Statistical Methods for Modeling Disease Progression

By

Jacquelyn Neal

Thesis

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

MASTER OF SCIENCE

in

Biostatistics

August 10, 2018

Nashville, Tennessee

Approved:

Professor Dandan Liu, Ph.D.

Professor Qingxia Chen, Ph.D.

DEDICATION

This thesis is dedicated to my sisters.

To the memory of Christina Elaine Neal, who would have been a formidable woman and supportive of this endeavor.

To Stephanie Michelle Neal, who overcomes adversity daily as a female engineer looks fantastic doing it.

To Alexandra Lynne Neal, first-rate accountant who beats chronic kidney disease every day with the brightest smile imaginable.

To Rebecca Suzanne Neal, current competitive cheerleader and future pharmacist, who has always terrified and amazed us and will continue to do so in the future.

I love you all.

ACKNOWLEDGEMENTS

I'd like to thank the faculty of the Department of Biostatistics for their teaching, mentoring, and support throughout this process. Specifically, my adviser, Dandan Liu, for all of her support, both academically and beyond. I truly would not have made it this far without her. Thanks go to Jonathan Schildcrout, Bryan Shepard, and Aaron Kipp for their involvement and input in my research work over the years, and Qingxia Chen for serving as the second reader of this thesis.

My time at Vanderbilt has been a major period of life-changing education and research. I would not have been able to complete this process without the friendship and support of the other students in the graduate program, especially my fellow cohort members, Allison Hainline and Sandya Lakkur. My time in the department is filled with fond memories, and I thank all of the students for their friendship.

Lastly, the support of my family has been invaluable. Thank you to my parents and sisters for their constant support and for raising my spirits during the tough parts. Thanks to my extended family for their support and love. To my Nashville friends and family, thank you for your friendship and non-school conversations. Finally, thank you to my mother, Lynne Buff Neal, for being my first role model of a woman in STEM and showing me the possibilities of all I could do in the future.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

For progressive diseases, building reliable and robust statistical models for disease progression is important yet challenging for several reasons. Risk factors can impact disease pathology in different ways during different disease stages, which might imply disease stage specific modeling. Risk factor effects on disease progression might be heterogeneous depending on the projection time period of interest (e.g. short-term vs long-term), which might lead to projection term specific inference. Accurate individualized disease progression prediction model will aid clinical decision making in early prevention, early identification and early treatment for disease management.

Disease progression could be quantified using either continuous measure of discrete measures. This paper focuses on discrete measures of disease progression such as clinical disease stages or severity of disease which is quite common in clinical practice. Modeling the transitions between multiple disease states is especially needed in progressive diseases like Alzheimer's disease (AD). The pathology of AD introduces unique issues in the statistical modeling of its progression. Clinicians use annual visits to assess cognition, functional activity, and other aspects of patient health to determine a patient's diagnosis. Generally speaking, there are two commonly used statistical modeling approaches for disease progression; each has its advantages and limitations. The first approach is cross-sectional. It is used to relate baseline characteristics of patients to cross-sectional clinical diagnosis evaluated at the most recent visit [2]. This method only includes disease progression measurement at one time point and thus ignores information collected between baseline and the cross-sectional follow-up time. The other approach is the multi-state model which models transition between disease states and is often used under the Markov assumption that future

evolution only depends on the current states. This approach takes advantage of the multiple assessments of potential risk factors and disease states, and thus might improve statistical efficiency of the model. Both transition intensities models and transition probability models have developed where the former models instantaneous transition rates and the latter models log-odds of transitions. It should be noted that due to the Markov assumption, only transitions between temporally adjacent measurements are modeled [7]. When the Markov assumption is violated, which is often the case in reality, the Markov multi-state model will result in biased estimations.

On the other hand, the partly conditional model proposed by Pepe [8] does not assume the Markov property, and directly models a future outcome given the current measurement with flexible time lags between the current measurement time and the future projection time. In the original papers, binary and continuous outcomes were discussed. This method could be easily extended to ordinal outcomes which is applicable to progressive disease with multiple disease stages.

The structure of the following thesis is structured in four parts. Section 2 is an introduction to the motivating disease, AD, and its pathology, which is necessary to understand for determining which methods to use for a particular scientific question. Section 3 discusses in detail three major types of statistical modeling methods for disease progression. Cross-sectional methods and multi-state methods have been used in AD literature, but the third method, partly conditional models, has not yet been used in this disease area. The results of applying each of these methods to an AD dataset are presented in Section 4, and Section 5 discusses the results and gives an overview of the strengths and limitations of each model type.

CHAPTER 2

MOTIVATING EXAMPLE

2.1 Introduction to AD

AD is a progressive disease with risk factors that have time-dependent/stage-dependent effect, with three general stages: Normal Cognition, MCI, and AD. Normal cognition is used for those with no evidence of cognitive impairment or deficits in activities of daily living due to cognitive impairment. MCI is the prodromal stage of AD and is characterized by report of a cognitive change by the patient, informant, or a clinician, reduced cognitive ability on neuropsychological testing, but an absence of dementia. AD is diagnosed when there are cognitive impairments in at least 2 domains and a decrease in daily functions due to these declines. There are other types of dementia that are not on the AD pathway, are not considered here. AD is currently considered an irreversible condition with no cure, and the best course of treatment is to delay the onset of the disease. Figure 2.1 below describes potential transitions between NC, MCI and AD [10]. Though AD is not reversible, a transition between MCI and NC exists as a potential pathway due to this possibility in the clinical context of AD.

2.2 Potential Time-Dependent Effects

The etiology of AD has not been fully established, but there is evidence the pathophysiological process of AD begins years before the presence of clinical symptoms appear [3]. A cascade hypothesis has been suggested to describe the progression of the disease, which is illustrated by Figure 2.2. The cascade hypothesis illustrates the measurement of a biomarker is dependent on both time and current disease stage of a participant, which suggests time-varying covariates are of interest when modeling

disease stage in AD. As the cascade hypothesis illustrates, the effectiveness of risk factors as predictors are expected to vary depending on the age of the patient and their current stage of disease.

