}_{ij} = 1)$. For example, for the IS-process, we fit:

$$\text{logit}(\pi_{ij}) = \eta_0 + \eta_1 \overline{\widetilde{m}}_{i1} + \eta_2 \Delta \overline{\widetilde{m}}_{ij} + b_i \tag{3.6}$$

in which $\overline{\widetilde{m}}_{ij} = \sum_{t=1+hj-w}^{hj} \widetilde{m}_{it}/w$, $\Delta\overline{\widetilde{m}}_{ij} = \overline{\widetilde{m}}_{ij} - \overline{\widetilde{m}}_{i1}$, and b_i is the individual specific random effect. Recall that $\overline{\widetilde{m}}_{ij}$ is the 30-day average walking speed for the IS-imputed digital biomarkers. For available and SI-imputed datasets, we replace values of $\overline{\widetilde{m}}_{ij}$ with their respective biomarker summary values. For the IS-imputed and SI-imputed data, we performed 10 multiple imputations and used Rubin's rules (Rubin, 1987) to pool results of all 10 imputations.

Table 3.1: Odd ratios and 95% confidence intervals for η_2 in Equation 3.6 for association of MCI with available, IS-imputed, and SI-imputed digital biomarker data. η_2 is the coefficient parameter for the longitudinal differences between a 30-day average digital biomarker value and the first 30-day average for an individual. These are the odds ratio of MCI for a 1 cm/s increase in average 30-day walking speed average.

Dataset	Odds ratio	Confidence interval
Available	0.908	(0.836, 0.987)
IS-imputed	0.993	(0.961, 1.025)
SI-imputed	0.978	(0.923, 1.036)

Similar to the simulation study, the main focus is the association with the longitudinal differences in 30-day average walking speed and MCI. Thus, we focus on the odds ratio from estimation of η_2. In Table 3.1, for each dataset, we present the odds ratios of MCI for a 1 cm/s increase in 30-day walking speed average. The IS and SI imputed datasets do not have significant odds ratios (confidence intervals span 1). The confidence interval of the odds ratios from the IS and SI-imputed data were smaller than the confidence interval of the odds ratios from the available dataset, but the odds ratios themselves are closer to 1 (0.993 and 0.978, respectively) than that of the available dataset (0.908). Overall, our results show that for 1 cm/s increase in 30-day average walking speed the odds of MCI decreases.

3.5 Discussion

In this chapter, we present an investigation of two processes to simultaneously impute and summarize high-frequency digital biomarkers. In the simulation study, we saw that longitudinal patterns of digital biomarkers and consecutive missing days affect how well the two different imputation processes can be applied to data. We provided evidence of this through performance measurements of coefficient estimates in traditional LDA methods. The simulation study showed that when data have large sections of consecutively missing digital biomarker values, the SI-imputed data better estimated the coefficient values in GLMM than the available and IS-imputed data.

For the simulation studies, we found that the IS-imputed data resulted in attenuated coefficient estimates for scenario 3 and 4 (with 60-day consecutive missing values). As percent missing data increased, the IS-imputed dataset produced relative bias closer to the prescribed coefficient value and with decreasing confidence intervals. Thus, the coefficient estimate for the IS-imputed dataset was close to 0 and had a small standard deviation compared to that of the complete dataset. This is a result of the increased variance among imputed daily digital biomarker data as compared to observed daily digital biomarkers. Thus, the variability among digital biomarker summary values for the IS-imputed data is greater than that of the complete data, resulting in increased variability for the IS-imputed predictors in the fitted GLMM. The IS-imputed data were more prone to attenuation bias than the SI-imputed data in scenarios with large stretches of missing data (as opposed to scenarios with one-day stretches of missing data). For example, if an individual is missing days 60 to 120, we must first use the imputation method described in Section 3.2 to impute the digital biomarker value at day 60. Then, the imputed value at day 61 will depend on the imputed value at day 60. However, when imputing day 60, we sample a value from a Normal distribution with mean μ_{it} and $\hat{\sigma}_t^2$, where $\hat{\sigma}_t^2$ is the mean-square error across individuals from the model fit in Equation 3.1. The estimate, $\hat{\sigma}_t^2$, across individuals is potentially an overestimate of an individual-specific variance. Thus, within an individual, the imputed values may have a larger variance than the observed values. As we progress from day 60 to day 120, we use previously imputed values with larger variance to sample later imputations. Thus, the IS-imputed daily data will have a larger variance than the observed data, resulting in 30-day averages with larger variances, finally resulting in a coefficient estimate shrunk towards zero.

The application to the ISAAC study has similar results. Neither the IS nor SI-imputed dataset had a similar coefficient estimate to that of the available data. For both imputation processes, it appeared that their respective estimates suffered from attenuation bias. The IS and SI-imputed dataset produced estimates close to 0 and small standard deviations compared to that of the available dataset. In the future, we can investigate two options to alleviate attenuation bias. First, we can incorporate individual-specific variance into the imputation process. Second, we can use a slope correction to adjust the coefficient estimate of the imputed dataset, similar to corrections used for measurement error (Berglund, 2012).

Lastly, assuming an missing completely at random (MCAR) missing mechanism, the results of our simulation study and case study indicate using the available dataset minimizes bias without losing efficiency. In the simulation study, at 40% missing data, the sampling standard deviation of the available dataset does not exceed 1.5 times the sampling standard deviation of the complete dataset. This small increase in standard deviation coupled with the attenuation bias of the imputed datasets in the ISAAC study may indicate that the association test of MCI with walking speed can be performed with the available dataset. However, we restricted our simulation study to MCAR missing patterns, thus using available data to analyze the association of MCI with digital biomarkers does not bias our coefficient estimates. For the case study, we may need to consider imputation methods for handling missing at random (MAR) patterns that may alleviate bias from other sources. Research involving the in-home sensors used within the ISAAC study conclude that missing data is not dependent on observed values because missing information is only caused by outages or sensor replacement (Dodge et al., 2015). However, MCAR is a naive missing pattern assumption that prevents the generalizability of our work. In the future, we plan to expand our research to missing mechanisms involving MAR and missing not at random (MNAR).

Within each IS and SI process, we use the five previous measurements to impute a current measurement. For the IS process, the five previous measurements are five *daily* measurements, while the SI process uses five previous *summary* values, which is equivalent to summary values over 150 days. Thus, the SI process incorporates 150 days' worth of digital biomarker data into an imputation while the IS process incorporate five days' worth of information. It may be more appropriate to impute daily values based on the past 150 daily measurements so that imputed summary values and daily values are more comparable. However, using 150 previous measurements to impute a current measurement would increase computation time and potentially limit the IS-process from its flexibility with quick pattern changes.

While outside of the scope of this paper, computation time is a factor when deciding which imputation process is applicable. In cases like scenario 1 and 2, where both imputation processes have low relative bias and decent coverage rate, computation time should be considered. We did not perform a complete investigation of computation time for each imputation process, but imputation for the IS-process took approximately 10 times longer than the SI-process. In the future, I would like to perform a more thorough investigation of computation time versus relative bias.

Overall, we demonstrate the importance of using the characteristics of high-frequency digital biomarker data and their missing data to consider the appropriate imputation and summarization process. These characteristics and conclusions are specific to analysis aims involving traditional LDA methods. By using the SI-process in datasets with long stretches of consecutively missing data, there is less attenuation bias of coefficient estimates than the of the IS-process. Thus, this chapter provides evidence that imputing at the finest time granularity is not necessarily the best option for all digital biomarker data. As high-frequency digital biomarkers are used more for association with or prediction of underlying biological processes, the data processing, including imputation and summarization, become increasingly important to thoroughly investigate and consider within analysis.

CHAPTER 4

Simultaneously Imputing and Summarizing Multiple Correlated High-Frequency Predictors

4.1 Introduction

Digital biomarkers are measurements of a person's physical characteristics, using technology like computing software or sensors, that may inform underlying biological phenomena (Vasudevan et al., 2022). Digital biomarkers are often measured longitudinally to construct specific features or summaries over time to indicate differences within biological processes. This is a common approach in studies using accelerometers to measure physical behavior, like gait, sleep quality, or physical activity. For example, a study following participants with Huntington disease (HD) over one week used wearable sensors to characterize significant differences in gait between participants with and without HD (Andrzejewski et al., 2016). If measured over a long period of time, digital biomarkers have the potential to aid in diagnosis of diseases with physical symptoms, like HD.

Unlike traditional biomarkers, digital biomarkers do not require appointments nor invasive procedures. Thus, digital biomarkers have the potential to be powerful and convenient proxies for burdensome measurements of processes like cognitive function. In Alzheimer's disease, the most effective measurements for diagnosis are cerebrospinal fluid (CSF) biomarkers because the fluid is directly linked to the brain (Niemantsverdriet et al., 2017). However, CSF biomarkers require invasive procedures, and thus, are not routinely measured for diagnosis. As a more feasible alternative, neuropsychological (NP) tests are an established gold standard for evaluating a person's cognitive function and disease progression non-invasively (Kourtis et al., 2019). However, in-person, clinically administered tests are limited in their frequency of administration because (1) there is potential for learning bias, and (2) resources, like clinician's time, are limited. Due to the limited testing, cognitive decline at the early stages of Alzheimer's disease, called mild cognitive impairment (MCI), is hard to detect. This stage of diagnosis is especially important because cognitive decline is still reversible (Toups et al., 2022). Thus, digital biomarkers that are easily measured at high frequencies can aid in early detection of cognitive decline.

To establish digital biomarkers as a proxy for cognitive function, we must establish a relationship between NP test results, which are measured through indicators of MCI, and digital biomarkers. Longitudinal data analysis (LDA) serves as an initial method to analyze association between MCI and digital biomarkers. As discussed in Chapter 2 and 3, there is a discrepancy between frequency of measurements of MCI, measured every 6-12 months, and digital biomarkers, measured every day. One way to accommodate this discrepancy is to summarize the digital biomarkers at corresponding MCI measurements. In the presence of missing digital biomarker data, the summarization process is complicated. Thus, it is important to establish an appropriate process for summarization and imputation of digital biomarkers to then analyze the association.

In the previous chapter, we limited our investigation of imputation and summarization processes to a single digital biomarker. Thus, we imputed and summarized only one digital biomarker for our analysis of MCI. In the context of Alzheimer's disease, research shows that cognitive decline is linked to more than one physical behavior change, including gait, fine motor skills, heart rate variability, sleep patterns, and more (Albers et al., 2015; Bliwise, 2004; Frewen et al., 2013; Kourtis et al., 2019; Rabinowitz and Lavner, 2014; Vitiello et al., 1990). Thus, analysis involving multiple digital biomarkers may result in more efficient, more accurate, and earlier detection of cognitive decline. Some research has successfully distinguished participants with and without a specific cognitive disease, like Huntington or Alzheimer's disease, using multiple digital biomarkers (Buegler et al., 2020; Harms et al., 2022; Zhan et al., 2018). However, these studies do not analyze cognitive function as a longitudinal outcome, meaning issues involving summarization and imputation of multiple digital biomarkers were not addressed.

In addition to using multiple digital biomarkers to improve analysis with cognitive function, potentially correlated digital biomarkers can be used to improve the imputations of missing digital biomarker values. Correlated data is an essential part of the imputation process of missing data. For example, in joint modelling within Multivariate Imputation by Chained Equations (MICE), the underlying assumption is that the complete data are generated from a multivariate normal with a mean and covariance matrix that must be estimated using observed data to impute the missing data (Bauer et al., 2017; Scheuren, 2005). Thus, there is an inherent assumption that leverages correlated data. Involving time-dependent variables, such as digital biomarkers, complicates the imputation process by introducing correlation among repeated measured within an individual. Within the field of time series analysis, there is thorough research on forecasting multivariate time series. Established methods include vector auto-regression (VAR) and recurrent neural network (RNN) (Elman, 1990; Holden, 1995). For example, in VAR, lags from all time series create a model to forecast the potentially correlated time series. In RNNs, a sequence of neural networks over time where the output of previous networks are integrated as inputs for proceeding networks. We aim to integrate methods similar to VAR into the work performed in Chapter 3.

This chapter aims to (1) update the imputation processes in Chapter 3 to leverage information from other potentially correlated digital biomarkers, and (2) investigate how varying levels of correlation and missingness affect imputation and downstream association analysis. In Section 4.2, we describe the updated imputation and summarization processes that involve multiple biomarkers. In Section 4.3, we outline and execute a simulation study involving varying levels of correlation and missingness between biomarkers. In Section 4.5, we apply our methods to select data from the Intelligent Systems for Assessing Aging Change (ISAAC) study to empirically demonstrate the potential benefit of correlated biomarkers in imputations. Finally, in Section 4.6, we summarize our findings and future directions for these methods.

4.2 Methods

We now explain two approaches to impute multiple high-frequency biomarkers. The two approaches directly build off the approaches in Chapter 3 with the introduction of a third dimension of the data to indicate multiple biomarkers. As explained in Chapter 3, any unit of time can be used, but we will assume time is measured in days. We have N individuals with K digital biomarkers, all of which can be followed for a maximum of T days. For individual $i = 1, 2, \dots N$, we let m_{itk} denote the value of the biomarker $k = 1, 2, \dots K$ at day $t = 1, 2, \dots T$. We let $\mathbf{m}_k$ denote the ($N \times T$) matrix of values for biomarker k for all N individuals and time points t, so that each matrix $\mathbf{m}_k$ resembles the singular digital biomarker matrix examined in Chapter 3. Thus, column t of $\mathbf{m}_k$ is equal to $\mathbf{m}_{\cdot tk}$, the vector of the kth biomarker values for all N individuals at time point t, while row i of $\mathbf{m}_k$ is equal to $\mathbf{m}_{i \cdot k}$, the vector of biomarker values for individual i for all time points. For individual i, we let $r_{itk} = 1$ if the value of m_{itk} is observed and $r_{itk} = 0$ if the value of m_{itk} is missing.

During the T days of maximum follow-up for each individual, the low-frequency outcome is measured every h days. For individual $i = 1, 2, \ldots N$, we let Y_{ij} denote the value of the low-frequency outcome measured at the end of an h-day period $j = 1, 2, \ldots J$. We let $\mathbf{Y}$ denote the $(N \times J)$ matrix of binary outcomes for all N individuals, with column j of $\mathbf{Y}$ equal to $\mathbf{Y}_{\cdot j}$, the vector of outcomes for all N individuals at time point t, and row i of $\mathbf{Y}$ is equal to $\mathbf{Y}_{i \cdot}$, the vector of outcomes for individual i for all time points.

Our goal is to examine the association of the high-frequency biomarkers with a low-frequency outcome. Thus, we ultimately want to condense the vector of T values in $\mathbf{m}_{i \cdot k}$ such that we have a summary of values of each digital biomarker in T that correspond to every one of the J values in $\mathbf{Y}_{i \cdot}$. We choose to do this by computing an average of the biomarker values that occur at the same time or precede each outcome measure in a window of $w \leq h$ days. For example, suppose the outcome is measured once every 30 days and we set $w = 7$ days. Thus, corresponding to the first outcome measured at day 30, we would compute an average of each biomarkers' values measured on days 24, 25, . . . , 30, and corresponding to the second outcome measured at day 60, we would compute a corresponding average of biomarker values measured at days, 54, 55, . . . , 60.

The appropriate ordering of imputation and summarization of the biomarkers is uncertain, which motivates us to once again examine two different approaches to imputation:

1. **Impute then Summarize (IS)**: Impute biomarkers at their original granularity (days) while modelling correlation between biomarkers; then summarize each biomarker corresponding to the period of measurement for the outcome.

2. **Summarize then Impute (SI)**: First, summarize the biomarkers corresponding to the period of measurement for the outcome. Summaries will be considered as missing if less than a threshold of observed values exist in that window. Then, impute values for the missing summaries while modelling correlation between biomarkers.

