Work Status, Cognitive Function, and Dementia

Jonathon Hirschi

Statistical Consulting, Spring 2022

Introduction

Alzheimer's and related dementias are among of the leading causes of death in the United States. As the population continues to age, there is increased focus and research on the risk factors associated with dementia. Various disparities in rates of dementia have been documented, including racial disparities.[1] Productive work is hypothesized to causally reduce the rates of incident dementia. This project analyzes the association between working for pay and the rates of incident dementia for older adults from the HABC study. Additionally, the difference in this effect by race is examined. The goals of this project are to load and format the HABC data, determine which adults developed dementia and when, and build a statistical model that estimates the relationship between working for pay and rates of incident dementia while controlling for other variables associated with the condition.

Study Description

The Health, Aging, and Body Composition (HABC) study followed 3,075 black and white adults over 11 years, beginning in 1997.[2]= All participants were considered "community-dwelling" and "highly functioning", indicating that they were physically 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 11, participants' cognitive functioning was examined with the Modified Mini-Mental State Exam[3] (3Ms). Additionally, various questionnaires and medical examinations were carried out over those years. The surveys included a question asking whether the participant had worked for pay in the past year.

Analysis and Results

A data set was constructed that combined longitudinal observations with patient characteristics observed in year 1.[4] A Cox Proportional Hazards ("PH") model was estimated with the categorization of dementia and the time until the onset as the response variables. In order to examine how dementia evolved over time, participants were only included in the modeling data set if they were free of dementia in year 1. The final model coefficients for the control variables corresponded to what was expected theoretically. However, some of the key assumptions of the Cox PH model were apparently violated.[5] Model results should therefore be interpreted with caution. Suggestions on how to remedy the model assumptions issue are presented in the Discussion section of this report.

[1] Yaffe, 2013

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

[3] Encyclopedia of Clinical Neuropsychology, 2011.

[4] See Appendix 1: Technical Methods for a more thorough discussion.

[5] See Appendix 1 for a more thorough discussion.

The final model coefficients are presented below in Table 1. The exponentiated coefficients[6] from the model are the so-called "hazard ratios". The hazard ratios can be interpreted as how many times greater an individual is at risk for dementia, while holding all other variables constant. If these ratios are larger than 1, they correspond to an increased risk in developing dementia. Correspondingly, values less than 1 are associated with a decreased risk of developing dementia. The exponentiated negative version of the model coefficient gives the opposite interpretation of a hazard: 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) and associated p-values are included.

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

Predictors[7] with p-values less than 0.05 are considered to have relatively strong evidence that they had an effect on the rate of incident dementia.[8] The p-value being less than 0.05 directly corresponds to a 95% CI that includes 1. Working for pay was associated with a slightly decreased risk in dementia, but this effect was not significantly different from no effect.[9] The 95% CI is very wide, indicating substantial variability in this effect. Race was significantly different from 1 at the 0.05 level.[10] The interaction effect between race and paid work was also significantly different from 1 at the 0.05 level.[11] 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 observed to be at a higher risk of incident dementia than men.[12] The e4 variant of the Apoe protein was also associated with an increased risk for dementia. Participants that had either a high school education or postsecondary education were at a lower risk of developing incident dementia than participants with less than a high school education,[13] but 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 1.[14]

[6] These values are calculated by taking the coefficient value β and using the formula e^{β}, where e is Euler's number.

[7] *Terminology Note:* The language "predictor" will be used, but other terms are used in the statistical literature, such as "covariate" or "independent variable".

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

[9] The associated coefficient had a 95% CI of 0.77-1.17, with a mean estimate of 0.95

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

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

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

[13] 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.

[14] The associated coefficient had a 95% CI of 0.75-0.0.93, with a mean estimate of 0.84.

Discussion

The diagnostics of the Cox PH model were poor, suggesting that the data used to estimate the model do not satisfy the model's mathematical assumptions. Therefore, all the findings should considered cautiously and further research should focus on identifying and resolving these issues. One potential resolution would be to add time-varying predictors, a more complicated form of survival analysis. Alternatively, one could consider a different survival 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 those models similarly violated assumptions.

With the reservations about model diagnostics in mind, many of the effects observed from the model matched the theoretical expectations. The scientific literature suggests that women are at a higher risk of developing dementia than men, 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. The effect of site was significant in the model. This variable is a proxy for many location specific effects such as proximity to community resources, environmental conditions, or other location-specific effects.

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 white participants in the study group. Future research would be needed to investigate whether this effect is due to black and white adults responding differently to paid work, or whether this effect reflects underlying differences in the type of work or the quantity of work that black and white participants engaged in. The confidence interval associated with the interaction effect ranged from 1.01 to 1.67. So, the most conservative estimate is that white participants who worked for pay were 1.01 times more likely to develop dementia than black participants who worked for pay. Statistical significance does not entail clinical significance, since an effect might be real but so small that it is not of great interest. A review of the dementia literature would be needed to determine whether a risk ratio of 1.01 represent a clinically significant result. It is rare that real world data perfectly satisfy statistical assumptions. Since the model effects discussed above are in line with the theoretical expectations from the broader scientific literature, it is possible that the estimated effect of paid work may still be an informative result.