Some risk factors will be more effective for predicting disease progression over shorter time periods while others have a smaller effect that amplifies over time. Amyloid beta (Aβ) accumulation is thought to be the first step of AD pathology, beginning years before clinical symptoms manifest. As illustrated in Figure 2.2, changes in $A\beta$ may be useful in identifying patients at risk from transitioning from normal cognition to MCI, while memory measures may be more useful when the patient has already transitioned to the MCI stage. Based on the cascade hypothesis, Aβ measurements would be more useful for prediction for a patient earlier in the disease, though the long-term effect of the build-up may not be visible until decades later. Another measure of interest, cognitive complaint status, has the potential to be useful in prediction throughout AD progression. It measures whether a participant and/or their informant notices issues in the patient's memory, with four levels: "No Complaint", "Self complaint only", "Informant complaint only", and "Both complaints". Having a cognitive complaint at a younger age is a sign of deteriorating memory and would appear earlier in the disease. However, a major change in complaint status over a short period of time, such as a patient having no complaint at one visit and then both types of complaint at the next is a marker of a more rapid decline in cognitive ability and an increase in disease progression. Predictors like cognitive complaint may be useful for prediction models when the potential time-dependent importance is incorporated.

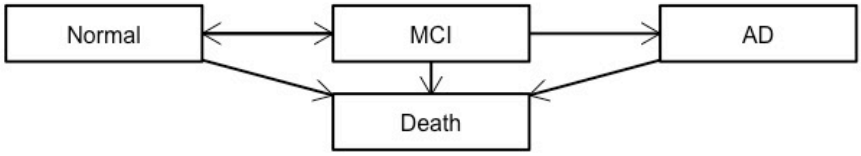


Figure 2.1: Diagram of AD Disease Progression

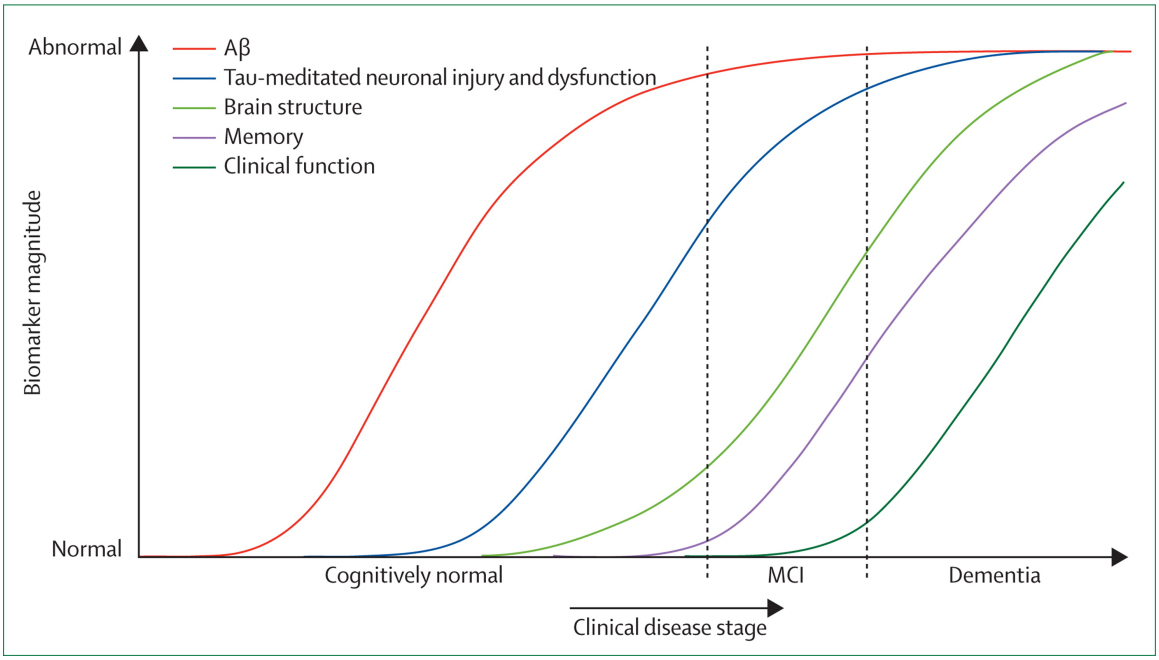


Figure 2.2: Hypothetical model of dynamic biomarkers of the AD pathologyJack et al. [3]

CHAPTER 3

METHODS

3.1 Overview

Examining risk factors of disease progression is a preliminary step to building a strong predictive model for progression. Several statistical methods have already been established that can be applied to studying disease progression, each with their own strengths, assumptions, and limitations. Three general categories of methods will be examined: cross-sectional methods, which use one record per patient, multi-state Markov models, and partly conditional longitudinal models. Each method has particular scientific questions it can answer and specific assumptions that influence parameter interpretation.

Examining risk factors over time requires longitudinal data collection. In the case of AD, most patients are examined annually, and building a longitudinal dataset with enough observations per person to use traditional longitudinal models can take decades. In the interim, the relationship between disease stage and risk factors over time can be examined using cross-sectional methods, where the dataset is reduced to one record per patient. This can be done using the most recent or last diagnosis per patient as the outcome, while adjusting for time in study or difference between current age and baseline age.

There are additional considerations which need to be investigated for a progressive disease like AD. Due to the nature of AD and the study population, death must be considered as a competing risk. Ignoring the competing risk introduces bias into the models which can affect prediction accuracy, so the competing event of death must be included in the model [12].

3.2 Cross-sectional Methods

3.2.1 Introduction

Cross-sectional statistical methods are commonly used for modeling the relationship between disease progression and risk factors, where the dataset is reduced to one record per patient by cross-sectionally looking at the dataset and comparing with a chosen baseline record. For example, a record could be built by evaluating disease progression based on the most recent outcome relative to the baseline outcome, where risk factors from baseline will be used while adjusting for time between baseline and the most recent visits. For categorical or ordinal measures of disease progression, logistic regression or proportional odds models could be used [2]. A pre-defined projection period can also be used to define the cross-section time. For example, to build a model for t year progression, the record with the follow-up time closest to t years could be used to construct disease progression measures relative to baseline disease stages.

3.2.2 Notation

Let t_{ij} denote the j-th follow-up time for subject i, $j = 1, \ldots, m_i$, where m_i is the number of visits for subject i and $t_{i1} = 0$. Let $Y_i(t)$ denote diagnosis for subject i at time t, where $Y_i(t) \in \{k : 1, \ldots K\}$. Let $X_i(t)$ denote risk factors for subject i at time t.