4.2.1 Impute then Summarize (IS) Process

For the IS process, we follow a similar overall procedure to Chapter 3. We first impute values at the original granularity of the biomarker. However, we now incorporate information from other digital biomarkers in the imputations, which is an important change from Chapter 3. We then summarize each biomarker with averages over consecutive days prior to an outcome measurement. Imputation is performed using the MICE library in R (vanBuuren and Groothuis-Oudshoorn, 2011). We now describe the process for producing a single imputation for all of the missing values of one or more biomarkers for each of the N individuals.

Recall in Chapter 3, we used a linear regression model to impute missing marker values at time point t using a specified number of previous days' values from the same marker. In this Chapter, we want to leverage information from other, potentially correlated digital biomarkers. Thus, we simultaneously impute missing marker values at time point t using a linear mixed effects model (LMM) with a random intercept. We regress observed values of all biomarkers at time t on their corresponding observed or imputed biomarker values at times $(t-1), (t-2), \dots, (t-p)$, where p is a pre-defined number of prior days. We connect the digital biomarkers by incorporating a random effect for each individual to adjust for correlation between markers within an individual at a given time point. We further discuss the correlation between a marker k at time t and marker k' at time t for a given individual in Section 4.3.5.

Let $R_{tk} = \sum_{i=1}^{N} r_{itk}$ denote the number of individuals with observed values of biomarker k at time t, and let $m_{v_k tk}$ denote the value for individual $v_k = 1, 2, \dots, R_{tk}$. Let $\dot{\mathbf{m}}_{\cdot tk} = (\dot{m}_{1tk}, \dot{m}_{2tk}, \dots, \dot{m}_{R_{tk}tk})$ denote the vector of observed values at time point t for marker k in $\mathbf{m}_{\cdot tk}$. Note that $\dot{\mathbf{m}}_{\cdot tk}$ at each time point is comprised of observed values for only biomarker k and that the size of vector $\dot{\mathbf{m}}_{\cdot tk}$ changes for each biomarker k.

Recall from Chapter 3 that imputations are sequential so past values in a biomarker's history can be observed or imputed. Thus, for individual $i = 1, 2, \dots N$, we let $\widetilde{m}_{itk}$ denote the value of the imputed or observed biomarker k at day $t = 1, 2, \dots T$. At time point t, for each value in $\dot{\mathbf{m}}_{\cdot tk}$, we need to identify the imputed or observed biomarker history at times $(t-1), (t-2), \dots, (t-p)$, which we denote $\dot{m}_{v_k(t-s)k}$, $v_k = 1, 2, \dots R_{tk}$ and $s = 1, 2, ..., p$.

For a given time point t, we then fit a single LMM for the observed values for all K digital biomarkers

$$\dot{m}_{v_k tk} = \beta_{t0} + \sum_{\ell=1}^{p} \beta_{t\ell} \dot{m}_{v_k(t-\ell)k} + b_{v_k t} + \epsilon_{v_k tk} \tag{4.1}$$

in which $b_{v_k t}$ is the random subject effect at time point t, $b_{1t}, b_{2t}, \ldots, b_{R_{tk}t}$ are independent and normally distributed with mean 0 and variance $\sigma^2_{b_t}$, and $\epsilon_{1tk}, \epsilon_{2tk}, \ldots \epsilon_{R_{tk}tk}$ are independent and normally distributed with mean 0 and variance σ^2_t. We also assume $b_{v_k t}$ and $\epsilon_{v_k tk}$ are mutually independent. This model assumes that the covariance between digital biomarkers within an individual follows a compound symmetric pattern.

Within biomarker k, for each individual i missing a biomarker value ($r_{itk} = 0$), we impute their biomarker value, m^*_{itk}, using a single draw from a normal distribution with conditional variance $\widehat{\sigma}^2_t$, the mean-squared error from the fitted model in Equation 4.1, and conditional mean $\mu_{itk} = \widehat{\beta}_{t0} + \sum_{\ell=1}^{p} \widehat{\beta}_{t\ell} m_{i(t-\ell)k} + \widehat{b}_{it}$, in which $\widehat{\beta}_{tq}$ is the least-squares estimate of $\beta_{tq}, q = 0, 1, \ldots, p$ and $\widehat{b}_{it}$ is the empirical best linear unbiased predictor (BLUP) of b_{it}. If all biomarkers are missing for an individual at time point t, then the individual-specific random effect, b_{vt}, cannot be predicted. In this case, we use $b_{vt} = 0$, meaning we use the population average of b_{vt}. For individual $i = 1, 2, \ldots N$ and biomarkers $k = 1, 2, \ldots, K$, we let $\widetilde{m}_{itk}$ denote the value of the imputed or observed biomarker k at day $t = 1, 2, \ldots T$, i.e. $\widetilde{m}_{itk} = m_{itk}$ when the marker is observed and $\widetilde{m}_{itk} = m^*_{itk}$ when the marker is imputed.

This process is recursive, meaning that we start at $t = 1$ and move forward in time; see Chapter 3 for a further explanation of the recursive process. At each time point, we impute all K biomarkers. We defined the imputation process using only past biomarker values. Thus, at each time point, we impute each biomarker simultaneously. For example, at $t = 10$, we impute the all biomarkers' missing values using their own values at $t = 10 - p, 10 - (p - 1), \ldots, 9$ and predicted, individual random effect. Imputed values for biomarker k at $t = 10$ does not affect imputations of biomarker $k \neq k'$. However, information for biomarker k at $t = 10$ is needed for imputations at $t = 11$ for biomarker $k \neq k'$.

Once the imputations are complete, we summarize the imputed and observed biomarker values. Specifically, for $j = 1, 2, 3, ..., J$, we compute $\overline{\widetilde{m}}_{ijk} = \sum_{t=1+hj-w}^{hj} \widetilde{m}_{itk}/w$, with $\overline{\widetilde{\mathbf{m}}}_{\mathbf{k}}$ denoting the $(N \times J)$ matrix of average biomarker values for biomarker k.

4.2.2 Summarize then Impute (SI) Process

For the SI process, we follow the same general process as described in Chapter 3 with the following changes to accommodate multiple biomarkers. Once again, imputation is performed at the summary level using the MICE library in R. Below we describe the process for producing a single vector of imputed summary values for biomarker k for all N individuals.

Recall from Chapter 3, for the SI process, some summary values are deemed missing if the window contains an insufficient number of observed biomarker values. Specifically, we define p_w as the minimum proportion of observed markers required in a window. We let s_{igk} indicate whether window g meets the threshold of observed biomarker values for individual i and biomarker k, and thus, if the corresponding summary is computed. We let $s_{igk} = 1$ if $\sum_{t=1+w(g-1)}^{wg} r_{itk} \geq p_w w$ and $s_{igk} = 0$ otherwise. This is the same threshold process described in Chapter 3 with the addition of an index for biomarker-specific indicators for each window g.

For individual $i = 1, 2, \ldots N$, we let n_{igk} denote the value of the biomarker summary for window $g = 1, 2, \ldots T/w$ and biomarker $k = 1, 2, \ldots, K$. We let $\mathbf{n}_k$ denote the $(N \times G)$ matrix of values for biomarker k for all N individuals and windows g, so that each matrix $\mathbf{n}_k$ resembles the singular digital biomarker matrix examined in Chapter 3. Thus, column g of $\mathbf{n}_k$ is equal to $\mathbf{n}_{\cdot gk}$, the vector of the kth biomarker values for all N individuals at windows g, while row i of $\mathbf{n}_k$ is equal to $\mathbf{n}_{i \cdot k}$, the vector of biomarker values for individual i for all windows. For $s_{igk} = 1$, we compute the window summary:

$$n_{igk} = \frac{\sum_{t=1+w(g-1)}^{wg} m_{itk} r_{itk}}{\sum_{t=1+w(g-1)}^{wg} r_{itk}} \tag{4.2}$$

If $s_{ig} = 0$ then we need to impute the summary value for the window.

Similar to the IS process, we impute missing marker summary data for window g and biomarker k using LMM with a random intercept. We impute by first regressing observed summary values of all biomarkers at window g on their corresponding observed or imputed summary values at windows $(g-1), (g-2), \ldots, (g-p)$, where p is a pre-defined number of prior summary values. We incorporate a random effect for each individual to adjust for correlation between marker summary values within an individual at a given window.

Let $S_{gk} = \sum_{i=1}^{N} s_{igk}$ denote the number of individuals with computed summary values for window g and biomarker k, and let $n_{l_k gk}$ denote the value for individual $l_k = 1, 2, \ldots, S_{gk}$. Let $\dot{\mathbf{n}}_{\cdot gk} = (\dot{n}_{1gk}, \dot{n}_{2gk}, \ldots, \dot{n}_{S_{gk}gk})$ denote the vector of observed values for marker k in $\mathbf{n}_{\cdot gk}$. Note that $\dot{\mathbf{n}}_{\cdot gk}$ for each window g and biomarker k is comprised of summary values for only biomarker k and that the size of vector $\dot{\mathbf{n}}_{\cdot gk}$ changes for each biomarker k.

Recall from Chapter 3 that imputations are sequential such that past values can be computed or imputed. Thus, for individual $i = 1, 2, \ldots N$ and biomarker $k = 1, 2, \ldots, K$, we let $\widetilde{n}_{igk}$ denote the value of the imputed or computed biomarker summary value for window $g = 1, 2, \ldots T/w$. At window g, for each value in $\dot{\mathbf{n}}_{\cdot gk}$, we need to identify the following summary histories at windows $(g-1), (g-2), \ldots, (g-p)$, which we denote $\dot{n}_{l_k(g-s)k}$ for $l_k = 1, 2, \ldots S_{gk}$ and $s = 1, 2, ..., p$.

For a given window g, we then fit the LMM using the computed biomarker summary values for all K digital biomarkers

$$\dot{n}_{lgk} = \alpha_{g0} + \sum_{\ell=1}^{p} \alpha_{g\ell} \dot{n}_{l(g-\ell)k} + a_{lg} + \gamma_{lgk} \tag{4.3}$$

in which a_{lg} is the random subject effect at window g, $a_{1g}, a_{2g}, \ldots, a_{S_{gk}g}$ are independent and normally distributed with mean 0 and variance $\sigma^2_{a_t}$; and $\gamma_{1gk}, \gamma_{2gk}, \ldots \gamma_{S_{gk}gk}$ are independent and normally distributed with mean 0 and variance σ^2_g. We also assume a_{lg} and γ_{lgk} are mutually independent. This model assumes the covariance between digital biomarker summaries within an individual follows a compound symmetric pattern.

Within biomarker k, for each individual i missing a biomarker summary value ($s_{igk} = 0$), we impute their biomarker value, n^*_{igk}, using a single draw from a normal distribution with conditional variance $\widehat{\tau}^2_g$, the mean-squared error from the fitted model in Equation 4.3, and conditional mean $\nu_{igk} = \widehat{\alpha}_{g0} + \sum_{\ell=1}^{p} \widehat{\alpha}_{g\ell} n_{i(g-\ell)k} + \widehat{a}_{ig}$, in which $\widehat{\alpha}_{gq}$ is the least-squares estimate of α_{gq}, $q = 0, 1, \ldots, p$ and $\widehat{a}_{ig}$ is the empirical BLUP of a_{ig}. Similar to the IS process, if all biomarker summaries are missing for an individual at window g, then the individual-specific random effect, a_{lg}, cannot be predicted. In this case, we use $a_{lg} = 0$, meaning we use the population average of a_{lg}. For individual $i = 1, 2, \ldots N$ and biomarkers $k = 1, 2, \ldots, K$, we let $\widetilde{n}_{igk}$ denote the value of the imputed or computed biomarker k summary at window $g = 1, 2, \ldots T/w$. Thus, $\widetilde{n}_{igk} = n_{igk}$ when the marker summary is computed and $\widetilde{n}_{igk} = n^*_{igk}$ when the marker summary is imputed. This is a recursive process, meaning we start imputing at $g = 1$ and move forward in time. For more information on the recursive process, see Chapter 3.

At this point, we have an $\widetilde{\mathbf{n}}_{\mathbf{k}}$ matrix for each biomarker k of size $N \times G$ of computed and imputed summary biomarker values. We now need to reduce the number of columns of each $\widetilde{\mathbf{n}}_{\mathbf{k}}$ from G to J to match the dimensions of Y. Within each $\widetilde{\mathbf{n}}_{\mathbf{k}}$, we do this by comparing h, the frequency of the outcome, to w, the frequency of each summary. We keep marker summary values for window g if g is a multiple of h/w. For example, suppose the outcome is measured once every $h = 60$ days and we want to average over $w = 30$ days. In this case, $h/w = 2$, thus we keep marker summaries for windows $g = 2, 4, 6, ..., J$. We let $\overline{\mathbf{n}}_{\mathbf{k}}$ denote this reduced set of imputed and computed biomarker summaries for biomarker k for all N individuals over all J outcome measurements times.

4.3 Simulation Methods

We present a simulation study to demonstrate the effectiveness of the imputation processes while imputing multiple, potentially correlated digital biomarkers. The simulation focuses on characteristics of the third simulation scenario from Chapter 3, with a large sine period ($d = 480$) and 60-day sections of missing data. In this simulation study, we focus on scenarios with incremental changes in the correlation between digital biomarkers, in order to examine how this affects the imputed values, and ultimately, how this affects estimating the association of the outcome with a single digital biomarker. We also investigate how increasing the number of digital biomarkers, whether partially missing or fully observed, impacts the imputation precision and downstream analysis. The simulation process generally follows the outline in Chapter 3, and each step is further explained in the following sections.

4.3.1 Simulating Complete Digital Biomarkers and Outcome Data

Similar to Chapter 3, for each complete simulation, we generate data for multiple high-frequency markers (m_{itk}), and lower-frequency binary outcomes (Y_{ij}). We continue to use notation from the methods for these variables. We simulate $N = 40$ individuals with $T = 480$ days of follow-up.

For an individual i and time point t, we simulate $\mathbf{m}_{it}$, where $\mathbf{m}_{it} = (m_{it1}, m_{it2}, \ldots, m_{itK})$, from a multivariate normal distribution with a mean vector $\boldsymbol{\mu}_{it}$, of size K, and variance-covariance matrix, $\mathbf{\Sigma}$ with $K \times K$ dimensions. The mean vector, $\boldsymbol{\mu}_{it}$, is generated from a continuous sine function for each kth component of the vector, where

$$\boldsymbol{\mu}_{it} = \sin \frac{t - \mathrm{s}}{480\mathrm{l}} + a_i \tag{4.4}$$

in which $\mathrm{s} = (s_1, s_2, \ldots, s_K)$ is a vector of size K of biomarker specific phase shifts of the sine function, and $\mathrm{l} = (l_1, l_2, \ldots, l_K)$ is a vector of size K of biomarker specific vertical shifts of the sine function. In Table 4.1, for each biomarker, we identify the values used for s_k and l_k within the sine function for up to four digital biomarkers. The individual-specific intercept, a_i, is normally distributed with a mean of zero and standard deviation of one. The variance-covariance matrix for the digital biomarkers within an individual at time point t, $\mathbf{\Sigma}$, is a K-dimensional matrix with diagonal elements equal to 1 and off-diagonal elements equal to a pre-specified correlation, ρ, of 1, which determines correlation between markers that varies in our simulations.

