Appendix 1: Methods

Jonathon Hirschi

Data and Processing

Terminology Note: The language "predictor" will be used, but other terms are used such as "covariate" or "independent variable".

Response Variable Construction

The response variable are a binary diagnosis of dementia and the associated year of onset. This response variable had to be constructed using hospital visits, medical records, and score on the 3Ms test. In years 1, 3, 5, 8, 10, and 11, participant medication prescriptions were recorded. This list of medications was searched for prefixes associated with dementia medications, including: "donep", "arciept", "gelanta", "razadyne", "rivastig", "exelon", "excelon", "mema", "namenda", and "namaenda". The first year that dementia medications were prescribed was recorded. Next, participant medical examinations were searched for doctor diagnoses of dementia. Again, the earliest year associated with a dementia diagnosis was recorded. Finally, participants 3Ms scores were collected over the same time period, and participants were considered to have dementia if they had a 3Ms score of less than 90. If any of these conditions were met for dementia, the earliest year associated with any of these methods was used as the onset year of dementia.

Paid Work Variable Construction

The number of hour worked was too sparse to be used. The paid work status of a participant typically changed over time, and there were missing observations in various years. For these reasons, a binary variable indicating whether an individual worked for pay at all over the 11 year time period was constructed. If an observation was missing in a given year, it is assumed that they did not work for pay. This methodology would only mislabel individuals who did not work for pay in any of the years observed, but did work for pay in the year with missing data, which is presumed to be a negligible or nonexistent proportion of the participants. In the different years of the study, this variable had different names and was located in different data sets. Therefore, pattern matching with `regex` was required. The pattern "curj" was searched for in the various yearly sub-directories.

Other Predictor Variables

The other predictor variables were all recorded in year 1 of the study and considered fixed in time. Age at year 1 of the study was the only continuous predictor. Both race and gender were coded as binary variables in the data. Education level was coded as 3 categories: less than high school education, high school graduate, and postsecondary education. Family income was divided into 4 categories: less than $10 thousand, $10 to $25 thousand, $25 to $50 thousand, and greater than $50 thousand. APOE allele type e4 is the variant associated with the increase risk of dementia. The genetic study broke down 6 categories based off allele type freuency, so this variable was converted into a binary indicator of whether the allele type contained e4 (which corresponds to classes 5 and 6 in the data). Missing observations included 166 missing values for APOE and 374 missing observations of family income level. These participants were fully removed from the analysis, resulting in 522 participants being removed due to missing data (as there was some overlap in these numbers). A critical challenge of this study was that variables were encoded with different names and stored across different data sets. Pattern matching on the variable suffixes was necessary to construct the data.

Statistical Modeling

Cox PH Model

The main assumption of the model is that hazards are proportional. This means that the ratio of the hazards is constant in time for any two individuals. This assumption can be assessed with "Kaplan-Meier" curves. The curves should have consistent shapes in time, rather than crossing or leveling out in inconsistent ways. Significance tests, which examine the proportional hazards assumption, will also be applied. Additionally, there is an assumed linear relationship between the log hazard and the predictor variables. The linearity assumption will be tested on the continuous variables using "Martingale residuals".

The set of predictor variables used was developed apriori based on the literature and client request. A model with race and paid work as fixed effects was compared to a model with an interaction term between the two predictors using a log-likelihood ratio test. The model with the interaction term resulted in a statistically significant increase in the log-likelihood (p-value of 0.041).

Mathematical Specifications

The Cox PH model fits a *hazard function* to the data. This is a function of time which corresponds to the risk of an event (in this case, a diagnosis of dementia) at a given time (in this case, the unit is years). The model specification with the given predictor variables is:

$$h(t) = h_0(t) \cdot \exp(\beta_1 \cdot \text{age} + \beta_2 \cdot \text{site} + \beta_3 \cdot \text{gender} + \beta_4 \cdot (ApoE4) + \beta_5 \cdot \text{education} + \beta_6 \cdot \text{income} + \beta_7 \cdot \text{race} + \beta_8 \cdot \text{work} + \beta_9 \cdot \text{race} \times \text{work})$$

Definitions:

- t: time in years.
- $h_0(t)$: baseline hazard, or the hazard if all categorical predictors are at the reference level and continuous variables are at the mean.
- $\beta_1, \beta_2, ...$: model coefficients
- race × work: interaction effect between race and paid work status.

The *hazard ratios* are defined as: $\exp(\beta_i)$. A value greater than 1 corresponds to an increased risk of dementia, and a value less than 1 corresponds to a decreased risk of dementia. These are estimated by holding all predictor variables constant except for the ith predictor variable. The ratio of the hazard functions with the change in the predictor variable of interest gives us the estimate for the hazard ratio.

Simplifying Assumptions

Due to the complexity of the data and the project, competing risk of other diseases and attrition due to death are not considered at this stage. If a participant died for any reason over the course of the study before receiving a dementia diagnosis, they were considered to not have gotten dementia.

Model Diagnostics

For the proportional hazards assumption, a statistical significance test based on the classic chi-squared test of residuals is presented below. Plots are presented below that investigate the same assumption.

