

Work Status, Cognitive Function, and Dementia

Jonathon Hirschi

Statistical Consulting, Spring 2022

Introduction

Alzheimer's & related dementias are an area of increased research focus with an aging population. Various disparities in rates of dementia have been documented, including racial disparities¹. Productive work is hypothesized to reduce the rates of incident dementia. In this project, the association between having a paid job and the rates of dementia for older adults is analyzed. Additionally, the difference in this effect by race is examined. The goal of this report is to model the time until incident dementia and examine the effect of working for pay. Also, the how the effect of working for pay interacts with the race of the participant will be studied.

Study Description

The Health, Aging, and Body Composition (HABC) study followed 3,075 black and white adults over 11 years, beginning in 1997. All participants were considered "community-dwelling" and "highly functioning", indicating that they were independent and not medically institutionalized. The study was conducted in one of two sites: the areas surrounding Pittsburgh, Pennsylvania and Memphis, Tennessee. In year 1, participant demographic characteristics were collected, including race (black or white), gender (male or female) family income level and years of education completed. Additionally, a genetic study examined a protein called apolipoprotein E (APOE), known to be associated with Alzheimer's disease. This analysis of the protein assigned allele types to the participants. From years 1 to 11², participant's were examined with the Modified Mini-Mental State Exam (3Ms)³. Additionally, various questionnaires and medical examinations were carried out over those years. A survey question asking whether the participant had worked for pay in the past year was collected. Summary values for the variables used are presented below.

Analysis and Results

A Cox Proportional Hazards (PH) model was fit to the HABC data. Although the model coefficients made sense theoretically, some of the key assumptions of the Cox PH model were apparently violated⁴. Model results should therefore be interpreted with caution. In order to examine how dementia evolved over time, participants were only included in the model if they were free of dementia in year 1.

The final model coefficients are presented below in Table XX. The exponentiated negative version of the model coefficient gives the opposite interpretation of a hazard defined above: values less than 1 correspond to increased risk in incident dementia and values less than 1 correspond to decreased risk. The upper and lower values from the 95% confidence interval (CI) are included, and the associated p-values are related to whether these intervals cross 1 (i.e. no effect).

¹Yaffe 2013

²Some participants were followed longer in sub-studies which were not utilized in this analysis.

³[https://link.springer.com/referenceworkentry/10.1007/978-0-387-79948-3_530#:~:text=The%20Modified%20Mini%2DMental%20State,and%](https://link.springer.com/referenceworkentry/10.1007/978-0-387-79948-3_530#:~:text=The%20Modified%20Mini%2DMental%20State,and%20)

⁴See Appendix 1 for a more thorough discussion.

Table 1: Final Model Coefficients

	Exp. Coef. Value	Exp. Negative Coef. Value	Lower 95% C.I. Value	Upper 95% C.I. Value	P- Value
Year 1 Age	1.004	0.996	0.985	1.024	0.673
Site	0.837	1.195	0.750	0.934	0.001
Gender (Female)	1.124	0.890	1.002	1.260	0.047
APOE-e4	1.170	0.854	1.032	1.328	0.015
Education (Post-HS)	0.791	1.264	0.654	0.957	0.016
Education (HS Grad)	0.885	1.130	0.730	1.072	0.211
Income-Class 1	0.965	1.036	0.768	1.213	0.760
Income-Class 2	1.004	0.996	0.879	1.147	0.951
Income- Class 4	1.067	0.937	0.920	1.238	0.390
Race (White)	0.829	1.207	0.707	0.971	0.020
Worked for Pay (True)	0.949	1.054	0.767	1.173	0.627
Interaction (White+Worked)	1.300	0.769	1.010	1.672	0.042

Predictors⁵ with p-values less than 0.05 are considered to have relatively strong evidence that they had an effect on the rate of incident dementia⁶. Working for pay was associated with a slightly decreased risk in dementia, but this effect was not significantly different from no effect. The 95% CI is very wide, indicating substantial variability in this effect. Race was significantly different from 1⁷. The interaction effect between race and paid work was also significantly different from 1 at the 0.05 level⁸. In other words, given that an individual worked for pay, they were expected to be at 1.3 times greater risk of developing dementia if they were white than if they were black.