We also generate binary outcomes that are measured at a lower frequency than the biomarker. As described in Chapter 3, we simulate the binary outcome, Y_{ij}, from a Bernoulli distribution with probability π_{ij} where

$$logit(\pi_{ij}) = \eta_0 + \eta_1 \overline{m}_{i11} + \eta_2 \Delta \overline{m}_{ij1} + e_i \tag{4.5}$$

in which $\overline{m}_{ijk} = \sum_{t=1-w+jh}^{jh} m_{itk}/w$, $\Delta \overline{m}_{ijk} = \overline{m}_{ijk} - \overline{m}_{i1k}$, and the random intercepts, $e_1, e_2, \ldots e_N$, are independent and normally distributed with mean 0 and standard deviation of 0.7. Note that only the first digital biomarker ($k = 1$) is used to generate the outcome. We set $\eta_0 = 0.05$, $\eta_1 = 0.45$, and $\eta_2 = 0.9$. Our goal is to estimate the three coefficients in Equation 4.5. For individuals $i = 1, 2, ...40$, biomarkers $k = 1, 2, \ldots, K$, and corresponding outcome

Table 4.1: Values for biomarker (k) specific phase (s_k) and vertical (l_k) shift of the sine function within the marker generating function (Equation 3.4).

k	s_k	l_k
1	0	1
2	0	2
3	20	1
4	d	1

measurements $j = 1, 2, 3, 4$, the complete data will consist of $\overline{m}_{ijk}$ and Y_{ij}.

4.3.2 Simulating Missing Biomarker Data

We create missingness in the biomarker data as follows. After the initial w days, we generate an indicator of missingness for each biomarker value. We exclude the first w days because our imputations require past marker data to inform the current imputed value, and if the first marker value, at the daily or summary level, were missing, we could not perform imputation.

For individual $i = 1, 2, \ldots N$, we let r_{itk} denote the missing indicator at day $t = 1, 2, \ldots T$ and biomarker $k = 1, 2, 3, 4$. We let $r_{itk} = 1$ if the value of m_{itk} is observed and $r_{itk} = 0$ if the value of m_{itk} is missing. To construct missing indicators across a section of time, we divide the total follow-up time into sections of size 60 with sections $z = 1, 2, ...Z$. For each individual $i = 1, 2, ..., N$, each biomarker $k = 1, 2, \ldots, K$, and each section $z = 1, 2, ..., Z$, we denote a missing indicator, v_{izk}, where $v_{izk} = 1$ if all marker k values in section z are observed and $v_{izk} = 0$ if all marker k values in section z are missing. Each v_{izk}, given $60z > w$, is generated independently from a Bernoulli distribution with a biomarker-specific probability p_k to produce data with $100 * p_k\%$ missing. We use the missing section indicator to construct the missing daily indicator. Thus, we let $r_{itk} = v_{izk}$ for $1 + 60(z - 1) < t \leq 60z$.

4.3.3 Summarizing Observed Biomarker Data

We summarize the observed daily marker values according to the process outlined in Section 4.2.2. For each biomarker k and at each outcome j, if the proportion of observed biomarker values within the window is greater than $p_w = 0.6$, then we compute the window summary value, $n'_{ijk} = \sum_{t=1+hj-w}^{hj} m_{itk} r_{itk} / \sum_{t=1+hj-w}^{hj} r_{itk}$. Otherwise, the window summary for biomarker k is considered missing.

4.3.4 Imputing Missing Biomarker Data

For the IS-LMM process, we impute missing marker values at time point t using the methods described in Section 4.2.1. We set $p = 2$, and thus use the previous observed or imputed values of all biomarkers at $t - 1$ and $t - 2$ to impute the missing digital biomarker value for biomarker k at time point t. Note that this chapter uses $p = 2$, while Chapter 3 used $p = 5$. We summarize the imputed and observed biomarker values. For $j = 1, 2, 3, 4$ and biomarkers $k = 1, 2, \ldots, K$, we compute individual averages over a $w = 30$ window.

For the SI-LMM process, we summarize the each biomarker k over consecutive $w = 30$ day windows as described in Section 4.2.2. We set $p_w = 0.6$ to determine whether the summary biomarker value for a given window is sufficiently observed to be computed. We set $p = 2$, and thus use the previous observed or imputed summary values at window $g - 1$ and $g - 2$ to impute the missing digital biomarker summary value at window g. Then we keep the computed and imputed biomarker summary values at windows $g = 4, 8, 12, 16$ that correspond to the $h = 120$ day frequency of the outcome.

We compare our above imputation process incorporating potentially correlated digital biomarker values to an imputation process that treats the biomarkers as independent. The result is two additional imputed datasets: (1) IS-MLR imputed data, and (2) SI-MLR imputed data. Specifically, we fit a MLR model where markers from the same individual at the same time are considered independent. Thus, the multiple linear regression equation resembles Equation 4.1 with b_{it} removed, and Equation 4.3 with a_{ig} removed.

4.3.5 Correlation of Digital Biomarkers

For all imputation processes, we aim to use other potentially correlated digital biomarkers, in addition to the past information of a missing digital biomarker, to improve the imputation precision of a missing marker value. Thus, we need to examine the two sources of correlation in the data: (1) correlation within a digital biomarker, which we refer to as auto-correlation, and (2) correlation between digital biomarkers.

To quantify within-individual correlation of a digital biomarker, we denote ACF$_k(\tau)$ as the auto-correlation function (ACF) for biomarker k at time points t and $t - \tau$, where τ is time lag between the two time points. The auto-correlation of a marker function is independent of the time point t, thus, it is only a function of τ (Karl, 1989). The ACF for our specific markers (Equation 4.4) is:

$$\text{ACF}_k(\tau) = \cos\frac{\tau}{480} \tag{4.6}$$

For more information on how this ACF was calculated, see Appendix A.

We also examine the within-individual correlation of a pair of marker values at a given time point t. Recall from Section 4.3.1, that we simulate $\mathbf{m}_{it}$ from a multivariate normal distribution with a mean vector, μ_{it}, defined in Equation 4.4 and variance-covariance matrix, $\mathbf{\Sigma}$. By our definition of $\mathbf{m}_{it}$ and $\mathbf{\Sigma}$, the correlations of multiple markers for an individual i at time point t are the off-diagonal components of $\mathbf{\Sigma}$, which we defined with a correlation coefficient, ρ. Thus, the correlation of a pair of marker values, corr$(m_{itk}, m_{itk'}) = \rho$. For the derivation of this correlation, please see Appendix B.

4.3.6 Statistical Analysis

To evaluate the performance of each dataset, we model the probability of the binary outcome using a generalized linear mixed effects model (GLMM) with the same covariates used to generate the outcome in Equation 4.5 in Section 4.3.1. For example, for the complete data, we fit Equation 4.5. For available and imputed datasets, we replace values of $\overline{m}_{ij1}$ with their respective biomarker summary values.

For each simulation, we focus on the estimates, $\hat{\eta}_0$, $\hat{\eta}_1$, and $\hat{\eta}_2$, for the fixed effects. We are most interested in $\hat{\eta}_2$, the fixed effect for the changes from baseline. Recall, we prescribed true values for each coefficient in Section 4.3.1. Coefficient estimates for the fixed effects from analyses using each dataset are compared to one another and the true coefficient parameter value ($\eta_2 = 0.9$). Within each simulation, for all imputed data, we performed five multiple imputations and used Rubin's rules (Rubin, 1987) to pool results of all five imputations. For each dataset, we calculate the relative bias of $\hat{\eta}_2$ across 1,000 simulations.

4.3.7 Design Settings and Datasets

The simulation study aims to (1) compare the imputation methods described in Section 4.2 to methods that ignore correlation, (2) compare imputations of one digital biomarker when potentially correlated digital biomarkers have 0% or 40% missing data, (3) investigate the relationship between correlation and percent missing data of all digital biomarkers, and (4) compare imputations using two and four digital biomarkers. We map each of these aims to a unique simulation design setting, which include a specific datasets that help examine the aim.

For each design setting, we focus on a subgroup of the below eight generated datasets:

1. **Complete dataset**: all observed values with 0% missing values
2. **Available dataset**: observed data with 20-60% missing values
3. **IS-LMM-imputed dataset**: observed and imputed values using the IS-process while modelling correlation of observed digital biomarkers through LMM (uses methods from Section 4.2)
4. **SI-LMM-imputed dataset**: observed and imputed values using the SI-process while modelling correlation of observed digital biomarkers through LMM (uses methods from Section 4.2)
5. **IS-MLR-imputed dataset**: observed and imputed values using the IS-process ignoring correlation through multiple linear regression (MLR) model; i.e. assuming all random intercepts are zero
6. **SI-MLR-imputed dataset**: observed and imputed values using the SI-process ignoring correlation through MLR model; i.e. assuming all random intercepts are zero
7. **IS-1DB-imputed dataset**: observed and imputed values using only the digital biomarker of interest in IS-process; i.e. exact process from Chapter 3 except now with lag of two days
8. **SI-1DB-imputed dataset**: observed and imputed values using only the digital biomarker of interest in SI-process; i.e. exact process from Chapter 3 except now with lag of two days

Note that missing or imputed values mentioned in each dataset refer to a single digital biomarker, the biomarker used in GLMM to model e probability of the lower frequency outcome. We refer to this digital biomarker as marker one. All other biomarkers will be referred to as marker two, three, or four, if simulated. While other biomarkers' missing values may be imputed, marker one is the only marker analyzed.

For our first design setting, we aim to compare imputed datasets that model between-biomarker correlation or ignore between-biomarker correlation. We simulate a setting with two digital biomarkers, in which marker one has 40% missing data and marker two is completely observed. In this simulation setting, we examine varying levels of correlation, $\rho = 0, 0.25, 0.5, 0.75, 0.9$, between the two digital biomarkers. We present the box plots for relative bias of $\hat{\eta_2}$ for datasets 1-6 above, which include the LMM-imputed datasets and the MLR-imputed datasets.

For our second design setting, we aim to compare the imputations of marker one when marker two has varying levels of missingness, and to benchmark the LMM-imputed datasets with the 1DB-imputed datasets from Chapter 3. In addition to the simulated data from the first aim, we simulate data in which marker one has 40% missing data, marker two has 40% missing data, and at varying levels of correlation, $\rho = 0, 0.25, 0.5, 0.75, 0.9$, between the two digital biomarkers. We present the box plots for relative bias of $\hat{\eta_2}$ for datasets 1, 2, 3, 4, 7, and 8 above, which include the LMM-imputed datasets and the 1DB-imputed datasets.

For our third design setting, we aim to investigate the relationship between relative bias, between-biomarker correlation, and percent missing values of two digital biomarkers. We perform simulation settings with: (1) correlation values at 0, 0.25, 0.5, 0.75, and 0.9, (2) percent missing data for marker one at 20, 40, and 60%, and (3) percent missing data for marker two at 0, 20, 40, and 60%. We present the mean relative bias of $\hat{\eta_2}$ and the 95% confidence interval for datasets 3 and 4 above, which include the IS and SI LMM-imputed datasets.

For our final design setting, we aim to compare relative bias from imputations using two and four digital biomarkers. Similar to design setting two, we benchmark the LMM-imputed datasets with the 1DB-imputed datasets from Chapter 3. We simulate data in which marker one has 40% missing data, and all other markers have 0% missing data. Thus, for two digital biomarkers, marker one will have 40% missing data and marker two will have 0% missing data. For four digital biomarkers, marker one will have 40% missing data and marker two, three, and four will have 0% missing data. We also simulate data in which marker one has 40% missing data, and all other markers have 40% missing data. Thus, for two digital biomarkers, both marker one and two have 40% missing data, and for four digital biomarkers, all four markers have 40% missing data. Again, we examine varying levels of correlation, $\rho = 0, 0.25, 0.5, 0.75, 0.9$, between the two or four digital biomarkers. We present the box plots for relative bias of $\hat{\eta_2}$ for datasets 3, 4, 7, and 8, which include the LMM-imputed datasets for two and four digital biomarkers and the 1DB-imputed datasets.

4.4 Simulation Results

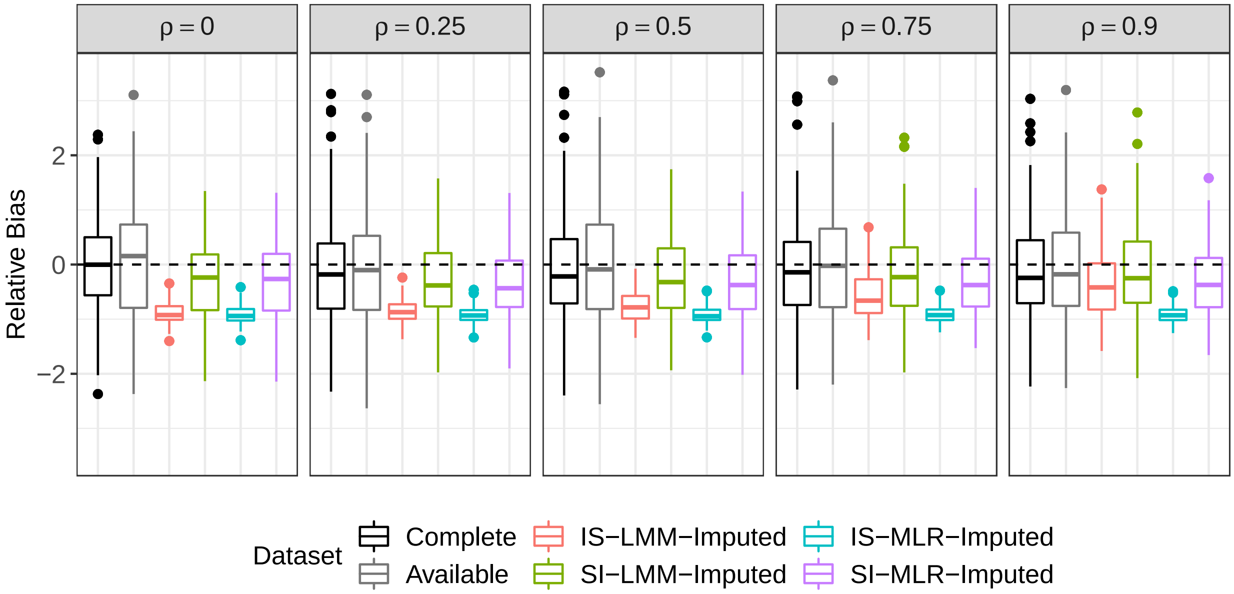
We examine the relative bias of $\hat{\eta}_2$ across simulations for the four design settings. All results examine varying levels of correlation and are summarized across 1,000 simulations. Each section will examine the results of each design setting and build off the previous setting's results.

4.4.1 Setting 1: Two Digital Biomarkers, One Completely Observed

In Figure 4.1, we present the relative bias using a box plot across simulations for each dataset. The figure is divided into various correlation coefficients between the two digital biomarkers: $\rho = 0, 0.25, 0.5, 0.75, 0.9$. Each color in the figure represents a different dataset: (1) complete data (no missing data for marker one), (2) available data (missing data, no values imputed), (3) imputed data using IS process and LMM model, (4) imputed data using SI process and LMM model, (5) imputed data using IS process and MLR model, and (6) imputed data using SI process and MLR model.

In Figure 4.1, when $\rho = 0$, the complete dataset produces a median relative bias of 0 with an interquartile range (IQR) of -0.56 to 0.50. The box plot for the available data, which has 40% observed data, produces a median relative bias of 0.16, but has a larger IQR than the complete data (-0.79 to 0.73). For $\rho = 0$, we see that each imputation processes (IS and SI), regardless of their imputation model (LMM versus MLR), closely align. This finding is because the MLR model correctly treats each digital biomarker for individual i and time point t as independent, and the LMM will result in an estimated random intercept variance close to zero. Finally, at $\rho = 0$, we see that the relative bias of IS and SI-imputed data corroborate the results seen in Chapter 3. That is, the IS-imputed data have greater attenuation bias than that of the SI-imputed data. The IS-imputed data, from the MLR or LMM model, produce a negative median relative bias (-0.92 and -0.94, respectively) and an IQR of (-1.01, -0.76) and (-1.02, -0.81), respectively. While the SI-imputed data, from the MLR or LMM model, also produce a negative median relative bias (-0.24 and -0.26, respectively), the IQRs span across zero, with ranges (-0.83, 0.19) and (-0.84, 0.20), respectively.