Table 1: Chi-Square Test assessing Proportional Hazards

	Test Stat.	D.O.F.	P-Value
cv1age	11.914451	1	0.0005570
site	13.477655	1	0.0002414
gender	5.621428	1	0.0177422
apo4	3.233901	1	0.0721287
educ	56.601841	2	0.0000000
faminc	37.284055	3	0.0000000

	Test Stat.	D.O.F.	P-Value
race	39.322830	1	0.0000000
any_work	3.015683	1	0.0824627
race:any_work	15.792650	1	0.0000707
GLOBAL	139.900776	12	0.0000000

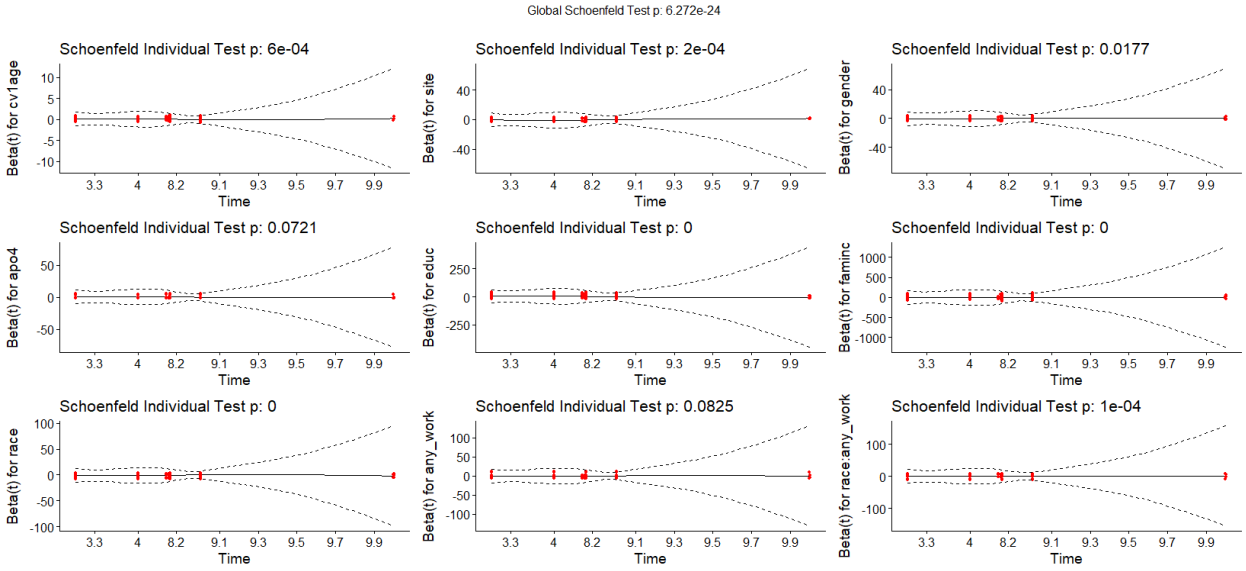

Figure 1: Proportional Hazards Assumption Plots

For the linearity assumption of the Cox PH model to hold, a plot of the Martingale residuals should be approximately linear. As can be seen in Figure XX below, this assumption is likely violated.

Code

All coding was done in R using the packages survival, survminer, dplyr, ggplot2. For the full code, visit the Github page: https://github.com/jh-206/

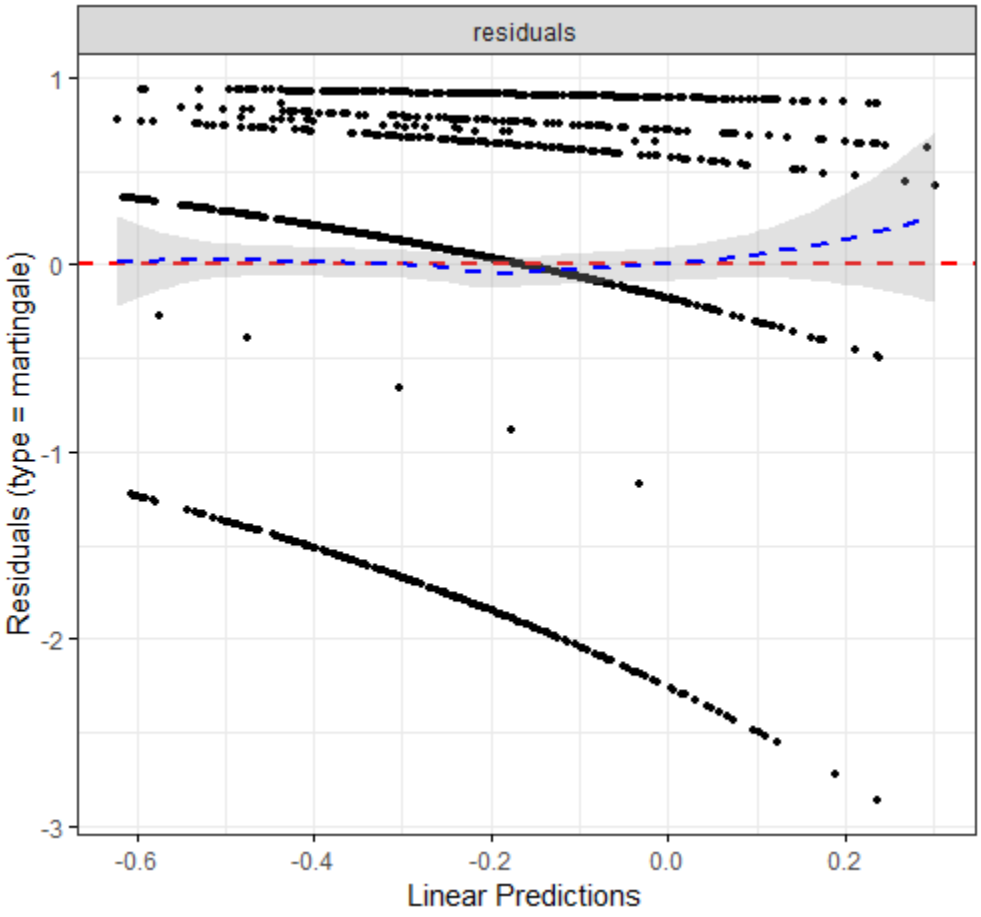

Figure 2: Residuals Plot-Linearity Assumption