Women were at a higher risk of incident dementia than men⁹. The e4 variant of the APOE protein was associated with an increased risk. Participants that had either a high school education or postsecondary education were at a lower risk for developing incident dementia than participants with less than a high school education¹⁰, though the effect of postsecondary education was not significantly different from 1. None of the variables associated with family income categories were significantly different from 1. The variable associated with study site was significantly different from 1¹¹.

Discussion

Since the diagnostics of the Cox PH model were poor, all findings should be considered cautiously and further research should focus on identifying and resolving these issues. One potential resolution would be to add time-varying predictors, or consider a different survival analysis model entirely that makes a less restrictive set of assumptions. Another potential area of research would be to review influential papers in the field that utilized Cox PH models, such as Yaffe 2013, to verify if the models similarly violated assumptions.

With the reservations about model diagnostics in mind, many of the effects observed in the model matched the theoretical expectations. The scientific literature suggests that women are at a higher risk of developing

⁵ *Terminology Note:* The language “predictor” will be used, but other terms are used such as “covariate” or “independent variable”.

⁶ When interpreting a variable effect size, all other predictors should be considered as held constant.

⁷ The associated coefficient had a 95% CI of 0.71-0.97, with a mean estimate of 0.83

⁸ The associated coefficient had a 95% CI of 1.01-1.67, with a mean estimate of 0.1.30

⁹ The associated coefficient had a 95% CI of 1-1.26, with a mean estimate of 1.12

¹⁰ The associated coefficient had a 95% CI of 0.66-0.96, with a mean estimate of 0.79 for post high school education, and from 0.73-1.10 for postsecondary, with a mean estimate of 0.88. For these two coefficients associated with education, the effect is relative to the category of less than a high school education.

¹¹ The associated coefficient had a 95% CI of 0.75-0.93, with a mean estimate of 0.84.

dementia, the APOE protein e4 variant is associated with an increased risk of dementia, higher education is associated with a protective effect against dementia, and black adults are at a higher risk of developing dementia. In the statistical model, all of these effects were statistically significant and had positive or negative associations that matches these theoretical expectations. It is rare that real world data perfectly satisfy statistical assumptions. Since the coefficients discussed above satisfy theoretical expectations, it is possible that the coefficient estimates for paid work may still yield useful results.

While the effect of working for pay could not be distinguished from no effect, the interaction with race was statistically significant. Black participants who worked for pay were at a lower risk than their white counterparts. Future research would be needed to investigate whether this is due to black and white adults responding differently to paid work, or whether these differences reflect underlying differences in the type of work or the quantity of work. The confidence interval associated with the interaction effect ranged from 1.01 to 1.67, so at the low end white participants who worked for pay were 1.01 times more likely to develop dementia than black participants who worked for pay. A review of the dementia literature would be needed to determine whether these are clinically significant results.

The treatment of paid work was coarse. The data could not be used to meaningfully distinguish between participants with years of full-time paid work versus participants that only worked part time for a short duration. Further research could examine whether the quantity of work or the type of work is related to the rates of dementia. Other engaging activity such as childcare, volunteer work, or other community involvement were not accounted for, and these lifestyles are also hypothesized to have similar protective effects from dementia as paid work.

Future analysis could build on this study with the same data and general modeling framework. Attrition due to death was not accounted for in this study, and it could be accounted for in future work. Competing risk analysis could also be utilized to account for other health conditions. Observational studies such as this cannot distinguish correlation from causality.

Methods

Data and Processing

The response variables were the onset of incident 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. 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. Participants were filtered out if they were categorized as having dementia in year 1. There was sufficient data to construct the response variable in years 1, 3, 5, 8, 10, and 11 of the study.

The main predictor of interest is whether the participant worked for pay. Other predictor variables were included in an attempt to control for effects that are known to be correlated with rates of dementia. 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 thousand¹².