3.2.3 Model and Estimation

For progressive disease with multiple disease stages, $Y_i(t_{im_i})$ can be considered as an ordinal outcome with K potential values. In application, a proportional odds model for diagnosis at the latest follow-up will be used given baseline outcome and adjusting for baseline measurements of all covariates, $X_i(0)$. Time in the study of the patient is used as a covariate in the model. It is important to note the time or

visit number at which a change in diagnosis occurred is not taken into account with this model. The proportional odds model for disease progression from $Y_i(0) = j$ to $Y_i(t_{im_i}) = k$ could then be specified as

$$\text{logit}(P(Y_i(t_{im_i}) > k | Y_i(0) = j, X_i(0))) = \alpha_{jk} + X_i(0)^T \beta_{jk} + t_{im_i} \delta_{jk}, j < k \tag{3.1}$$

where β_{jk} is the vector of coefficients for baseline covariates, and δ_{jk} is the coefficient for cross-sectional follow-up time. The model is conditioned on the baseline disease stage for simplified modeling and comparison between methods. Logistic regression is commonly used in practice where the outcome as the last follow-up measurement was collapsed to two states: stable vs progression.

3.2.4 Calculating Risk

A calculable quantity to compare between different models of disease progression is needed. The model estimates of each method cannot be directly compared as they differ in type of quantity. The cross-sectional and partly conditional models give estimates as odds ratio, while multi-state models model transition intensities as hazard ratios. Though the cross-sectional methods and longitudinal methods both estimate odds ratios, they have different interpretations and cannot be directly compared either. Because the interest of this study lies in t year prediction, calculating predicted probabilities of progression for specific patient profiles is used for comparison. The formula used for cross-sectional methods is $P(Y(t) > k | Y(0) = j, X(0) = \text{invlogit}(\alpha_{jk} + X(0)^T \beta jk + t\delta_{jk})$, where $\text{invlogit}(x) = \frac{exp(x)}{1+exp(x)}$.

3.3 Multi-state Markov Model

3.3.1 Introduction

Multi-state methods model transitions between multiple states. Under the Markov assumption, only transitions between consecutive visits are included in the likelihood formulation. These methods are relatively new-to-use in AD literature, and have been used only for the last decade. Some studies directly model transition probabilities [11] using polytomous logistic regression, whereas some studies model transition intensities [1, 4]. A multi-state model based on transition intensity is the focus of this paper. The multi-state model based on log-odds of transition is a special case of partly-conditional model and will be discussed in the next subsection. Figure 2.1 displays the transitions of interest between AD stages. Several assumptions are imposed for commonly used multi-state methods in application. The Markov assumption assumes that the future evolution only depend on the current states and might not be valid in practice. In addition, the model assumes equally spaced consecutive time intervals and the actual follow-up times are not used in the modeling, which is termed ignorable observation time process. Since AD is a slowly progressive disease and patients are typically assessed annually, this assumption might be reasonable for studies with pre-specified regular visits.

3.3.2 Notation

Let the intensity matrix Q be the matrix of transition intensities, with the element in the l-th row and k-column denoting transition intensities $q_{lk}(t, X(t))$ moving from state l to state k given covariates measured at time t, $X(t)$.

3.3.3 Model and Estimation

Multi-state Markov models for multi-state outcomes where the states are observed at finite series of time $t_{ij},\ j = 1,\ldots,m_i$ were described by Kalbfleisch and Lawless [6]. The full likelihood is written as the product of probabilities of transition between all consecutive visits over all individuals i. This likelihood models the transition intensities between stages, which can be transformed into hazard ratios. These hazard ratios can then be used to calculate transition probabilities between disease stages of interest.

The model can be specified with the following equations. Covariates can be time-independent or time-dependent.

$$q_{lk}(X(t)) = q_{lk}(0)\exp(X(t)^T\beta_{lk}) \tag{3.2}$$

$$q_{lk}(t,X(t)) = \lim_{\delta t\to 0} P(Y(t+\delta t) = k|Y(t) = l)/\delta t \tag{3.3}$$

$$q_{ll} = -\sum_{k\neq l} q_{lk} \tag{3.4}$$

The likelihood for the model is calculated using the transition probability matrix. Equation 3.2 describes the multi-state model with covariates, where q_{lk} are the transition probabilities for previous state l and current state k. Equation 3.3 represents the instantaneous risk of moving from state l to state k. Each row of Q sums to 0, and diagonal probabilities are described in Equation 3.4.

Multi-state methods can handle the inclusion of time-dependent covariates, though their inclusion complicates parameter interpretations. However, when these model estimates are used to calculate predicted probabilities, the time-dependent covariates should not be used. As discussed in the following section on calculating risk, predicted probabilities are calculated by raising the estimated transition matrix to the power of the time period of interest, t years.

3.3.4 Markov Assumption

The multi-state Markov method requires the use of the Markov assumption such that the likelihood only includes transitions between consecutive visits. The assumption states the future disease stage does not depend on previous history of disease progression given the current stage. Written mathematically, the assumption is $q_{lk}(X_i(t), \mathscr{F}_t^-) = q_{lk}(X_i(t))$, where $\mathscr{F}_t^-$ is full disease history up to time t. For diseases with possible progression in a single direction, this assumption may hold. However, when disease progression can move in multiple directions, the probability of transition to certain states may be dependent on disease history beyond solely their current state. It is important to be aware of this assumption in the modeling process and determine whether it is appropriate for the disease of interest.

3.3.5 Calculating Risk

Predicted probabilities of transition between disease stages after t future time periods can be easily calculated using the estimated transition probabilities [5]. Predicted probabilities for specific patient profiles for t time periods in the future can be done using the formula $P(t) = exp(tQ)$, which is calculated by taking the matrix exponential of the scaled transition intensity matrix. Because these calculations rely on exponential matrices, the larger number of values for l and k require more difficult calculations. The multi-state Markov model at its simplest requires the assumption the transition matrix does not change with time. Therefore, when the time period is one year, calculating predicted probabilities 3 years into the future is 3 transitions, $P(3) = exp(3Q)$. It is important to note the risk prediction for this method is reliant on the assumption of equally spaced visits.

3.4 Partly Conditional Longitudinal Model

3.4.1 Introduction

Under the Markov assumption, the multi-state model only considers short time span between consecutive visits and thus might not be used to provide long-term risk prediction or identify long-term risk factors For progressive disease with complex pathology, understanding potential heterogeneous time effects and being flexible about different projection time spans are crucial.