Figure 4.1: Relative bias for simulation with two digital biomarkers: marker one with 40% missing data and marker two with 0% missing data. ρ denotes amount of correlation between biomarkers.



As ρ increases, Figure 4.1 indicates that relative bias with complete data varies little, and a similar finding occurs when using available data. In Figure 4.1, the magnitude of the median relative bias decreases towards zero as ρ increases for both IS and SI-imputed data based on LMM. Additionally, the spread of the relative bias for the IS-LMM-imputed data increases as correlation increases. This occurs because the imputed values of marker one are more precise as the correlation increases, which can be attributed to the conditional variance, $\widehat{\sigma}_t^2$, and conditional mean, $\widehat{\mu}_{itk}$, of the normal distribution from which a single imputed value is drawn. Recall in our simulations, we set the diagonal values for the variance-covariance matrix, Σ, which is the sum of the variance for the random effect, σ_b^2 and the residuals, σ_t^2. Therefore, when correlation increases, the normal distribution from which we draw an imputation has a conditional mean closer to the individual-specific marker value and a smaller conditional variance. This leads to increased precision of imputations and thus, less variable predictors in the model.

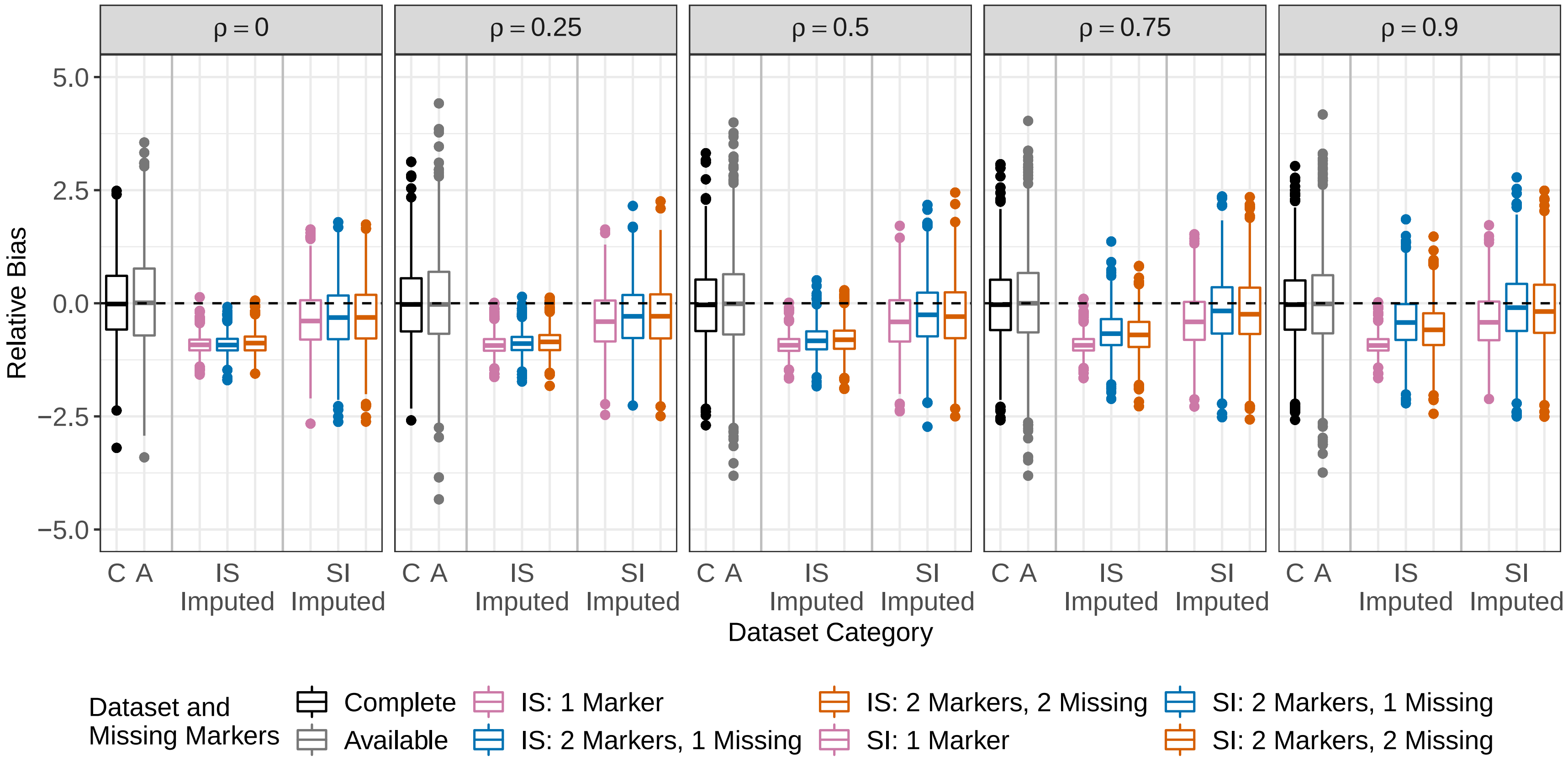
In the MLR-imputed data, we do not see a decreased magnitude in relative bias as correlation increases. For both the IS-MLR-imputed and SI-MLR-imputed data, the median relative bias stays consistently negative regardless of correlation. Once again, this can be attributed to the conditional variance, $\widehat{\sigma}_t^2$, and conditional mean, $\widehat{\mu}_{itk}$, of the normal distribution from which a single imputed value is drawn. The MLR model treats each biomarker as independent, meaning the model ignores within-individual variability. Thus, the total variability from σ_t^2 and σ_b^2 are not separately estimated, but combined within the estimated $\widehat{\sigma}_t^2$. Therefore, regardless of correlation, the normal distribution from which we draw a MLR imputation has a population-averaged conditional mean and constant conditional variance, leading to consistently biased results.

4.4.2 Setting 2: Two Digital Biomarkers, Both Missing

In the previous section, we established that increased correlation improves imputations of marker one when marker two is fully observed. We now explore the relative bias of $\hat{\eta_2}$ when both markers have missing data. We will compare the relative bias for imputed marker one when marker two has 0% missing data (from previous section) and when marker two has 40% missing data. As a final comparison, we include the relative bias for the imputation processes described in Chapter 3, in which only marker one is incorporated into imputations. Thus, we compare the relative bias of eight different datasets as described in Section 4.3.7: (1) complete dataset, (2) available dataset, (3) IS-imputed data using methods from Chapter 3, (4) SI-imputed data using methods from Chapter 3, (5) IS-imputed data using LMM modeling with one missing and one complete marker, (6) SI-imputed data using LMM modeling with one missing and one complete marker, (7) IS-imputed data using LMM modeling with both markers missing 40% data, and (8) SI-imputed data using LMM modeling with both markers missing 40% data.

In Figure 4.2, we present the relative bias using box plots over 1,000 simulations. The figure is divided into different correlation coefficients, labelled at the top of each separated plot. The x-axis contains the dataset category: complete data (C), available data (A), IS-imputed data, and SI-imputed data. We further divide the dataset categories into colors, representing a different dataset or number of missing markers: (1) complete data (no missing digital biomarker data) is represented in black, (2) available data (missing data, no values imputed) is represented in grey, (3) imputed data (both IS and SI) using methods from Chapter 3 are represented in pink, (4) imputed data (both IS and SI) using LMM modeling with one missing and one complete marker are represented in blue, and (5) imputed data (both IS and SI) using LMM modeling with two missing markers are

Figure 4.2: Relative bias for simulation with one and two digital biomarkers, where marker one has 40% missing data. Box plots are divided into categories of: complete data (C), available data (A), IS-imputed data, and SI-imputed data. Colors indication the simulation settings: simulations with one marker is represented in pink, simulations with two markers with marker two at 0% missing data is in blue, and with marker two at 40% missing data in orange. The dashed black line marks zero relative bias. ρ denotes amount of correlation between biomarkers. At each ρ, relative bias for each of the eight datasets are presented.



represented in orange.

We include the results from Chapter 3 as a benchmark for imputations involving two digital biomarkers. In Figure 4.2, across all ρ values and imputation methods, imputations from Chapter 3 have a higher magnitude for median relative bias. Regardless of ρ, the median relative bias from methods in Chapter 3 varies little. Similar to the MLR method examined in the previous section, imputing only using marker one means between-biomarker correlation is ignored. Additionally, as ρ increases, the difference of median relative bias increases between methods from Chapter 3 and methods introduced in Section 4.2. Thus, the addition of a second correlated digital biomarker can help alleviate the attenuation bias in results from the methods from Chapter 3 when markers are correlated.

Next, we compare the relative bias when the second marker has 0% or 40% missing data. For the IS-imputed datasets at $\rho = 0$ and $\rho = 0.25$, the relative bias is slightly closer to zero when marker two has 40% missing values than when it has 0% missing values. For the SI-imputed datasets at $\rho = 0$ and $\rho = 0.25$, the relative bias is close when marker two has 40% missing values and 0% missing values. For both IS and SI-imputed datasets at $\rho = 0.75$ and $\rho = 0.9$, the relative bias is closer to zero using 0% versus 40% missing marker two. This indicates that when the two markers are highly correlated, more observations of marker two increase the precision of marker one's imputations, and subsequently increase relative bias towards zero, but more observations of marker two has the opposite effect when between-biomarker correlation is low. Once again, this is attributed to the conditional variance and conditional mean of the normal distribution for the imputation draws. For 40% missing values for marker two, at higher ρ values, the imputation draws for marker one are from a normal distribution that has an overestimated conditional variance and a conditional mean estimated from less, highly correlated, data than when marker two has 0% missing.

Additionally, the difference in relative bias at the lower and higher ρ is larger for the IS-imputed data than the SI-imputed data. This is because the percent missing marker two is prescribed on the daily level, which translates to a lower level of missing data on the summary-level. Thus, the difference in sample size of marker two summary values is smaller on the summary-level than the daily-level, meaning there is not as large of a lose of precision as on the daily-level imputations.

4.4.3 Setting 3: Two Digital Biomarkers, Varying Correlation and Missingness

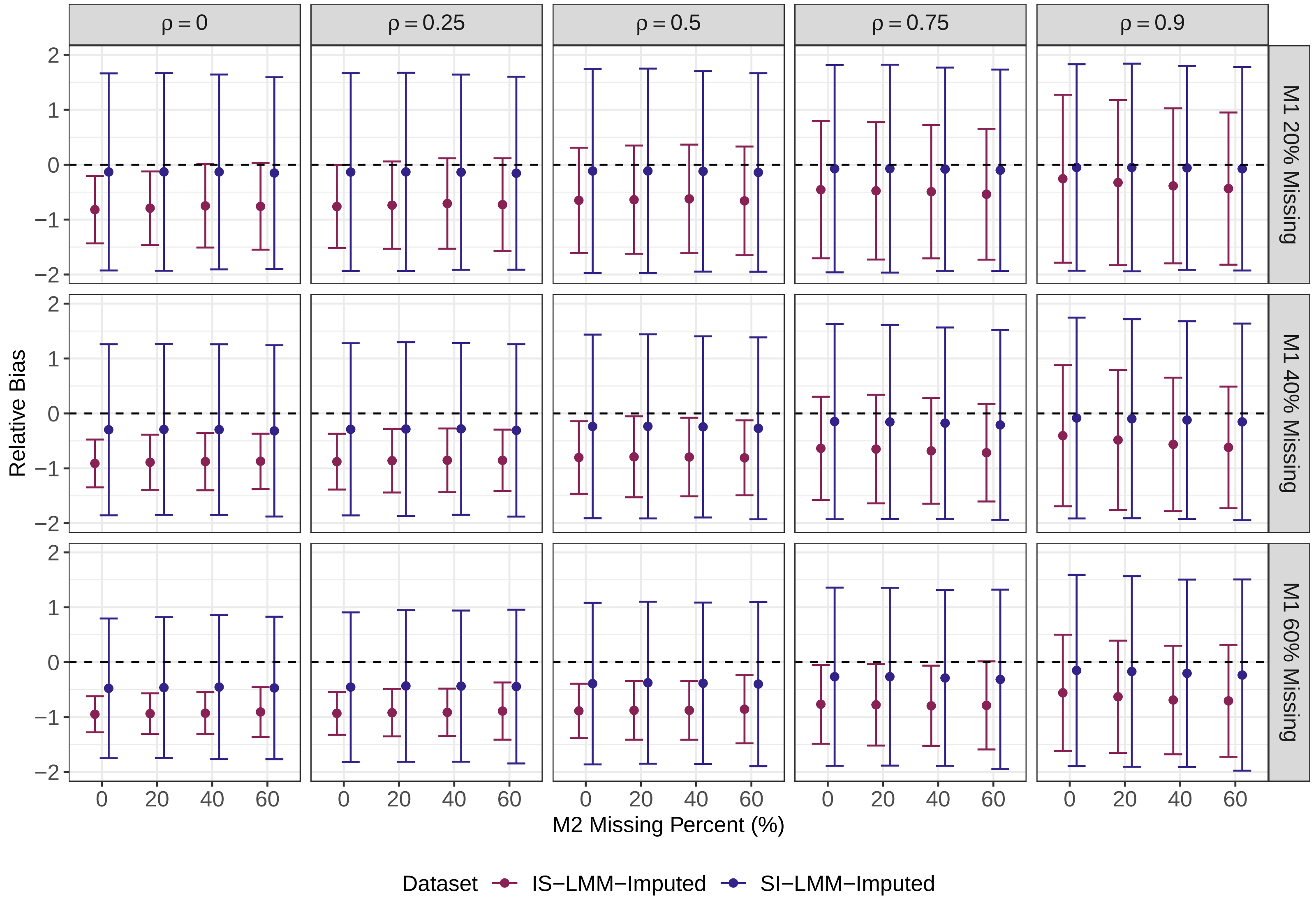
Recall, the third simulation design includes simulation settings investigating (1) correlation levels, with $\rho = 0, 0.25, 0.5, 0.75, 0.9$, (2) percent missing data for marker one at 20, 40, and 60%, and (3) percent missing data for marker two at 0, 20, 40, and 60%.

In Figure 4.3, we present the mean relative bias and the 95% confidence interval across 1,000 simulations per setting. The columns are divided into different ρ values, labelled at the top of each separated plot. The rows are divided into percent missing values for the first digital biomarker, referred to as M1 in the figure. The x-axis contains percent missing values for the second digital biomarker, referred to as M2 in the figure. We present two imputed datasets: (1) IS-LMM-imputed data represented in maroon, and (2) SI-LMM-imputed data represented in purple.

Starting with the first plot, for $\rho = 0$ and 20% missing values for marker one, for both IS and SI-imputed data, there is a slight decrease of the magnitude for the mean relative bias as the percent missing of marker two increases. This indicates that the imputed values of marker one are more precise as the percent missing of marker two increases. For the normal distribution from which imputations are drawn, the estimated conditional mean and the conditional variance are more accurate when there is more missing marker two data. Thus, at $\rho = 0$, an increase in percent missing values for marker two leads to increased precision of the imputation draws from the normal distribution.