¹²These respectively are class 1, class 2, class 3, and class 4

Table 2: Participant Characteristics (n=1,638)

Paid Work	No	Yes
Number of Participants	1,132	506
White race	849 (75%)	335 (66%)
Rate of Incident Dementia	913 (81%)	438 (87%)
Mean (se) Onset Year of Dementia	7 (2.45)	7 (2.41)

Table 3: Other Predictor Variables (n=1,638)

Variable Name	Variable Type	Summary
Year 1 Age	Continuous	74 (2.83)
Study Site	Binary	Site 2: 825 (50%)
Gender	Binary	Female: 860 (53%)
APOE Type 4	Binary	Type 4: 379 (23%)
Education Level	Categorical	No HS: 188, HS: 900, Post HS: 550
Family Income Level	Categorical	Cat. 1: 625, Cat. 2: 126, Cat. 3: 542, Cat. 4: 345

Statistical Modeling

The model used is a Cox Proportional Hazards model, a type of statistical survival analysis. The goal of the model is to estimate how the predictor variables affect the rate of onset of dementia. The model estimates a hazard ratio for each predictor. If this ratio is greater than 1, then this predictor is associated with an increased rate in the onset of dementia. If the ratio is less than 1, there is an associated decrease in the rate of onset of dementia. The model assumes that hazards are consistent in time, which as discussed above was likely violated in this case.

Appendix I: Technical Methods

Data and Processing

The data structure was very complicated. The data containing 3Ms score and work status were stored in SAS files in years 1 through 11. Each year had many different data sets associated with it, and different variables were observed in different years. In year 1, many of the patient characteristics were recorded. Additionally, the genetics study was from a different data source. The files were in a SAS format and required the **haven** R package to load and process. Iterative procedures were used to loop through the various years and search for required variables. The variable indicating whether an individual worked for pay had the suffix “curj” which was used for pattern matching.

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 in separate data files. This list of medications was searched for prefixes associated with dementia medications, including: “donep”, “ariciept”, “gelanta”, “razadyne”, “rivastig”, “exelon”, “exelon”, “mema”, “namenda”, and “namaenda”. If any of these sequences were found using pattern matching, the year of the prescription was recorded as the year of dementia. The hospital record data had a separate variable for each study year indicating whether the participant was diagnosed with dementia in that year’s hospital visit. The earliest year of the three different methods was used as the year of onset 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 frequency, 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 Cox PH model was fit using the **survival** package in R. 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 scientific literature. 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:

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

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.
- β_1, β_2, \dots : model coefficients
- $\text{race} \times \text{work}$: 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 i th 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 4: Chi-Square Test assessing Proportional Hazards

	Test Stat.	D.O.F.	P-Value
cvlage	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
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

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/>

References

1. *Modified Mini-Mental State Examination*, Encyclopedia of Clinical Neuropsychology: https://link.springer.com/referenceworkentry/10.1007/978-0-387-79948-3_530#:~:text=The%20Modified%20Mini%2DMental%20