Pepe and Couper [8] and Pepe et al. [9] developed partly conditional models for continuous and binary outcomes, where the method aims to quantify predictive distribution of $Y(t)$, the outcome evaluated at a future time $t > s$ using $X(s)$, variables of interest measured at or up to t, for a range of t and s of interests, with $t > s$. For continuous outcomes, the partly conditional model is specified as $\mathrm{E}\{Y(t)|Y(s),X(s)\} = X(s)^T\beta(s,t)$. It allows risk factor effects to vary with follow-up time s and projection period $u = t - s$. Partly conditional model with constant s and varying t examines risk factors effect over different projection period, whereas the model with constant $u = t - s$ allows the use of time-dependent risk factors $X(s)$ in predicting progression within a fixed projection period. If $X(s)$ is highly predictive of $Y(t)$, the predictive distributions involving these covariates can be used to select subjects likely to have poor prognosis, or in the case of AD, are more likely to progress to further disease stages. However, if some components of $X(s)$ are found not to be predictive, less emphasis could be placed on those follow-up measurements. These models can be fit using existing GEE methods with some modifications to the setup of the data. Partly conditional models can be implemented in data where the number and timing of visits vary across individuals, which allows for their use with observational data sets.

3.4.2 Model and Estimation

In this section, we extend partly conditional models to ordinal outcomes describing disease progression specified as $\text{logit}(P(Y(t) > k|Y(s) = l, X(s))) = \alpha_{lk}(s,t) + X(s)^T\beta(s,t)$. With the objective of predicting a fixed u_0-year of disease progression and approximating averaged risk factor effects over u_0-years of projection period, the model could be simplified as

$$\text{logit}(P(Y(s+u) > k|Y(s) = l, X(s))) = \alpha_{lk} + f(u) + X(s)^T\beta, j < k \tag{3.5}$$

where $u = t - s$ is the prediction period satisfying $|u - u_0| < c$ for a constant c, and $f(\cdot)$ is a function of u either completely unspecified or with simple form such as linear function of u. It should be noted that the parameters in this model $(\alpha_{lk}, \beta, f(\cdot))$ is specific to u_0 and are assumed to not depend on s because the interest of application and comparison across different statistical modeling is u_0-year prediction of disease progression. Time-dependent effects parameterization could be included in the partly conditional model to address a different scientific question of interest, but is considered in this paper. To fit this model, the dataset must be augmented by obtaining all possible pairs of observations from each subject within domain of interests $\mathscr{D} = \{(s,t) : |t - s - u_0| < c\}$. In other words, for each subject, all pairs with observation time difference falling in the bandwidth of c from u_0 are included in the analysis. We chose appropriate bandwidth such that sufficient number of pairs will be included in the analysis, yet the time-invariant effect assumption with the domain is still valid. For example, in prediction 3-year disease progression, we choose a bandwidth of 1-year such that all paired observations with time difference between 2-4 years from each subject used in the analysis. The interpretation of the parameters are specific to 3-year of risk prediction. If we choose a different projection period, e.g. 5-years, different paired observations will be chosen and the parameter estimates will

have interpretations for 5-year prediction. Variance of β coefficients can be estimated using the sandwich estimator. With the data manipulation as described above, standard methods for fitting a GEE model with independence covariate structure with robust standard errors will fit this model [8].

Partly conditional models take advantages of multiple observations per subject and take into account time-varying risk factors values. It does not have the Markov assumption that disease progression is independent of previous history given current states. With hypothetical equally spaced observation times, multi-state Markov model for transition probabilities could be considered as a special case of partly conditional models where short term disease progression prediction (e.g. annually disease progression for studies with annual visits) is of interests.

3.4.3 Calculating Risk

Calculations of u_0-year predicted probabilities of progression are similar to those with cross-sectional methods with the formula $\text{invlogit}(\alpha_{jk} + f(u_0) + X(s)^T\beta)$ to calculate the predicted probabilities of progression. Confidence intervals can be calculated using the model-estimated covariance matrix.

3.5 Comparison between Three Methods

Cross-sectional methods model the disease stage of the patient at a specific follow-up time, often the final visit. The disease progression within the patient is not directly modeled or incorporated with this method, as the time at which a patient transitions between disease states is not taken into account. The disease stage of the patient at their final follow-up is known, but not the time at which this transition occurred. The multi-state method directly models transition intensities between consecutive time points, but assumes future evolution does not depend on the previous disease history given the current disease stage. This method could account for competing risk

of death by taking death as an absorbing state. To calculate predicted probabilities, this model requires the transition matrix remain constant over time. Using this method for prediction may be difficult due to this restriction, and time-dependent covariates cannot be included if the model goal is prediction. Partly conditional models directly model log-odds of disease progression using generalized estimation equations. This method does not directly model transitions to death, but death can be incorporated using inverse probability weighting. Partly conditional models can incorporate time-dependent covariates in the realm of prediction, as the calculation for the predicted probabilities does not require the odds of transition remain constant over time, unlike the multi-state method.

CHAPTER 4

APPLICATION

4.1 National Alzheimer's Coordinating Center Dataset

The National Institute on Aging created a task force to developed a uniform set of assessment procedures to characterize individuals with mild Alzheimer disease and mild cognitive impairment in comparison with nondemented aging. The resulting Uniform Data Set (UDS) defines a common set of clinical observations collected longitudinally on participants at Alzheimer Disease Centers (ADCs). The UDS was implemented at all ADCs on September 1, 2005. Data obtained with the UDS are submitted to the National Alzheimer's Coordinating Center. The primary goals for NACC are to develop a database that captures and integrates data on all ADC participants and promotes collaborative research among the ADCs. Data needed to be sufficiently comprehensive to allow phenotyping of each individual's cognitive, behavioral, functional, and medical status, but not too burdensome for implementation. The protocol includes detailed guidelines for administration with standard definitions and terminology, allowing for findings to be compared over ADCs. A common set of clinical observations were developed for use on all ADC participants, collected longitudinally in a uniform manner. Other goals are to improve clinical assessment and diagnosis, track change over time, provide data in support of current projects, and stimulate research. Thirty-one ADCs are currently reporting or have reported information to the NACC UDS in the past. Data has been collected by the ADCs since 2005, with over 30,000 participants in the database.