Focusing on the plot for $\rho = 0.9$ and 20% missing values for marker one, the magnitude of the mean relative bias increases as the percent missing for marker two increases. Thus, the higher correlation between marker one and two, the greater precision marker two contributes to the imputation of marker one, and that precision is further aided by more observed values of marker two. For $\rho = 0.25, 0.5, 0.75$ and 20% missing values for marker one, the relationship between the percent missing values of marker two and the mean relative bias is less apparent. At $\rho = 0.25$, we see that the best relative bias occurs at 40% missing values of marker two for the IS-imputed data (-0.71) and at 20% missing marker two for the SI-imputed data (-0.13). This is also the case at $\rho = 0.5$, but the mean relative bias across marker two's percent missing varies less, with mean relative bias of -0.65, -0.64, -0.62, and -0.66 for IS-imputed and -0.11, -0.11, -0.12 and -0.14 for

Figure 4.3: Mean relative bias and 95% confidence intervals across simulations with two digital biomarkers. Plots are divided into columns for each correlation level and rows for each percent missing for marker one (M1). X-axis represents the percent missing for marker two (M2). ρ denotes amount of correlation between biomarkers.



SI-imputed. At $\rho = 0.75$, the mean relative bias also vary little, but now the mean relative bias is closest to zero at 0% missing marker two for the IS-imputed data (-0.46) and 20% for the SI-imputed data (-0.07). For other levels of percent missing values of marker one, we see a similar shift in the mean relative bias trend from $\rho = 0$ to $\rho = 0.9$. At each percent missing of marker one at $\rho = 0$, mean relative bias is closest to zero when marker two has 60% missing values, and at $\rho = 0.9$, when marker two has 0% missing values.

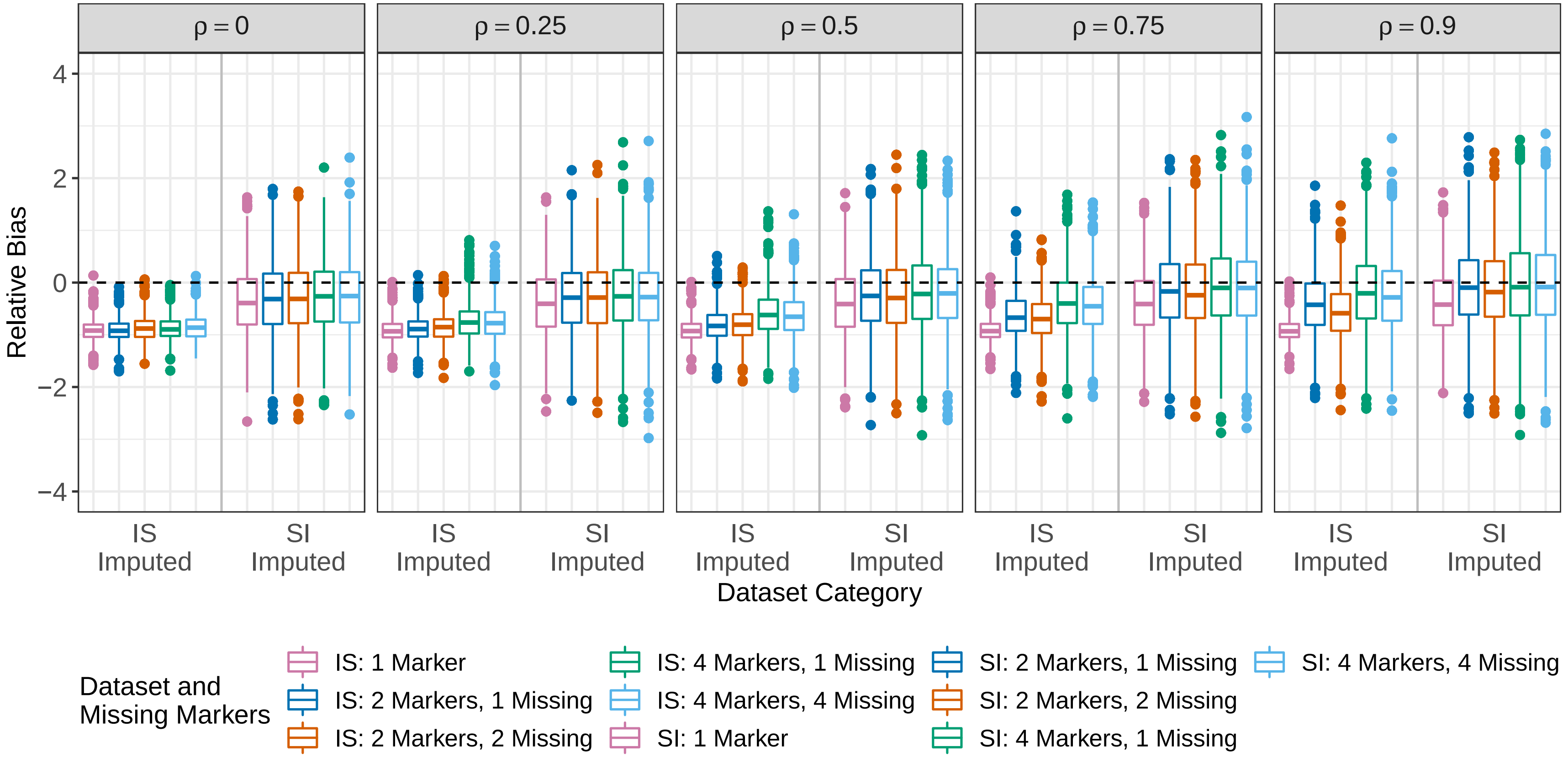
Finally, as we increase the percent missing values for marker one, the mean relative bias shifts towards one for every combination of correlation and percent missing values for marker two. Similar to the simulation results in Chapter 3, attenuation bias increases as the percent missing of marker one increases because with more missing data, the stretches of sequential imputations leads to an overestimate of the conditional variance. See Section 3.5 in Chapter 3 for more explanation.

4.4.4 Setting 4: Four Digital Biomarkers

Now we extend our results to the case where $K = 4$, meaning we simulate and impute four digital biomarkers. We compare the results of four digital biomarker to the results of two digital biomarkers in Section 4.4.2 and one digital biomarker in Chapter 3. For the IS process, we compare the relative bias from five different datasets: (1) IS-imputed data using methods from 3, in which marker one is imputed with only marker one data, (2) IS-imputed data using LMM modeling with one missing and one complete marker, (3) IS-imputed data using LMM modeling with two markers missing 40% data, (4) IS-imputed data using LMM modeling with one missing and three complete markers, (5) IS-imputed data using LMM modeling with all four markers missing 40% data. For the SI process, we compare the relative bias using the same datasets as the IS process.

In Figure 4.4, we present the relative bias using box plots across 1,000 simulations. The plots are divided into different correlation coefficients, labelled at the top of each separated plot. The x-axis contains the dataset category: IS-imputed data and SI-imputed data. We further divide the dataset categories into colors, representing a different dataset or number of missing markers: (1) imputed data (both IS and SI) using methods from 3, in which marker one is imputed with only marker one data, are represented in pink, (2) imputed data (both IS and SI) using LMM modeling with marker one missing 40% and marker two missing 0% are represented in dark blue, and (3)

Figure 4.4: Relative bias for simulation with one, two, and four digital biomarkers, where marker one has 40% missing data. Box plots are divided into categories of IS-imputed and SI-imputed data. Colors indicate the simulation settings. For example, orange represents a simulation of two digital biomarkers where both have 40% missing data. For example, At each correlation coefficient, relative bias for each of the ten datasets are presented. ρ denotes amount of correlation between biomarkers.



imputed data (both IS and SI) using LMM modeling with marker one and two missing 40% are represented in orange, (4) imputed data (both IS and SI) using LMM modeling with marker one missing 40% and markers 2-4 missing 0% are represented in green, and (5) imputed data (both IS and SI) using LMM modeling with marker 1-4 missing 40% are represented in light blue. There is a black, dashed line marking a relative bias of zero.

We continue to see the effect of correlation on relative bias with four digital biomarkers. As ρ increases, the magnitude of the relative bias decreases towards zero. However, there are discrepancies between the rate of change in relative bias as ρ increases. For the SI-imputed data, the relative bias of the simulations with four digital biomarkers slowly approaches a relative bias of zero as ρ increases. This rate of change is almost indistinguishable from the rate for the SI-imputed data with two digital biomarkers. For the for IS-imputed data, as ρ increases, the median relative bias for simulations involving four digital biomarkers increase towards zero at a faster rate than that of two digital biomarkers. For example, at $\rho = 0$, the simulation for four missing markers has a median relative bias of -0.86 and the simulation involving two missing markers has a similar median relative bias, -0.88, making a difference of 0.02 between the median relative biases. At $\rho = 0.25$, that difference increases to 0.08, then 0.15 at $\rho = 0.5$, 0.25 at $\rho = 0.75$, and finally 0.30 at $\rho = 0.9$. This indicates that increased correlation has a greater positive effect on the median relative bias for simulations with four versus two digital biomarkers. For simulations where all biomarkers can be missing, the probability that any given time point has two out of two biomarkers missing is 16%, while the probability of four out of four biomarkers missing is 2.6%. Thus, mores imputations are drawn from a normal distribution with an overestimated conditional variance for two digital biomarkers than for four digital biomarkers.

4.5 Case Study

As an extension of Chapter 3, we continue to examine data from the ISAAC study, a longitudinal cohort study collected through the Oregon Center for Aging and Technology (ORCATECH) at the Oregon Health and Science University. The study consists of 158 individuals who are followed up to 11 years with measurements including cognitive function through NP tests and physical characteristics measured through digital biomarkers. More information can be found in previously published work through ORCATECH (Dodge et al., 2015; Hayes et al., 2014).

We aim to analyze the association of cognitive function with digital biomarkers. Thus, we examine two sets of variables: (1) digital biomarkers, the predictors within our analysis and (2) MCI, the cognitive indicator serving as the binary outcome. We have four digital biomarkers: (1) mean walking speed, (2) total sleep time within one day, (3) total time out of the house within one day, (4) and daily home computer usage. Information on how mean walking speed is measured is described in Chapters 2 and 3. For total sleep time within one day, an individual was considered asleep if movement sensors were fired in the bedroom and were followed by inactivity in the bedroom sensor and all other in-home sensors.

For total time out of the house, sensors measured door openings and activity within the home. If a door opened and was followed by a fifteen minute period without activity in the home, then those fifteen minutes were considered time out of the house. Fifteen minute periods in which individuals were out of the house were summed over the day. For daily home computer usage, computer mouse movement was tracked within five minute periods. Each five minute period was considered in use or not in use, and the five minute periods considered in use were summed over the day. Finally, the binary outcome, MCI, is a clinically administered NP test given to each individual every six months to one year; tests are scored with a clinical dementia rating (CDR). MCI indicates that an individual received a CDR of 0.5 or higher, which indicates that an individual has questionable dementia. More information on each digital biomarker and NP tests is previously published (Dodge et al., 2015; Kaye et al., 2014; Kaye et al., 2011).

In Table 4.2, we present key characteristics of each digital biomarkers' missing data. All digital biomarkers except daily computer usage time have approximately 30-40% missing values at the daily granularity and 25-35% missing values at the summary-level granularity. Daily computer usage has 62.3% missing values at the daily granularity and 51.9% at the summary-level granularity. The mean consecutive missing days for all digital biomarkers is 6-10 days, with mean walking speed with the highest mean at 10 days. Furthermore, the percent of period for missing mean walking speed that are larger than 10 days is 9.9%, indicating that some individuals have very long stretches of missing data. In Table 4.3, we present the Pearson correlation coefficient between digital biomarkers. The maximum magnitude of correlation between digital biomarkers is 0.14 for daily computer usage time and total time out of the house. Mean walking speed is most

Table 4.2: Information on missing data for each digital biomarker within the case study.

Digital Biomarker	% Missing Daily	% Missing Summary	Mean consecutive missing days	% Consecutive missing ≤ 2 days	% Consecutive missing > 10 days
Mean Walking Speed	32.4%	24.8%	10 days	62.1%	9.9%
Total Sleep Time	40.5%	32.3%	9 days	46.9%	11.4%
Total Time Out of House	35.6%	25.8%	7 days	71.9%	7.2%
Daily Computer Usage	62.3%	51.9%	6 days	65.2%	6.3%

Table 4.3: Pearson correlation coefficients between digital biomarkers in the case study. Correlation coefficient is calculated from observed values between each pair of digital biomarkers.

Digital Biomarker	Mean Walking Speed	Total Sleep Time	Total Time Out of House
Total Sleep Time	-0.05		
Total Time Out of House	0.02	0.13	
Daily Computer Usage	0.12	-0.09	-0.14

correlated with daily computer usage with a correlation coefficient of 0.12.

We examine two analyses using the methods from Section 4.3: (1) imputing log-transformed mean walking speed incorporating log-transformed sleep time, and (2) imputing log-transformed mean walking speed incorporating log-transformed sleep time, time out of house, and daily computer usage. For the first analysis, we refined each individual's follow-up time according to observed measurements of total sleep time. Our analysis consists of 144 individuals with an average follow-up time of 1,733 days (approximately 4.75 years). Notice that only one individual has been excluded from the sample in the Chapter 3 case study because they were missing 100% of their total sleep time data. There are a total of 639 MCI measurements with an average of 4.44 MCI measurements per individual.

We aim to compare methods within this chapter and those of Chapter 3 to assess the impact of imputing with a single digital biomarker versus multiple digital biomarkers. We construct seven datasets consisting of summarized walking speed over 30-day windows and MCI:

1. Available data

2. IS-imputed data using only log-transformed mean walking speed

3. SI-imputed data using only log-transformed mean walking speed

4. IS-imputed data using log-transformed mean walking speed and log-transformed sleep time

5. SI-imputed data using log-transformed mean walking speed and log-transformed sleep time

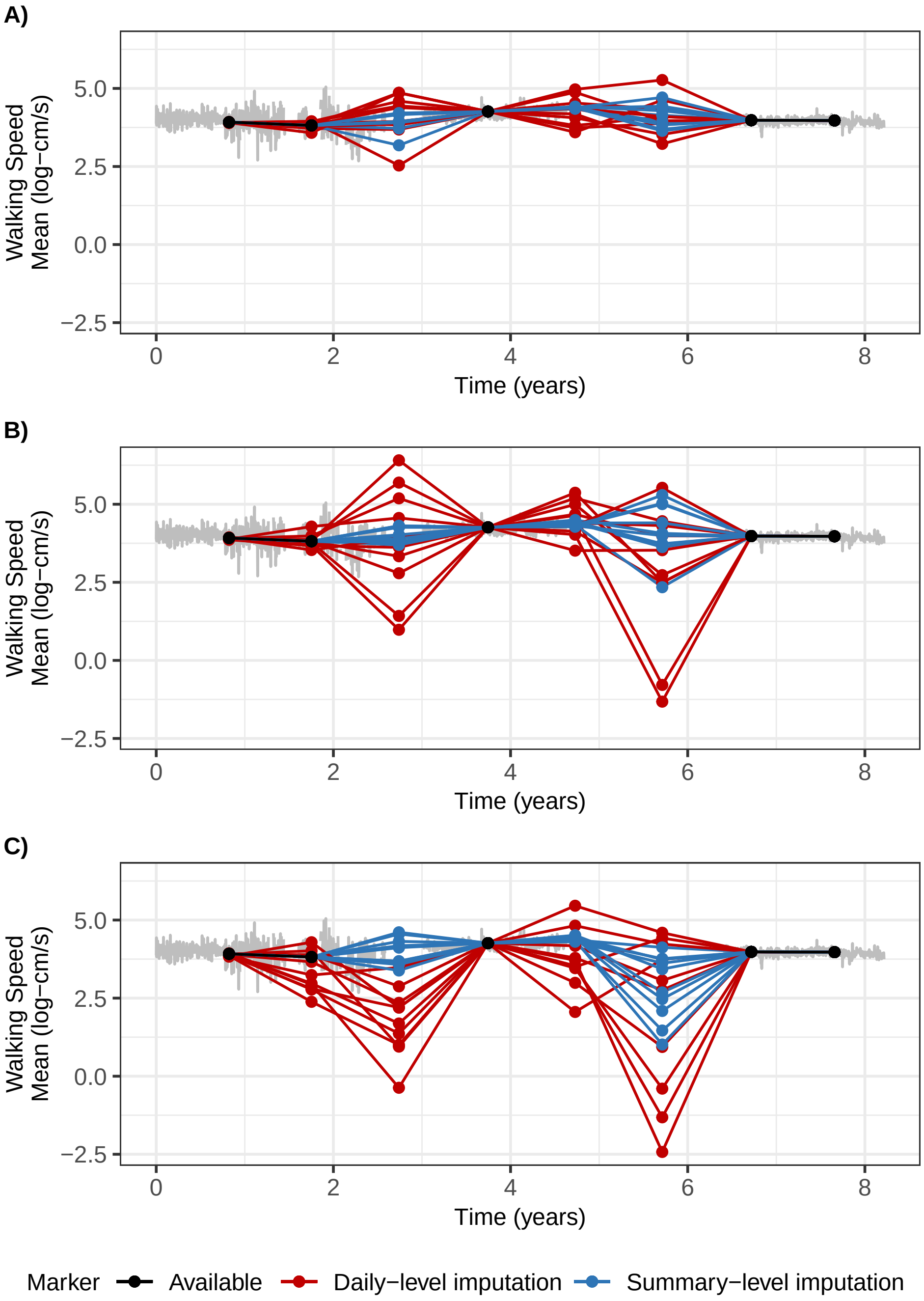
6. IS-imputed data using log-transformed mean walking speed and log-transformed sleep time, time out of house, and daily computer usage