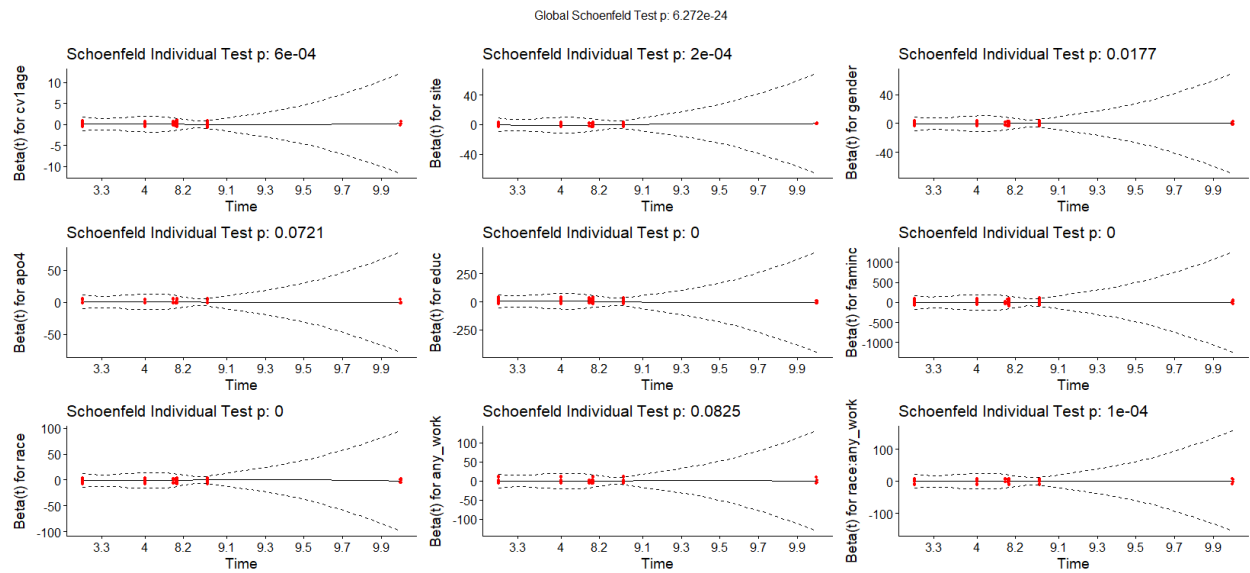


Figure 1: Proportional Hazards Assumption Plots

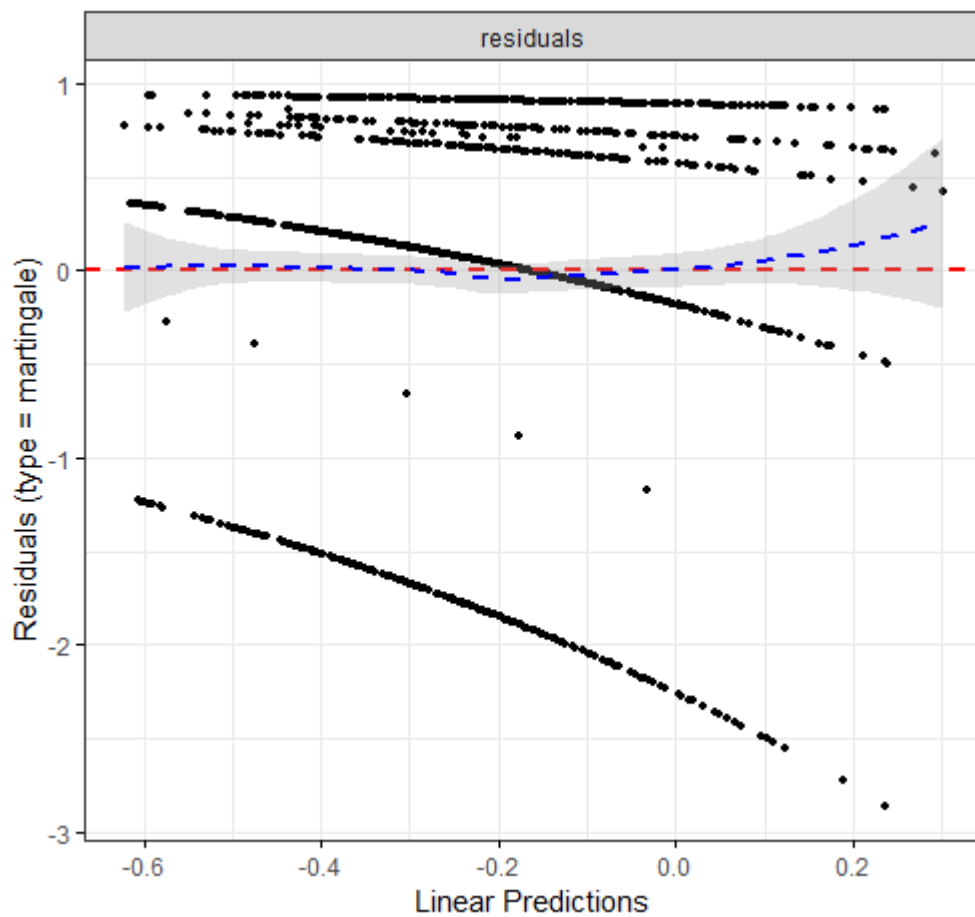


Figure 2: Residuals Plot-Linearity Assumption

2. *Effect of socioeconomic disparities on incidence of dementia among biracial older adults: prospective study*, Yaffe 2013.
3. *Cox Proportional Hazards Model*, STHDA:<http://www.sthda.com/english/wiki/cox-proportional-hazards-model>