Potential covariates available in NACC include time in study, cognitive complaint, demographics, measures of general health, and neuropsychological test results. Cog-

nitive complaint status refers to whether a participant and/or their informant notices issues in the patient's memory. It has four levels: "No Complaint", "Self complaint only", "Informant complaint only", and "Both complaints". Age, education, Framingham Stroke Risk Profile (FSRP), MMSE, cognitive complaint status, sex, and race were included as covariates in both the cross-sectional and longitudinal models. Age, FSRP, MMSE, and cognitive complaint status are time-dependent covariates measured at each follow-up visit. APOE E4 status (positive or negative), a genetic covariate, is also included as a covariate [13]. For NACC, left-ventricular hypertrophy was not collected, so FSRP is the modified FSRP definition excluding that particular criterion, and called mFSRP throughout the results.

Time in UDS (years) is included as a covariate in cross-sectional models and the time in UDS (years) at each visit is included in longitudinal models. Covariates used in the models presented here are based on those included in a previous cross-sectional model examining diagnosis as a binary outcome [2]. The cross-sectional and multi-state methods both use only baseline values for all covariates, while the partly conditional model uses the time-dependent values for any covariate with measurements at every follow-up visit: Age, mFSRP, MMSE, and cognitive complaint. Comparison between the three methods will be based on 3-year prediction of disease progression.

4.2 Checking Participants with Diagnostic Reversion

The statistical methods considered in this paper assume disease progression is not reversible. This is true in the pathological context of AD, but might not be true in the clinical context of AD. To put the assumption in the clinical context of AD, it is assumed that participants with clinical diagnosis of MCI will not be diagnosed as cognitive normal in any future clinic visits. This might not be true in reality.

There are many reasons where diagnostic reversion might happen. For example, a participant not in a good mood at the clinic visit might be misdiagnosed as MCI, and then correctly diagnosed as cognitive normal in the next visit.

Participants with diagnostic reversion might have different characteristics compared to participants without diagnostic reversion. If we ignore such potential heterogeneous underlying patient characteristics, risk factors effects estimation from disease progression modeling might be biased, especially for multi-state Markov model using consecutive observations. In addition, the Markov assumption might be violated because a cognitive normal participant with previous diagnosis of MCI might be more likely to progress comparing to participants without clinical diagnosis of MCI in the past. To check this assumption, statistical analysis were conducted using NACC participants who are cognitively normal or MCI at enrollment and diagnosed with normal cognition at any time in their follow-up. The first clinic visit with cognitive normal clinical diagnosis was used as the new "baseline'' to create a hypothetical "baseline'' cognitive normal cohort. Proportional odds regression model was used to assess "baseline'' risk factor effects in relation to the cross-sectional clinical diagnosis (Normal, MCI or AD) at the last clinic visits. In addition to the covariates considered in section 4.1, an indicator variable for previous diagnosis of MCI was included as the variable of interest. Table 4.1 shows the odds ratios, 95% confidence intervals, and p-values from this analysis. Patients with a previous history of MCI are 2.1(p-value=0.0002) times more likely to progress compared to those without previous diagnosis of MCI. Results from additional analysis stratified by previous diagnosis of MCI (Table 4.2) showed that the effect of sources of complaint differs between two different cohort. Based on these results, in order to make the three statistical methods comparable, participants with diagnostic reversion will be excluded from the analysis dataset.

Table 4.1: Parameter Estimates from Proportional Odds Model Adjusting for Previous History of MCI

	Estimate	95% CI	P-value
Time	1.12	(1.04, 1.20)	0.002
Baseline Age	1.07	(1.05, 1.10)	<0.0001
Female (vs Male)	1.00	(0.76, 1.33)	0.99
White (vs Non-white)	1.16	(0.81, 1.68)	0.41
Education	0.97	(0.93, 1.02)	0.24
APOE4+	1.86	(1.42, 2.43)	<0.0001
MMSE	0.79	(0.72, 0.86)	<0.0001
mFSRP	1.04	(0.99, 1.08)	0.06
Complaint: ref=No Complaint			
Self Complaint Only (vs No Complaint)	1.97	(1.41, 2.76)	<0.0001
Informant Complaint Only (vs No Complaint)	2.20	(1.24, 3.93)	0.007
Both Complaints (vs No Complaint)	3.38	(2.29, 4.99)	<0.0001
Previous History of MCI	2.18	(1.44, 3.28)	0.0002

Table 4.2: Parameter Estimates for Proportional Odds Model, Stratified by Previous History of MCI

	No Previous History of MCI, N = 2750			Previous History of MCI, N = 199		
	Estimate	95% CI	P-value	Estimate	95% CI	P-value
Time	1.10	(1.03, 1.19)	0.007	1.37	(1.03, 1.81)	0.03
Age	1.07	(1.04, 1.10)	<0.0001	1.05	(0.98, 1.12)	0.20
Female (vs Male)	0.96	(0.71, 1.30)	0.79	1.38	(0.62, 3.06)	0.42
White (vs Non-white)	1.14	(0.76, 1.71)	0.51	1.47	(0.57, 3.83)	0.43
Education	0.97	(0.92, 1.02)	0.20	1.00	(0.89, 1.13)	0.91
APOE4+	2.00	(1.50, 2.67)	<0.0001	1.00	(0.46, 2.23)	0.98
MMSE	0.79	(0.71, 0.88)	<0.0001	0.74	(0.58, 0.93)	0.01
mFSRP	1.04	(0.99, 1.08)	0.12	1.11	(0.99, 1.23)	0.07
Complaint: ref=No Complaint						
Self Complaint Only	2.16	(1.51, 3.11)	<0.0001	0.89	(0.36, 2.22)	0.80
Informant Complaint Only	1.90	(0.98, 3.69)	0.057	3.94	(0.91, 17.03)	0.07
Both Complaints	3.80	(2.49, 5.82)	<0.0001	1.99	(0.73, 5.42)	0.18

4.3 Analysis Dataset

Those missing covariates at baseline or follow-up were excluded for this application, leaving a sample size of 2750 baseline normal participants with information for all methods examined. Table 4.3 gives basic demographics of participants included in the applied methods. The majority of the participants are white, female, and have more than 3 follow-up visits. The analysis sample was restricted to those with a baseline diagnosis of normal cognition with more than one follow-up visit, and excluding those with medical conditions or a history of head trauma. The baseline normal participants have an average follow-up time of 4.4 years. 30% of the participants are male, and there are statistically significant differences between all of the cognitive scores of and 85% of participants are white.