7. SI-imputed data using log-transformed mean walking speed and log-transformed sleep time, time out of house, and daily computer usage

For the available data, we summarize walking speed using the process described in Section 4.3.3. The available data consist of MCI and partially missing 30-day walking speed averages. For IS-imputed and SI-imputed data using only walking speed, we implement the imputation processes described in Section 4.2 on the log-transformed walking speed mean. For the IS-imputed and SI-imputed data using walking speed and sleep time, we implement the imputation processes described in Section 4.2 on the log-transformed walking speed mean and log-transformed sleep time. For all four imputation processes, the final summarized walking speed consists 30-day averages corresponding to windows preceding MCI measurements. Thus, for the imputed data, each MCI measurement will have a corresponding 30-day walking speed average.

In Figure 4.5, we present an example of one individual's imputed digital biomarker data, log-mean walking speed, using one, two, and four digital biomarkers. We present the original, daily walking speed mean values using a grey line. Summary values averaged over 30 days that correspond to an outcome are shown. We also present the imputed summary values for ten multiple imputations that were computed using the IS and SI-imputed data. For both the IS and SI-imputed processes, the summary values are more variable as the number of digital biomarkers increases, except for the second set of IS-imputed summary values in Figure 4.5C. These values are less variable than the set of values for two digital biomarkers, but they do not align with the other set of imputed summary values in 4.5A-B. Also, the SI-imputed summary values are less variable than the IS-imputed summary values for each respective number of digital biomarkers incorporated into imputations.

Figure 4.5: An example of one individual's digital biomarker data. (A) log-transformed mean walking speed is imputed using the methods described in Chapter 3, (B) log-transformed mean walking speed is imputed from log-transformed sleep time, and (C) log-transformed mean walking speed is imputed from log-transformed sleep time, log-transformed total time out of house, and daily compute usage. The light grey line represents the observed, daily walking speed mean. The black dots indicate the available summary data, red dots are the ten IS-imputed summary values, and the blue dots are the ten SI-imputed summary values. Summary values that correspond to the outcome are displayed.



We analyze MCI similar to the process described in Chapter 3. Once again, we fit MCI using a GLMM seen in Equation 4.5 in Section 4.3.1. For individual $i = 1, 2, \ldots 144$, we let MCI$_{ij}$ denote the MCI at the individual's jth measurement and $\pi_{ij} = P(\text{MCI}_{ij} = 1)$. For example, for the IS-process, we fit:

$$\text{logit}(\pi_{ij}) = \eta_0 + \eta_1 \overline{\widetilde{m}}_{i1} + \eta_2 \Delta \overline{\widetilde{m}}_{ij} + b_i \tag{4.7}$$

in which $\overline{\widetilde{m}}_{ij} = \sum_{t=1+hj-w}^{hj} \widetilde{m}_{it}/w$, $\Delta\overline{\widetilde{m}}_{ij} = \overline{\widetilde{m}}_{ij} - \overline{\widetilde{m}}_{i1}$, and b_i is the individual specific random effect. Recall that $\overline{\widetilde{m}}_{ij}$ is the 30-day average of the log-transformed mean walking speed for the IS-imputed digital biomarkers. For available and SI-imputed datasets, we replace values of $\overline{\widetilde{m}}_{ij}$ with their respective biomarker summary values. For the IS-imputed and SI-imputed data, we performed ten multiple imputations and used Rubin's rules (Rubin, 1987) to pool results of all ten imputations.

Similar to the simulation study, the main focus is the association of MCI with the longitudinal differences in 30-day average log-transformed walking speed. In Table 4.4, for each dataset, we present the odds ratios of MCI for a one log-cm/s increase in 30-day walking speed average. Similar to Chapter 3, all IS and SI-imputed datasets, regardless of number of digital biomarkers, do not have a significant odds ratio. The odds ratios for the SI-imputed datasets are lower compared to the odds ratios for their respective IS-imputed datasets. Additionally, as the number of digital biomarkers incorporated into imputations increase, the odds ratio increases and the range of the 95% confidence interval decreases. This indicates that The IS and SI-imputed datasets using four digital biomarkers have the odds ratios closest to one (1.05 and 0.86, respectively). Results from the simulation study indicate that this may be the least biased imputed dataset. However, the example imputations from Figure 4.5 indicate that the imputations using four digital biomarkers have a greater variance than the imputations using two digital biomarkers, indicating that the imputed datasets using four digital biomarkers may have attenuation bias. With the exception of one dataset, our results show that for a one log-cm/s increase in 30-day walking speed average, the odds of MCI decrease.

Table 4.4: Odd ratios and 95% confidence intervals for η_2 in Equation 4.7 for association of MCI with available, IS-imputed, and SI-imputed digital biomarker data. Each dataset is specified by the number of digital biomarkers used for imputations, #DBs.

Dataset	# DBs	Odds Ratio	Confidence interval
Available	1	0.02	(0.001, 0.404)
IS-imputed	1	0.40	(0.082, 1.969)
	2	0.75	(0.355, 1.576)
	4	1.05	(0.657, 1.683)
SI-imputed	1	0.24	(0.032, 1.802)
	2	0.55	(0.145, 2.081)
	4	0.86	(0.409, 1.813)

4.6 Discussion

In this chapter, we investigate the benefits of simultaneously imputing multiple, potentially correlated, digital biomarkers. We simulated digital biomarkers using varying levels of correlation, missingness, and number of digital biomarkers. Our simulation study concluded that correlation most influenced the relative bias, specifically that high levels of correlation resulted in the lowest magnitudes of relative bias. At lower correlation levels, we found that increased percent missing of the additional marker resulted in lower magnitudes of relative bias. At higher correlation levels, we found the opposite relationship: increased percent missing of the additional marker resulted in higher magnitudes of relative bias. Finally, we found that an increase from two to four digital biomarkers led to a decrease in magnitude of relative bias.

In the simulation study, our results showed the advantage of simultaneously imputing multiple correlated digital biomarkers. We compared our methods to two alternative imputation methods: an MLR model that ignores correlation and single marker imputation from Chapter 3. Over all levels of correlation, our method had reduced bias compared to the other two methods. However, at lower correlation values, the improvement was very small. At $\rho = 0$ and $\rho = 0.25$, all methods experience attenuation bias, and especially those that imputed on the finest time granularity. As discussed in Section 4.4, attenuation bias is partially due to poorly estimated conditional means, $\widehat{\mu}_{itk}$. Within the conditional mean, the population-averaged mean profile is estimated from all digital biomarkers, which leads to coefficient estimates that have high standard error at lower correlation levels. Thus, it may be beneficial to implement an alternative model in which correlation is modelled through the fixed effects instead of random effects. We could use a regression model similar to VAR models, where each observed biomarker at a given time point is a weighted regression of its own past values and other digital biomarkers' past values. We did not use this model because the number of estimated coefficients quickly increased with the number of digital biomarkers and lags. For example, if we impute four digital biomarkers with a lag of two, we would have eight coefficients to estimate. If we had five digital biomarkers with a lag of three, we would need to estimate 15 coefficients. Thus, we may not have enough power to detect significant coefficients. Another problem with the model based on VAR is that correlation of values within a digital biomarker is greater than the correlation between digital biomarkers. For example, we found that for a lag of one, the auto-correlation of a digital biomarker is 0.99, while the prescribed correlation with other digital biomarkers is between 0 and 0.9. Thus, when we fit the model based on VAR with the simulated data, the coefficient estimates for other digital biomarker are approximately zero. However, the model based of VAR may improve upon the case study results from Chapter 3, and is worth investigating further in the future.

In the simulation study, attenuation bias was mitigated for higher correlation values. However, as the case study shows, digital biomarkers measured in the real world may have lower correlation values. The highest correlation coefficient between any two digital biomarkers in the case study was 0.14. The estimated odds ratio for the available data was the only significant estimate, which could indicate issues with attenuation bias for the imputed datasets. Also, as the number of digital biomarkers included in the imputation model increased, the odds ratios increased toward one for IS and SI-imputations, respectively. In the simulation study, we found that even at zero correlation, the additional biomarkers helped reduce attenuation bias, but that does not seem to hold for the case study. One potential reason for the discrepancy between the simulation study and the case study is that the simulation study randomly assigned missingness for each digital biomarker. However, in the ISAAC study, certain measurements are recorded with the same sensor system. Thus, if the sensor system is down, then all four digital biomarkers are likely missing. Thus, the benefit of additional markers to aid in the prediction of missing values is negligible.

Additionally, we may not see reduced attenuation bias in the case study for four digital biomarkers because the imputation method for multiple digital biomarkers assumes a compound symmetric correlation structure. The compound symmetric correlation structure is not generalizable to digital biomarkers with varying correlations. In our simulation study, we limited the data to have equal correlation between digital biomarkers, but this a simplification of real world digital biomarkers. As seen in our case study, correlation coefficients were not equal, and thus, our imputation method misclassifies the correlation structure of the data. In the future, we will build on our current model to fit an unstructured correlation matrix (Cai et al., 2011; Tasca et al., 2009). We could construct an individual-specific response and a digital biomarker-specific response through random intercepts.

Within the case study, the estimated odds ratio for the available data were significant and had a estimate closer to zero compared to the estimates from imputed data. This may be a result of our missing completely at random (MCAR) assumption. Recall from the discussion in Chapter 3 that the MCAR assumption within the simulation study and case study limits the generalizability of our imputation methods, and potentially misclassifies the missing mechanism within the case study. Refer to Section 3.5 for more discussion about missing mechanisms within our study. In the future, we plan to extend our missing mechanism assumption to missing at random (MAR) and missing not at random (MNAR).

Overall, we provide evidence that imputing multiple digital biomarkers using imputation methods that model correlation alleviates some of the attenuation bias from single digital biomarker imputations in Chapter 3. High levels of correlation, balanced samples of markers, and increased number of digital biomarkers resulted in less attenuation bias within the simulation study. Additionally, the SI-process was more robust than the IS-process in cases with low correlation, unbalanced samples, and regardless of number of digital biomarkers. We corroborate results from Chapter 3 that imputing at the finest time granularity is not necessarily the best approach, and that imputing digital biomarkers using other digital biomarkers will not necessarily produce more precise imputations. As research continues to use digital biomarkers to help find associations or predictions of diseases, it is important to continue understanding the consequences of ignoring missing data and explore the most effective imputation practices.

CHAPTER 5

Summary and Future Work

In this dissertation, we have presented work on summarizing and imputing high-frequency digital biomarkers, leading to an association analysis with a binary lower-frequency outcome using traditionally-taught longitudinal analysis, like generalized linear mixed effects model (GLMM). Our work of summarizing digital biomarkers demonstrates how the choices in window size for summary values effect the balance between accuracy and efficiency of those summaries in relation to the underlying, true signal. Our work on simultaneously summarizing and imputing a single digital biomarker in the context of longitudinal data analysis demonstrates how accuracy of imputations and downstream association analysis depends on the order in which missing values are imputed and summarized. Finally, our work incorporating multiple digital biomarkers within the imputation process demonstrates how imputations improve with the added information from additional, correlated biomarkers, leading to a more accurate association analysis.

In Chapter 2, we described factors to be considered when summarizing digital biomarker trajectories for longitudinal analysis. We established a guide for choosing an appropriate time granularity by determining a window size in which consecutive summaries are computed. We identified four key factors that affect the choice of appropriate summary window size: (i) duration of follow-up, (ii) variables of interest in analysis, (iii) pattern detection, and (iv) signal to noise ratio (SNR). For each factor, we discussed, in statistical detail, how the characteristics of digital biomarkers affect the accuracy and efficiency of summaries for varying window sizes. We then demonstrated the complex relationship among the listed factors, and with time granularity using average root mean square error (RMSE). We provided examples of data where more frequent pattern changes resulted smaller window sizes for a summary, and with an increase in follow-up time and decrease in frequency of pattern changes, larger window sizes were appropriate. Lastly, we incorporated data from the Intelligent Systems for Assessing Aging Change (ISAAC) study to demonstrate how one would navigate real world data and the factors to decide on a time granularity.

In the future, we plan to accompany this work with an R shiny application that enables users to input the factors discusses, i.e. pattern changes, follow-up time, and marker standard deviation. As a parallel to the summarization process, we will also research smoothing techniques within the field of functional data analysis (FDA) that are applicable to our digital biomarker data.

In Chapter 3, we demonstrate the importance of the process of imputation and summarization, particularly the order in which imputation and summarization are performed. In our simulation study of a single digital biomarker, we saw that longitudinal patterns of digital biomarkers and consecutive missing days affect the ordering of imputation and summarization. When data are missing for several consecutive days, imputing before summarization leads to imprecise covariates and thus, biased estimates within the association analysis. Thus, this chapter provides evidence that sequentially imputing at the finest time granularity is not necessarily the best option for all digital biomarker data. We finished this investigation by applying the imputation processes to a single digital biomarker, walking speed, within the ISAAC study.

In Chapter 3, we restricted our simulation study to missing completely at random (MCAR) missing patterns, thus using available observed data to analyze the association of MCI with digital biomarkers does not bias our coefficient estimates. Research involving the in-home sensors used within the ISAAC study conclude that missing data does not depend on observed values because missing information is only caused by outages or sensor replacement (Dodge et al., 2015). However, MCAR is a naive missing pattern assumption that prevents the generalizability of our work. For example, if we wanted to investigate digital biomarkers in more advanced dementia, subjects may occasionally be hospitalized as a result of falling. This would lead to a period of missing data that is related an observable variable. Thus, in the future, we plan to expand our research to missing mechanisms involving missing at random (MAR) and missing not at random (MNAR).

In Chapter 4, we expanded our imputation processes to incorporate information from other, potentially correlated, digital biomarkers. We created a new imputation method that models the correlation between digital biomarkers, while imputing sequentially over time, using a linear mixed effects model (LMM) with a random intercept. In our simulation study, we found two key factors that decrease relative bias for our imputation method: increased correlation and increased number of digital biomarkers. We also corroborated work in Chapter 3 that the summarize then impute (SI) process is more robust when these key factors are not met. In the case study, we found that the the addition of digital biomarker to the imputation process resulted in highly variable imputed values.

In the future, we plan to expand our imputation methods in two ways. First, we will further investigate a method similar to vector auto-regression (VAR) in which observed digital biomarker values are regressed on lags from all digital biomarkers (Holden, 1995). This will model correlated values through fixed effects without assuming a specific correlation structure. We also can fit a new regression model at each time point or include Bayesian priors for the coefficients to incorporate information from past regression estimates. We believe this method might be better suited to imputed the digital biomarkers in the ISAAC study. Second, we would like to incorporate smoothing methods from FDA into our imputations to decrease the variability of imputed values (Leroux et al., 2018; Morris et al., 2006). We can fit a functional mixed effects model or a function additive mixed model, in which we can combine the benefits of smoothed, non-parametric representations of the digital biomarkers and random effects within a LMM (Guo, 2002; Scheipl et al., 2015).

As high-frequency digital biomarkers are used more for association with or prediction of underlying biological processes, data processing, including imputation and summarization, become increasingly important to thoroughly investigate and consider within analysis. Thus, this dissertation serves as a critical initial step towards translating digital biomarkers into clinically meaningful analysis of underlying biological phenomena, like early cognitive decline for individuals developing Alzheimer's disease.

APPENDIX A

Auto-correlation Function of Digital Biomarkers

In signal processing, the auto-correlation function (ACF) of a sine wave is well defined. The normalized ACF is defined as:

$$
\begin{aligned}
\mathrm{ACF}_k(\tau) &= \frac{\int_{-\inf}^{\inf} 300 \sin \frac{t}{480} 300 \sin \frac{t-\tau}{480} \, dx}{\int_{-\inf}^{\inf} 300 \sin \frac{t}{480} 300 \sin \frac{t}{480} \, dx} \\
\mathrm{ACF}_k(\tau) &= \frac{\int_{a}^{a+960\pi} 300 \sin \frac{t}{480} 300 \sin \frac{t-\tau}{480} \, dx}{\int_{a}^{a+960\pi} 300 \sin \frac{t}{480} 300 \sin \frac{t}{480} \, dx} \\
\mathrm{ACF}_k(\tau) &= \frac{300^2 \int_{a}^{a+960\pi} \sin \frac{t}{480} \sin \frac{t-\tau}{480} \, dx}{300^2 \int_{a}^{a+960\pi} \sin \frac{t}{480} \sin \frac{t}{480} \, dx} \\
\mathrm{ACF}_k(\tau) &= \frac{\int_{a}^{a+960\pi} \sin \frac{t}{480} \sin \frac{t-\tau}{480} \, dx}{\int_{a}^{a+960\pi} \sin \frac{t}{480} \sin \frac{t}{480} \, dx} \\
\mathrm{ACF}_k(\tau) &= \frac{480\pi \cos \frac{\tau}{480}}{480\pi} \\
\mathrm{ACF}_k(\tau) &= \cos \frac{\tau}{480}
\end{aligned}
$$

APPENDIX B

Correlation of Digital Biomarkers

Calculated correlation between two markers within individuals between time point t and $t - 1$. Using Equation 3.4 for the simulated markers, we calculate the correlation of a pair of marker values within individuals at a given time point. We will refer to these markers as m_{itk} and $m_{i(t-1)k'}$.

Below is the initial setup for the correlation:

$$\begin{aligned}
\text{corr}(m_{itk}, m_{itk'}) &= \frac{\text{cov}(m_{itk}, m_{itk'})}{\sqrt{\text{var}(m_{itk})\text{var}(m_{itk'})}} \\
\text{corr}(m_{itk}, m_{itk'}) &= \frac{E[m_{itk}m_{itk'}] - E[m_{itk}]E[m_{itk'}]}{\sqrt{\text{var}(m_{itk})\text{var}(m_{itk'})}}
\end{aligned}$$

We will calculate the correlation by separating the above equation into several parts. First, we calculate $E[m_{itk}m_{itk'}]$:

$$\begin{aligned}
E[m_{itk}m_{itk'}] &= E\left[\left(300\sin\frac{t-s_k}{l_k d} + \omega_{itk}\right)\left(300\sin\frac{t-s_{k'}}{l_{k'}d} + \omega_{itk'}\right)\right] \\
E[m_{itk}m_{itk'}] &= E\Bigg[300^2\sin\frac{t-s_k}{l_k d}\sin\frac{t-s_{k'}}{l_{k'}d} + 300\omega_{itk'}\sin\frac{t-s_k}{l_k d} + \\
&\qquad 300\omega_{itk}\sin\frac{t-s_{k'}}{l_{k'}d} + \omega_{itk}\omega_{itk'}\Bigg] \\
E[m_{itk}m_{itk'}] &= 300^2\sin\frac{t-s_k}{l_k d}\sin\frac{t-s_{k'}}{l_{k'}d} + E[\omega_{itk}\omega_{itk'}] \\
E[m_{itk}m_{itk'}] &= 300^2\sin\frac{t-s_k}{l_k d}\sin\frac{t-s_{k'}}{l_{k'}d} + E[\omega_{itk}]E[\omega_{itk'}] + \text{cov}(\omega_{itk}, \omega_{itk'}) \\
E[m_{itk}m_{itk'}] &= 300^2\sin\frac{t-s_k}{l_k d}\sin\frac{t-s_{k'}}{l_{k'}d} + \Sigma_{k,k'}
\end{aligned}$$

We then calculate $E[m_{itk}]E[m_{itk'}]$:

$$\begin{aligned}
E[m_{itk}]E[m_{itk'}] &= E\left[300\sin\frac{t-s_k}{l_k d}+\omega_{itk}\right]E\left[300\sin\frac{t-s_{k'}}{l_{k'}d}+\omega_{itk'}\right] \\
E[m_{itk}]E[m_{itk'}] &= \left(300\sin\frac{t-s_k}{l_k d}\right)\left(300\sin\frac{t-s_{k'}}{l_{k'}d}\right) \\
E[m_{itk}]E[m_{itk'}] &= 300^2\sin\frac{t-s_k}{l_k d}\sin\frac{t-s_{k'}}{l_{k'}d}
\end{aligned}$$

Finally, we calculate var$[m_{itk}]$var$[m_{itk'}]$:

$$\begin{aligned}
\text{var}[m_{itk}]\text{var}[m_{itk'}] &= \text{var}\left[300\sin\frac{t-s_k}{l_k d}+\omega_{itk}\right]\text{var}\left[300\sin\frac{t-s_{k'}}{l_{k'}d}+\omega_{itk'}\right] \\
\text{var}[m_{itk}]\text{var}[m_{itk'}] &= \Sigma_{k,k}\Sigma_{k',k'}
\end{aligned}$$

We combine the separate calculations from above to finish the calculation of the correlation:

$$\text{corr}(m_{itk}, m_{itk'}) = \frac{\Sigma_{k,k'}}{\sqrt{\Sigma_{k,k}\Sigma_{k',k'}}}$$

where all diagonal elements of $\mathbf{\Sigma}$ are equal to 1 and all off diagonal elements of $\mathbf{\Sigma}$ are equal to ρ. Thus, the final correlation between two markers is:

$$\begin{aligned}
\text{corr}(m_{itk}, m_{itk'}) &= \frac{\rho}{\sqrt{1^2}} \\
\text{corr}(m_{itk}, m_{itk'}) &= \rho
\end{aligned}$$

Bibliography

Aggarwal, N. T., Wilson, R. S., Beck, T. L., Bienias, J. L., and Bennett, D. A. (Dec. 2006). Motor Dysfunction in Mild Cognitive Impairment and the Risk of Incident Alzheimer Disease. *Archives of Neurology*, 63 (12): 1763–1769.

Aisen, P., Touchon, J., Amariglio, R., Andrieu, S., Bateman, R., Breitner, J., Donohue, M., Dunn, B., Doody, R., Fox, N., Gauthier, S., Grundman, M., Hendrix, S., Ho, C., Isaac, M., Raman, R., Rosenberg, P., Schindler, R., Schneider, L., Sperling, R., Tariot, P., Welsh-Bohmer, K., Weiner, M., and Vellas, B. (2017). EU/US/CTAD Task Force: Lessons Learned from Recent and Current Alzheimer's Prevention Trials. *The journal of prevention of Alzheimer's disease*, 4 (2): 116–124.

Akl, A., Taati, B., and Mihailidis, A. (May 2015). Autonomous Unobtrusive Detection of Mild Cognitive Impairment in Older Adults. *IEEE Transactions on Biomedical Engineering*, 62 (5). Conference Name: IEEE Transactions on Biomedical Engineering: 1383–1394.

Albers, M. W., Gilmore, G. C., Kaye, J., Murphy, C., Wingfield, A., Bennett, D. A., Boxer, A. L., Buchman, A. S., Cruickshanks, K. J., Devanand, D. P., Duffy, C. J., Gall, C. M., Gates, G. A., Granholm, A.-C., Hensch, T., Holtzer, R., Hyman, B. T., Lin, F. R., McKee, A. C., Morris, J. C., Petersen, R. C., Silbert, L. C., Struble, R. G., Trojanowski, J. Q., Verghese, J., Wilson, D. A., Xu, S., and Zhang, L. I. (Jan. 2015). At the interface of sensory and motor dysfunctions and Alzheimer's disease. *Alzheimer's & Dementia*, 11 (1). Publisher: John Wiley & Sons, Ltd: 70–98.

Andrzejewski, K. L., Dowling, A. V., Stamler, D., Felong, T. J., Harris, D. A., Wong, C., Cai, H., Reilmann, R., Little, M. A., Gwin, J. T., Biglan, K. M., and Dorsey, E. R. (Jan. 2016). Wearable Sensors in Huntington Disease: A Pilot Study. en. *Journal of Huntington's Disease*, 5 (2). Publisher: IOS Press: 199–206.

Anoop, A., Singh, P. K., Jacob, R. S., and Maji, S. K. (June 2010). CSF Biomarkers for Alzheimer's Disease Diagnosis. *International Journal of Alzheimer's Disease*, 2010: 606802.

Austin, J., Klein, K., Mattek, N., and Kaye, J. (Mar. 2017). Variability in medication taking is associated with cognitive performance in nondemented older adults. *Alzheimer's & Dementia : Diagnosis, Assessment & Disease Monitoring*, 6: 210–213.

Bauer, J., Angelini, O., and Denev, A. (July 2017). *Imputation of Multivariate Time Series Data - Performance Benchmarks for Multiple Imputation and Spectral Techniques*. en. SSRN Scholarly Paper. Rochester, NY.

Berglund, L. (Aug. 2012). Regression dilution bias: Tools for correction methods and sample size calculation. *Upsala Journal of Medical Sciences*, 117 (3): 279–283.

Blennow, K., Hampel, H., Weiner, M., and Zetterberg, H. (Mar. 2010). Cerebrospinal fluid and plasma biomarkers in Alzheimer disease. eng. *Nature Reviews. Neurology*, 6 (3): 131–144.

Bliwise, D. L. (2004). Sleep disorders in Alzheimer's disease and other dementias. eng. *Clinical Cornerstone*, 6 Suppl 1A: S16–28.

Buchman, A. S., Boyle, P. A., Yu, L., Shah, R. C., Wilson, R. S., and Bennett, D. A. (2012). Total daily physical activity and the risk of AD and cognitive decline in older adults. en: 7.

Buegler, M., Harms, R. L., Balasa, M., Meier, I. B., Exarchos, T., Rai, L., Boyle, R., Tort, A., Kozori, M., Lazarou, E., Rampini, M., Cavaliere, C., Vlamos, P., Tsolaki, M., Babiloni, C., Soricelli, A., Frisoni, G., Sanchez-Valle, R., Whelan, R., Merlo-Pich, E., and Tarnanas, I. (Aug. 2020). Digital biomarker-based individualized prognosis for people at risk of dementia. *Alzheimer's & Dementia : Diagnosis, Assessment & Disease Monitoring*, 12 (1): e12073.

Butera, N. M., Li, S., Evenson, K. R., Di, C., Buchner, D. M., LaMonte, M. J., LaCroix, A. Z., and Herring, A. (July 2019). Hot Deck Multiple Imputation for Handling Missing Accelerometer Data. en. *Statistics in Biosciences*, 11 (2): 422–448.

Cai, L., Yang, J. S., and Hansen, M. (2011). Generalized full-information item bifactor analysis. *Psychological Methods*, 16. Place: US Publisher: American Psychological Association: 221–248.

Di, J., Demanuele, C., Kettermann, A., Karahanoglu, F. I., Cappelleri, J. C., Potter, A., Bury, D., Cedarbaum, J. M., and Byrom, B. (Feb. 2022). Considerations to address missing data when deriving clinical trial endpoints from digital health technologies. en. *Contemporary Clinical Trials*, 113: 106661.

Dodge, H. H., Mattek, N. C., Austin, D., Hayes, T. L., and Kaye, J. A. (June 2012). In-home walking speeds and variability trajectories associated with mild cognitive impairment. en. *Neurology*, 78 (24): 1946–1952.

Dodge, H. H. and Estrin, D. (Mar. 2019). Making sense of aging with data big and small. English (US). *Bridge*, 49 (1). Publisher: National Academy of Sciences: 39–46.

Dodge, H. H., Zhu, J., Mattek, N. C., Austin, D., Kornfeld, J., and Kaye, J. A. (Sept. 2015). Use of High-Frequency In-Home Monitoring Data May Reduce Sample Sizes Needed in Clinical Trials. en. *PLOS ONE*, 10 (9): e0138095.

Doherty, A., Jackson, D., Hammerla, N., Plötz, T., Olivier, P., Granat, M. H., White, T., Hees, V. T. v., Trenell, M. I., Owen, C. G., Preece, S. J., Gillions, R., Sheard, S., Peakman, T., Brage, S., and Wareham, N. J. (Feb. 2017). Large Scale Population Assessment of Physical Activity Using Wrist Worn Accelerometers: The UK Biobank Study. en. *PLOS ONE*, 12 (2). Publisher: Public Library of Science: e0169649.

Dollins, A. B., Zhdanova, I. V., Wurtman, R. J., Lynch, H. J., and Deng, M. H. (Mar. 1994). Effect of inducing nocturnal serum melatonin concentrations in daytime on sleep, mood, body temperature, and performance. en. *Proceedings of the National Academy of Sciences*, 91 (5): 1824–1828.

Dorsey, E. R., Papapetropoulos, S., Xiong, M., and Kieburtz, K. (2017). The First Frontier: Digital Biomarkers for Neurodegenerative Disorders. *Digital Biomarkers*, 1 (1). Publisher: Karger Publishers: 6–13.

Eby, D. W., Silverstein, N. M., Molnar, L. J., LeBlanc, D., and Adler, G. (2012). Driving behaviors in early stage dementia: A study using in-vehicle technology. *Accident Analysis & Prevention*, 49: 330–337.

Ehlers, D. K., Aguiñaga, S., Cosman, J., Severson, J., Kramer, A. F., and McAuley, E. (Oct. 2017). The effects of physical activity and fatigue on cognitive performance in breast cancer survivors. en. *Breast Cancer Research and Treatment*, 165 (3): 699–707.

Elman, J. L. (Apr. 1990). Finding structure in time. en. *Cognitive Science*, 14 (2): 179–211.

Enders, C. K., Mistler, S. A., and Keller, B. T. (June 2016). Multilevel multiple imputation: A review and evaluation of joint modeling and chained equations imputation. *Psychological Methods*, 21 (2). Publisher: American Psychological Association: 222–240.