Table 4.3: Descriptive statistics for participants

Variable	N	Descriptive Statistics
Baseline Age	2750	66.00 72.00 78.00 (71.63 ± 7.93)
Sex	2750	
Male		30% $\frac{823}{2750}$
Female		70% $\frac{1927}{2750}$
Race	2750	
Non-white		15% $\frac{420}{2750}$
White		85% $\frac{2330}{2750}$
Education	2750	14.00 16.00 18.00 (15.74 ± 2.87)
Baseline mFSRP	2750	7.00 10.00 14.00 (10.75 ± 4.41)
Baseline Complaint	2750	
No complaint		77% $\frac{2121}{2750}$
Self complaint only		13% $\frac{370}{2750}$
Informant complaint only		3% $\frac{87}{2750}$
Both complaints		6% $\frac{172}{2750}$

Table 4.3: (continued)

Variable	N	Descriptive Statistics
		$N = 2750$
Baseline MMSE	2750	29.00 29.00 30.00 (29.11 ± 1.16)
Time in Study	2750	3.09 4.37 6.33 (4.73 ±1.90)
CDR Global	2750	
0		94% $\frac{2578}{2750}$
0.5		6% $\frac{172}{2750}$

4.4 Cross-sectional Model

A proportional odds model was used as the cross-sectional model for comparison. The clinic visit closest to 3-years of followup was used to determine the cross-sectional outcome of clinical diagnosis. Table 4.4 shows the odds ratios, 95% confidence intervals, and p-values for the proportional odds model. Those with an self-complaint had 2.2 times the odds of disease progression than someone with no complaint. These odds increase for those with an informant complaint (2.3) and for those with both types of complaint (4.4).

4.5 Multi-state Markov Model

A multi-state Markov model was fit to examine consecutive transitions between disease stages. For the counts of these consecutive transitions, 93% involve baseline normal participants remaining at normal cognition. 3% of consecutive observations involve a transition from normal cognition to MCI and only 0.5% involve a participant transitioning directly from normal cognition to AD.

Table 4.5 shows the hazard ratios and 95% confidence intervals for the multi-state model illustrated in Figure 2.1. All covariates included in the model are baseline

Table 4.4: Estimates from Proportional Odds Model with Follow-up of 3 Years

	Odds	95% CI	p-value
Age	1.08	(1.05, 1.11)	<0.0001
Female (vs Male)	1.12	(0.78, 1.56)	0.54
White (vs Non-white)	1.13	(0.71, 1.80)	0.62
Education	0.987	(0.93, 1.04)	0.65
APOE4+	1.801	(1.29, 2.52)	0.001
MMSE	0.78	(0.69, 0.88)	<0.0001
mFSRP	1.04	(0.99, 1.09)	0.12
Follow-up Time	1.96	(1.43, 2.70)	<0.0001
Complaint: ref=No Complaint			
Self Complaint Only	2.23	(1.47, 3.37)	0.0001
Informant Complaint Only	2.28	(1.10, 4.73)	0.03
Both complaints	4.40	(2.74, 7.07)	<0.0001

covariates in order to calculate risk prediction appropriately. See section 3.3 for details. For the transitions between NC to MCI, each level of cognitive complaint has an increased hazard of disease progression compared to the referent group of "No complaint". Those with an initial self complaint has a hazard of 1.96 progressing from NC to MCI. Those with an initial informant complaint only have a hazard of 1.72, and 3.16 for those with both types of complaint at baseline. For MCI-AD transitions, only the baseline both complaint level has statistically significant difference in the hazard of disease progression compared to no complaint. This suggests the baseline complaint for a patient is a better predictor of disease progression for those in the earlier stages of the disease than later.

Table 4.5: Estimates for Multi-state Markov Model Results for Normal - MCI and MCI - AD Transitions

	NC - MCI	MCI - AD
Baseline Age	1.08 (1.06,1.10)	1.07 (1.02,1.13)
Female (vs Male)	1.08 (0.85,1.36)	1.04 (0.59,1.83)
White (vs Non-white)	1.15 (0.84,1.57)	1.11 (0.50,2.46)
Education	0.97 (0.94,1.01)	0.99 (0.91,1.08)
APOE4+	1.95 (1.56,2.44)	2.08 (1.23,3.52)
Baseline MMSE	0.86 (0.80,0.93)	0.89 (0.74,1.07)
Baseline mFSRP	1.02 (0.99,1.06)	0.96 (0.89,1.04)
Baseline Complaint: ref=No Complaint		
Self Complaint Only	1.96 (1.49,2.58)	1.21 (0.62,2.38)
Informant Complaint Only	1.72 (1.03,2.88)	1.24 (0.42,3.63)
Both Complaints	3.16 (2.32,4.30)	2.18 (1.15,4.11)

4.6 Partly Conditional Longitudinal Model

4.6.1 Construction of Augmented Dataset

An augmented dataset was created to fit the model for $u_0 = 3$, i.e. 3-year risk prediction. For each participant, all possible paired observations with 2 to 4 years of time difference were included.

4.6.2 Partly Conditional Model Results

Covariates from the first observation within each pair are used in the model. Table 4.6 provides the odds ratios, 95% confidence intervals, and p-values for a partly conditional model based on a 3-year time period. A patient with self complaint has 2.41 the odds of disease progression after 3 years than someone with no complaint. Someone with informant complaint has 8.58 times the odds, and someone with both complaints has 8.46 times the odds, of disease progression after 3 years compared to someone with no complaint.