Erler, N. S., Rizopoulos, D., Rosmalen, J. v., Jaddoe, V. W. V., Franco, O. H., and Lesaffre, E. M. E. H. (2016). Dealing with missing covariates in epidemiologic studies: a comparison between multiple imputation and a full Bayesian approach. en. *Statistics in Medicine*, 35 (17). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.6944: 2955–2974.

Fletcher, G. F., Ades, P. A., Kligfield, P., Arena, R., Balady, G. J., Bittner, V. A., Coke, L. A., Fleg, J. L., Forman, D. E., Gerber, T. C., Gulati, M., Madan, K., Rhodes, J., Thompson, P. D., Williams, M. A., and American Heart Association Exercise, Cardiac Rehabilitation, and Prevention Committee of the Council on Clinical Cardiology, Council on Nutrition, Physical Activity and Metabolism, Council on Cardiovascular and Stroke Nursing, and Council on Epidemiology (Aug. 2013). Exercise standards for testing and training: a scientific statement from the American Heart Association. eng. *Circulation*, 128 (8): 873–934.

Frewen, J., Finucane, C., Savva, G. M., Boyle, G., Coen, R. F., and Kenny, R. A. (Dec. 2013). Cognitive function is associated with impaired heart rate variability in ageing adults: the Irish longitudinal study on ageing wave one results. eng. *Clinical Autonomic Research: Official Journal of the Clinical Autonomic Research Society*, 23 (6): 313–323.

Goldberg, T. E., Harvey, P. D., Wesnes, K. A., Snyder, P. J., and Schneider, L. S. (Mar. 2015). Practice effects due to serial cognitive assessment: Implications for preclinical Alzheimer's disease randomized controlled trials. *Alzheimer's & Dementia : Diagnosis, Assessment & Disease Monitoring*, 1 (1): 103–111.

Gorus, E., De Raedt, R., Lambert, M., Lemper, J.-C., and Mets, T. (Sept. 2008). Reaction Times and Performance Variability in Normal Aging, Mild Cognitive Impairment, and Alzheimer's Disease. en. *Journal of Geriatric Psychiatry and Neurology*, 21 (3). Publisher: SAGE Publications Inc STM: 204–218.

Guo, C., Lu, M., and Chen, J. (Mar. 2020). An evaluation of time series summary statistics as features for clinical prediction tasks. *BMC Medical Informatics and Decision Making*, 20 (1): 48.

Guo, W. (2002). Functional Mixed Effects Models. *Biometrics*, 58 (1). Publisher: [Wiley, International Biometric Society]: 121–128.

Harms, R. L., Ferrari, A., Meier, I. B., Martinkova, J., Santus, E., Marino, N., Cirillo, D., Mellino, S., Catuara Solarz, S., Tarnanas, I., Szoeke, C., Hort, J., Valencia, A., Ferretti, M. T., Seixas, A., and Santuccione Chadha, A. (June 2022). Digital biomarkers and sex impacts in Alzheimer's disease management — potential utility for innovative 3P medicine approach. *The EPMA Journal*, 13 (2): 299–313.

Hayes, T. L., Riley, T., Mattek, N., Pavel, M., and Kaye, J. A. (2014). Sleep Habits in Mild Cognitive Impairment. *Alzheimer disease and associated disorders*, 28 (2): 145–150.

Holden, K. (1995). Vector auto regression modeling and forecasting. en. *Journal of Forecasting*, 14 (3). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/for.3980140302: 159–166.

Hsu, H.-H., Chang, C.-Y., and Hsu, C.-H. (Jan. 2017). Big Data Analytics for Sensor-Network Collected Intelligence. Elsevier Science & Technology.

Iaboni, A., Spasojevic, S., Newman, K., Schindel Martin, L., Wang, A., Ye, B., Mihailidis, A., and Khan, S. S. (2022). Wearable multimodal sensors for the detection of behavioral and psychological symptoms of dementia using personalized machine learning models. en. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, 14 (1). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/dad2.12305: e12305.

Jack, C. R. and Holtzman, D. M. (Dec. 2013). Biomarker Modeling of Alzheimer's Disease. en. *Neuron*, 80 (6): 1347–1358.

James, G. M. (2002). Generalized linear models with functional predictors. en. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64 (3): 411–432.

Ji, L., Chow, S.-M., Schermerhorn, A. C., Jacobson, N. C., and Cummings, E. M. (Sept. 2018). Handling Missing Data in the Modeling of Intensive Longitudinal Data. *Structural Equation Modeling: A Multidisciplinary Journal*, 25 (5). Publisher: Routledge _eprint: https://doi.org/10.1080/10705511.2017.1417046: 715–736.

Karl, J. H. (Jan. 1989). "7 - Application of the Fourier Transform to Digital Signal Processing". en. In: *Introduction to Digital Signal Processing*. Ed. by J. H. Karl. San Diego: Academic Press, pp. 127–164.

Kaye, J., Mattek, N., Dodge, H., Buracchio, T., Austin, D., Hagler, S., Pavel, M., and Hayes, T. (Feb. 2012). One walk a year to 1000 within a year: continuous in-home unobtrusive gait assessment of older adults. eng. *Gait & Posture*, 35 (2): 197–202.

Kaye, J., Mattek, N., Dodge, H. H., Campbell, I., Hayes, T., Austin, D., Hatt, W., Wild, K., Jimison, H., and Pavel, M. (Jan. 2014). Unobtrusive measurement of daily computer use to detect mild cognitive impairment. *Alzheimer's & Dementia*, 10 (1). Publisher: John Wiley & Sons, Ltd: 10–17.

Kaye, J. A., Maxwell, S. A., Mattek, N., Hayes, T. L., Dodge, H., Pavel, M., Jimison, H. B., Wild, K., Boise, L., and Zitzelberger, T. A. (July 2011). Intelligent Systems for Assessing Aging Changes: Home-Based, Unobtrusive, and Continuous Assessment of Aging. en. *The Journals of Gerontology: Series B*, 66B (suppl_1): i180–i190.

Keadle, S. K., Lyden, K., Staudenmayer, J., Hickey, A., Viskochil, R., Braun, B., and Freedson, P. S. (July 2014). The independent and combined effects of exercise training and reducing sedentary behavior on cardiometabolic risk factors. *Applied physiology, nutrition, and metabolism = Physiologie appliquee, nutrition et metabolisme*, 39 (7): 770–780.

Kourtis, L. C., Regele, O. B., Wright, J. M., and Jones, G. B. (Feb. 2019). Digital biomarkers for Alzheimer's disease: the mobile/wearable devices opportunity. en. *npj Digital Medicine*, 2 (1). Number: 1 Publisher: Nature Publishing Group: 1–9.

Law, L. L., Rol, R. N., Schultz, S. A., Dougherty, R. J., Edwards, D. F., Koscik, R. L., Gallagher, C. L., Carlsson, C. M., Bendlin, B. B., Zetterberg, H., Blennow, K., Asthana, S., Sager, M. A., Hermann, B. P., Johnson, S. C., Cook, D. B., and Okonkwo, O. C. (2018). Moderate intensity physical activity associates with CSF biomarkers in a cohort at risk for Alzheimer's disease. en. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, 10 (1). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1016/j.dadm.2018.01.001: 188–195.

Leroux, A., Xiao, L., Crainiceanu, C., and Checkley, W. (2018). Dynamic prediction in functional concurrent regression with an application to child growth. en. *Statistics in Medicine*, 37 (8). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.7582: 1376–1388.

Li, K.-C., Jiang, H., and Zomaya, A. Y. (Jan. 2017). Big data management and processing.

Likhachev, D. V. (Aug. 2017). Selecting the right number of knots for B-spline parameterization of the dielectric functions in spectroscopic ellipsometry data analysis. en. *Thin Solid Films*, 636: 519–526.

Lyons, B. E., Austin, D., Seelye, A., Petersen, J., Yeargers, J., Riley, T., Sharma, N., Mattek, N., Dodge, H., Wild, K., and Kaye, J. A. (2015). Corrigendum: Pervasive computing technologies to continuously assess Alzheimer's disease progression and intervention efficacy. eng. *Frontiers in Aging Neuroscience*, 7: 232.

Morris, J. S., Arroyo, C., Coull, B. A., Ryan, L. M., Herrick, R., and Gortmaker, S. L. (Dec. 2006). Using Wavelet-Based Functional Mixed Models to Characterize Population Heterogeneity in Accelerometer Profiles. *Journal of the American Statistical Association*, 101 (476). Publisher: Taylor & Francis _eprint: https://doi.org/10.1198/016214506000000465: 1352–1364.

Newell, J., McMillan, K., Grant, S., and McCabe, G. (Mar. 2006). Using functional data analysis to summarise and interpret lactate curves. eng. *Computers in Biology and Medicine*, 36 (3): 262–275.

Niemantsverdriet, E., Valckx, S., Bjerke, M., and Engelborghs, S. (2017). Alzheimer's disease CSF biomarkers: clinical indications and rational use. *Acta Neurologica Belgica*, 117 (3): 591–602.

Parker, B. J. and Wen, J. (Jan. 2009). Predicting microRNA targets in time-series microarray experiments via functional data analysis. en. *BMC Bioinformatics*, 10 (1). Number: 1 Publisher: BioMed Central: 1–10.

Rabinowitz, I. and Lavner, Y. (Aug. 2014). Association between finger tapping, attention, memory, and cognitive diagnosis in elderly patients. eng. *Perceptual and Motor Skills*, 119 (1): 259–278.

Ramsay, J. O. (2006). "Functional Data Analysis". en. In: *Encyclopedia of Statistical Sciences*. American Cancer Society.

Rubin, D. (1987). "Multiple imputation for nonresponse in surveys". In:

Ruppert, D. and Carroll, R. J. (2000). Theory & Methods: Spatially-adaptive Penalties for Spline Fitting. en. *Australian & New Zealand Journal of Statistics*, 42 (2): 205–223.

Scarmeas, N., Albert, M., Brandt, J., Blacker, D., Hadjigeorgiou, G., Papadimitriou, A., Dubois, B., Sarazin, M., Wegesin, D., Marder, K., Bell, K., Honig, L., and Stern, Y. (May 2005). Motor signs predict poor outcomes in Alzheimer disease. *Neurology*, 64 (10): 1696–1703.

Scheipl, F., Staicu, A.-M., and Greven, S. (Apr. 2015). Functional Additive Mixed Models. *Journal of Computational and Graphical Statistics*, 24 (2). Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/10618600.2014.901914: 477–501.

Scheuren, F. (Nov. 2005). Multiple Imputation. *The American Statistician*, 59 (4). Publisher: Taylor & Francis _eprint: https://doi.org/10.1198/000313005X74016: 315–319.

Seelye, A., Mattek, N., Sharma, N., Riley, T., Austin, J., Wild, K., Dodge, H. H., Lore, E., and Kaye, J. (2018). Weekly observations of online survey metadata obtained through home computer use allow for detection of changes in everyday cognition before transition to mild cognitive impairment. eng. *Alzheimer's & Dementia: The Journal of the Alzheimer's Association*, 14 (2): 187–194.

Shi, Z., Rundle, A., Genkinger, J. M., Cheung, Y. K., Ergas, I. J., Roh, J. M., Kushi, L. H., Kwan, M. L., and Greenlee, H. (June 2020). Distinct trajectories of moderate to vigorous physical activity and sedentary behavior following a breast cancer diagnosis: the Pathways Study. en. *Journal of Cancer Survivorship*, 14 (3): 393–403.

Silbert, L. C., Dodge, H. H., Lahna, D., Promjunyakul, N.-O., Austin, D., Mattek, N., Erten-Lyons, D., and Kaye, J. A. (2016). Less Daily Computer Use is Related to Smaller Hippocampal Volumes in Cognitively Intact Elderly. eng. *Journal of Alzheimer's disease: JAD*, 52 (2): 713–717.

Sperling, R. A., Aisen, P. S., Beckett, L. A., Bennett, D. A., Craft, S., Fagan, A. M., Iwatsubo, T., Jack Jr., C. R., Kaye, J., Montine, T. J., Park, D. C., Reiman, E. M., Rowe, C. C., Siemers, E., Stern, Y., Yaffe, K., Carrillo, M. C., Thies, B., Morrison-Bogorad, M., Wagster, M. V., and Phelps, C. H. (May 2011). Toward defining the preclinical stages of Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & Dementia*, 7 (3). Publisher: John Wiley & Sons, Ltd: 280–292.

Tackney, M. S., Cook, D. G., Stahl, D., Ismail, K., Williamson, E., and Carpenter, J. (June 2021). A framework for handling missing accelerometer outcome data in trials. *Trials*, 22 (1): 379.

Tan, F. E. S., Jolani, S., and Verbeek, H. (Oct. 2018). Guidelines for multiple imputations in repeated measurements with time-dependent covariates: a case study. en. *Journal of Clinical Epidemiology*, 102: 107–114.

Target Heart Rates Chart (2020). en.

Tasca, G. A., Illing, V., Joyce, A. S., and Ogrodniczuk, J. S. (July 2009). Three-level multilevel growth models for nested change data: A guide for group treatment researchers. *Psychotherapy Research*, 19 (4-5). Publisher: Routledge _eprint: https://doi.org/10.1080/10503300902933188: 453–461.

Toups, K., Hathaway, A., Gordon, D., Chung, H., Raji, C., Boyd, A., Hill, B. D., Hausman-Cohen, S., Attarha, M., Chwa, W. J., Jarrett, M., and Bredesen, D. E. (Jan. 2022). Precision Medicine Approach to Alzheimer's Disease: Successful Pilot Project. en. *Journal of Alzheimer's Disease*, 88 (4). Publisher: IOS Press: 1411–1421.

Ullah, S. and Finch, C. F. (Dec. 2013). Applications of functional data analysis: A systematic review. en. *BMC Medical Research Methodology*, 13 (1). Number: 1 Publisher: BioMed Central: 1–12.

vanBuuren, S. and Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, 45 (3): 1–67.

Vasudevan, S., Saha, A., Tarver, M. E., and Patel, B. (Mar. 2022). Digital biomarkers: Convergence of digital health technologies and biomarkers. en. *npj Digital Medicine*, 5 (1). Number: 1 Publisher: Nature Publishing Group: 1–3.

Verbeke, G. and Molenberghs, G. (Jan. 2008). Linear Mixed Models for Longitudinal Data. en. Google-Books-ID: TjrhBwAAQBAJ. Springer Science & Business Media.

Vitiello, M. V., Prinz, P. N., Williams, D. E., Frommlet, M. S., and Ries, R. K. (July 1990). Sleep disturbances in patients with mild-stage Alzheimer's disease. eng. *Journal of Gerontology*, 45 (4): M131–138.

Wakim, N. I., Braun, T. M., Kaye, J. A., and Dodge, H. H. (2020). Choosing the right time granularity for analysis of digital biomarker trajectories. en. *Alzheimer's & Dementia: Translational Research & Clinical Interventions*, 6 (1). _eprint: https://alz-journals.onlinelibrary.wiley.com/doi/pdf/10.1002/trc2.12094: e12094.

Wang, J.-L., Chiou, J.-M., and Müller, H.-G. (June 2016). Functional Data Analysis. en.

Zhan, A., Mohan, S., Tarolli, C., Schneider, R. B., Adams, J. L., Sharma, S., Elson, M. J., Spear, K. L., Glidden, A. M., Little, M. A., Terzis, A., Dorsey, E. R., and Saria, S. (July 2018). Using Smartphones and Machine Learning to Quantify Parkinson Disease Severity: The Mobile Parkinson Disease Score. *JAMA Neurology*, 75 (7): 876–880.

Zhang, Y., Li, H., Keadle, S. K., Matthews, C. E., and Carroll, R. J. (2019). A Review of Statistical Analyses on Physical Activity Data Collected from Accelerometers. *Statistics in biosciences*, 11 (2): 465–476.