Table 4.6: Estimates from Partly Conditional Model

	Odds Ratio	95 % CI	p-value
Age	1.12	(1.04,1.12)	0.002
Female (vs Male)	1.37	(0.65,2.89)	0.41
White (vs Non-white)	1.93	(0.60,6.22)	0.27
Education	0.91	(0.82,1.00)	0.07
APOE4+	3.79	(1.83,7.84)	0.0003
MMSE	0.73	(0.62,0.86)	0.0002
mFSRP	1.00	(0.91,1.11)	0.96
Time Difference	1.56	(1.11,2.31)	0.013
Self Complaint Only	2.41	(0.98,5.96)	0.056
Informant Complaint Only	8.58	(3.90,18.88)	<0.0001
Both Complaints	8.46	(3.68,19.46)	<0.0001

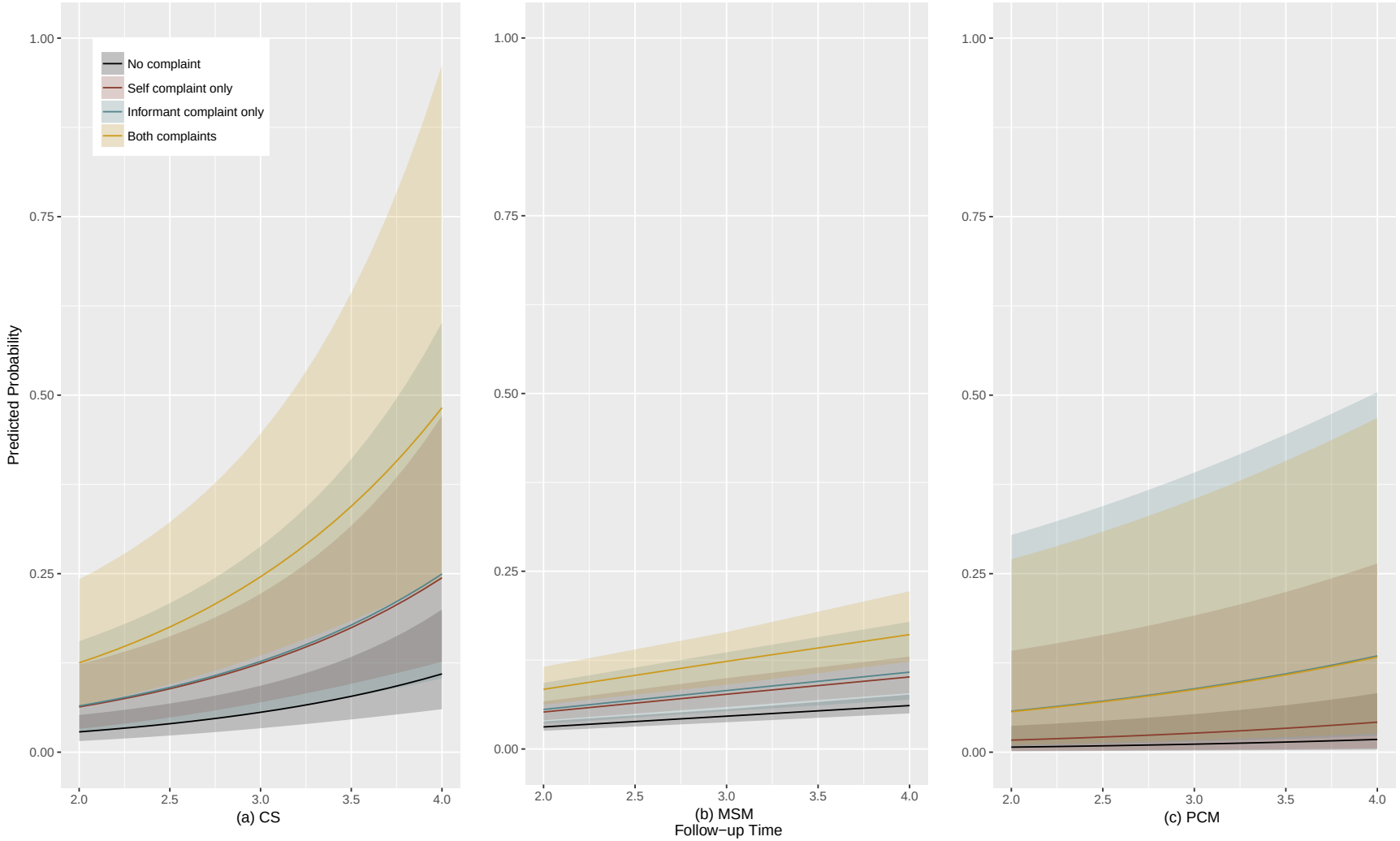
4.7 Comparison Across Methods

It is difficult to directly compare effects across methods, as the interpretation of parameters are different for the cross-sectional method, the multi-state Markov model, and the partly conditional model. Therefore, the predicted risk probabilities of disease progression are used for comparison. Figure 4.1 shows the calculated predicted probabilities of each model over a time period of 2 to 4 years for each level of cognitive complaint. Figure 4.1a shows the curved line formed by the predicted probabilities for the cross-sectional model, with both complaints at baseline having the highest probability of disease progression as expected. Figure 4.1b shows linear predicted calculations, as expected based on the calculations discussed in section 3.5 and the assumption of a constant transition matrix between time points. Both the cross-sectional and multi-state models show the similarity in predicted probability of disease progression for self complaint and informant complaint. The partly conditional model (Figure 4.1c) shows there is a closer similarity in the predicted probabilities for the informant complaint only and both complaint levels.

Each of the results of the methods show similar trends for the relationship between predicted probabilities of disease progression and complaint. The cross-sectional and multi-state methods show those with baseline self complaint only and informant complaint only are quite similar. However, the partly conditional model shows there is a closer similarity in the estimates for the predicted probability of disease progression for the informant complaint only and both complaint groups. The confidence intervals for the predicted probabilities of the cross-sectional methods, shown in Figure 4.1a, are much larger than either of the other two methods. Due to the use of baseline measurements and the reduction of information, this method is less efficient in its estimation of standard errors compared to longitudinal methods, and the standard errors are larger, as expected. Figure 4.1b displays the predicted probabilities of disease progression over time, but due to the need to exponentiate the transition probability

matrix, with follow-up time as the exponent, this was calculated only at 2, 3, and 4 years. Therefore, the predicted probabilities appear linear with time. The partly conditional model results are shown in Figure 4.1c. In Table 4.6, the partly conditional model results show larger odds ratios for complaint compared to the cross-sectional method, as shown in Table 4.4. The incorporation of the time-dependent complaint measurement in the partly conditional model results in the larger odds of disease progression for the informant complaint and both complaint groups. As shown in Table 4.3, the patient population used here is majority white, female, and college-educated, and therefore these results cannot be extended to the general public. These results are shown here for the purposes of comparing different methods for modeling disease progression.

Figure 4.1: Three-year Predicted Probabilities by Complaint Level for Cross-sectional, Multi-state, and Partly Conditional Models



CHAPTER 5

DISCUSSION

Examining the differences in statistical modeling of disease progression is important to determining how different risk factors contribute to the disease pathology. With the increasing use of electronic health records data and only sources of longitudinal patient data, comparing cross-sectional methods with differing longitudinal methods is important to support the use of methods to best answer questions about disease progression. Three general methods were applied to an AD dataset to explore the differences in methods of disease progression. Because the model estimates cannot be directly compared and the eventual goal is to build predictive models for disease progression, the predicted probabilities were calculated for each model for comparison.

Using cross-sectional models to examine transitions between disease stages, while popular and easy to perform, results in large amounts of information loss. It models the disease stage of the patient at a specific follow-up time, with an adjustment for follow-up time, giving the estimated model parameters a specific interpretation based on the follow-up time period chosen. This method does not take the disease progression within the patient into account. For example, a baseline normal participant can progress to MCI and then AD, but the cross-sectional model does not differentiate between the time at which a patient transitions between disease states. Only the disease stage of the patient at the end of their follow-up is known. Markov transition models allow for the examination of consecutive transitions, but rely on several assumptions in both the modeling process and calculating predicted risk. The Markov assumption is the strongest assumption required for the fitting of the model, assuming the entire disease history is known based on the current disease stage. The multi-state model can

directly model transitions between the disease stages and death, which is a strength of the method, especially when modeling disease progression in AD, as death is a major competing risk with the disease progression. Predictions based on this model requires the assumption the transition matrix remains constant over time, which may not be the reality of the disease. Eventually using this method for prediction may be difficult due to this restriction, and though time-dependent covariates can be used with this method, due the to requirements to calculate predicted probabilities, they cannot be used if the model is for prediction purposes. The partly conditional model directly models odds of disease progression using generalized estimation equations, which have been used in the statistical literature for several decades now, but have not yet been applied to disease progression. Unlike the multi-state method, this method does not directly model transitions to death, but there are additional ways to incorporate death information into the model, like inverse probability weighting. Partly conditional models can incorporate time-dependent covariates in the realm of prediction, which is important in AD due to the dynamic nature of its biomarkers.

The existing methods examined here rely on stratification by baseline diagnosis, with baseline normal participants modeled and baseline MCI participants excluded. Baseline MCI participants and their transition to AD are either excluded from models, or analyzed separately. This loss of information affects the coefficients and may lead to inaccurately determining the risk factors of transition [14]. Partly conditional models allows for the inclusion of baseline MCI participants in the model, which could lead to better prediction of participants at highest risk for transition, especially for the transition between MCI and AD.

Due to the nature of AD and the study population, death should be considered as a competing risk. Future work accounting for competing risk for the cross-sectional method and partly conditional methods are needed.

BIBLIOGRAPHY

[1] Bloudek, L. M., Spackman, D. E., Veenstra, D. L. and Sullivan, S. D. [2011], Cdr state transition probabilities in alzheimer's disease with and without cholinesterase inhibitor intervention in an observational cohort, Journal of Alzheimer's Disease 24(3), 599–607.

[2] Gifford, K. A., Liu, D., Lu, Z., Tripodis, Y., Cantwell, N. G., Palmisano, J., Kowall, N. and Jefferson, A. L. [2014], The source of cognitive complaints predicts diagnostic conversion differentially among nondemented older adults, Alzheimer's & Dementia 10(3), 319–327.

[3] Jack, C. R., Knopman, D. S., Jagust, W. J., Shaw, L. M., Aisen, P. S., Weiner, M. W., Petersen, R. C. and Trojanowski, J. Q. [2010], Hypothetical model of dynamic biomarkers of the alzheimer's pathological cascade, The Lancet Neurology 9(1), 119–128.

[4] Jack Jr, C. R., Therneau, T. M., Wiste, H. J., Weigand, S. D., Knopman, D. S., Lowe, V. J., Mielke, M. M., Vemuri, P., Roberts, R. O., Machulda, M. M. et al. [2016], Transition rates between amyloid and neurodegeneration biomarker states and to dementia: a population-based, longitudinal cohort study, The Lancet Neurology 15(1), 56–64.

[5] Jackson, C. H. et al. [2011], Multi-state models for panel data: the msm package for r, Journal of Statistical Software 38(8), 1–29.

[6] Kalbfleisch, J. and Lawless, J. F. [1985], The analysis of panel data under a markov assumption, Journal of the American Statistical Association 80(392), 863–871.

[7] Mandel, M., Gauthier, S. A., Guttmann, C. R. G., Weiner, H. L. and Betensky, R. A. [2007], Estimating time to event from longitudinal categorical data: an analysis of multiple sclerosis progression, Journal of the American Statistical Association 102(480), 1254–1266.

[8] Pepe, M. S. and Couper, D. [1997], Modeling partly conditional means with longitudinal data, Journal of the American Statistical Association 92(439), 991–998.

[9] Pepe, M. S., Heagerty, P. and Whitaker, R. [1999], Prediction using partly conditional time-varying coefficients regression models, Biometrics 55(3), 944–950.

[10] Sperling, R. A., Aisen, P. S., Beckett, L. A., Bennett, D. A., Craft, S., Fagan, A. M., Iwatsubo, T., Jack, C. R., Kaye, J., Montine, T. J. et al. [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), 280–292.

[11] Tyas, S. L., Salazar, J. C., Snowdon, D. A., Desrosiers, M. F., Riley, K. P., Mendiondo, M. S. and Kryscio, R. J. [2007], Transitions to mild cognitive impairments, dementia, and death: findings from the nun study, American journal of epidemiology 165(11), 1231–1238.

[12] Wei, S., Xu, L. and Kryscio, R. J. [2014], Markov transition model to dementia with death as a competing event, Computational statistics & data analysis 80, 78–88.

[13] Wisniewski, T. and Frangione, B. [1992], Apolipoprotein e: A pathological chaperone protein in patients with cerebral and systemic amyloid, Neuroscience Letters 135(2), 235 – 238.

[14] Yu, L., Tyas, S. L., Snowdon, D. A. and Kryscio, R. J. [2009], Effects of ignoring baseline on modeling transitions from intact cognition to dementia, Computational statistics & data analysis 53(9), 3334–3343.