The treatment of paid work in the data set was coarse. The data was sparse, and could not be used to meaningfully distinguish between participants with multiple years of full-time paid work versus participants that only worked part-time for a short duration. There was also no distinction made between the different types of paid work that a participant engaged in. Further research could examine whether the quantity of work or the type of work is related to the rates of dementia. Other mentally and socially engaging activities, such as childcare, volunteer work, or other community involvement, were not accounted for. 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 to get a better picture of rates of incident dementia. Competing risk analysis could also be utilized to account for other health conditions.

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. If dementia medications were prescribed at any point, the earliest year of prescription was used as a candidate for the year of onset of dementia. 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. Participants were considered to have dementia if they had a 3Ms score of less than 90. If any of these three conditions were met for dementia, the earliest year associated with any of these methods was used as the onset year of dementia. Participants were excluded from the data set used for modeling if they were categorized as having dementia in year 1, which accounted for 915 out of the original 3,075 adults. 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 at any point over the 11 year period. Other predictor variables were included in an attempt to control for effects that are known to be correlated with the risk of developing dementia. The other predictor variables were all recorded in year 1 of the study and are considered fixed in time. The participants' age at year 1 of the study was the only continuous predictor. All other variables were categorical. Both race and gender were recorded as binary variables in the original HABC data. Education level was coded as one of 3 categories: less than a high school education, a high school graduate education, and some 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.[15] Of the original 3,075 participants, 522 were missing some or all of the variables discussed above. For simplicity, participants with any missing observations were not included in this study.

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 (S.E.) 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

Appendix I: Technical Methods

Data and Processing

The data structure from HABC was very complicated. Each of the 3,075 participants were associated with a unique ID that was used to assemble the data set used for the statistical modeling. The data containing 3Ms score and work status were stored in SAS files in years 1 though 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 all 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 the 3Ms score and work status in each year. The variable indicating whether an individual worked for pay had the suffix "curj" which was used for pattern matching.

[15] These classes are designated class 1, class 2, class 3, and class 4, respectively

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, the medication prescriptions of the study participants were recorded in separate data files. This list of medications was searched for prefixes associated with dementia medications, including: "donep", "arciept", "gelanta", "razadyne", "rivastig", "exelon", "excelon", "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 was a single data set that 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 calculated from any of the three different methods above was used as the year of onset of dementia.

Paid Work Variable Construction

In years 1, 3, 5, 8, 9, 10, and 11, participants were asked whether they had worked for pay over the last year. In some of the study years, participants were asked how many total hours they worked for pay. The number of hours worked had too many missing values to be meaningfully used in the analysis. The paid work status of a participant typically changed over time, as many individuals left the work force at some point in their older age. There were also some 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 a 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. A survival analysis model that utilized time-varying predictors could be built in a future study using the work status variable, and then the hazard ratio associated with working for pay would more accurately reflect the quantity of work that a participant engaged in.

Other Predictor Variables

The other predictor variables were all recorded in year 1 of the study. 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 the plots seen in Figure 1.[16] The curves should be consistently flat in time, but the uncertainty grows substantially at later times and the curves therefore cannot therefore be assumed to be linear. Significance tests, which examine the proportional hazards assumption, are also displayed in these figures. The associated p-values indicate that for most of the predictor variables used, the data is not consistent with the proportional hazards assumption. 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 that included an additional variable representing the interaction effect between race and paid work. These two models were compared using a log-likelihood ratio test. The model with the interaction term resulted in a statistically significant increase in the log-likelihood .[17]

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. Whether a participant worked for pay was treated as a binary variable, and no distinction was made between individuals that worked full-versus part time.

Model Diagnostics

As the name suggests, the Cox PH model assumes that hazards are proportional. If an individual is at 2 times the risk of another individual in year 1 of the study, they should still be at 2 times the risk at any later time. 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.

[16] See Appendix 3: Numbered Figures.

[17] The associated p-value was 0.041.

Table 4: 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
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 2,[18] this assumption is likely violated. An additional assumption is that there are no outliers or extremely influential observations. Since the observed time period was restricted to 11 years, there was no possibility for an outlier that had an extremely long time until the onset of dementia.

Code

All coding was done in the R programming language using the packages `survival`, `survminer`, `dplyr`. For the full code complete with annotations, visit the Github page: https://github.com/jh-206/. In order to fully replicate the analysis, the necessary SAS files from the HABC dataset need to be saved locally.

Appendix 2: 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%2

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

Appendix 3: Numbered Figures

[18] See Appendix 3: Numbered Figures

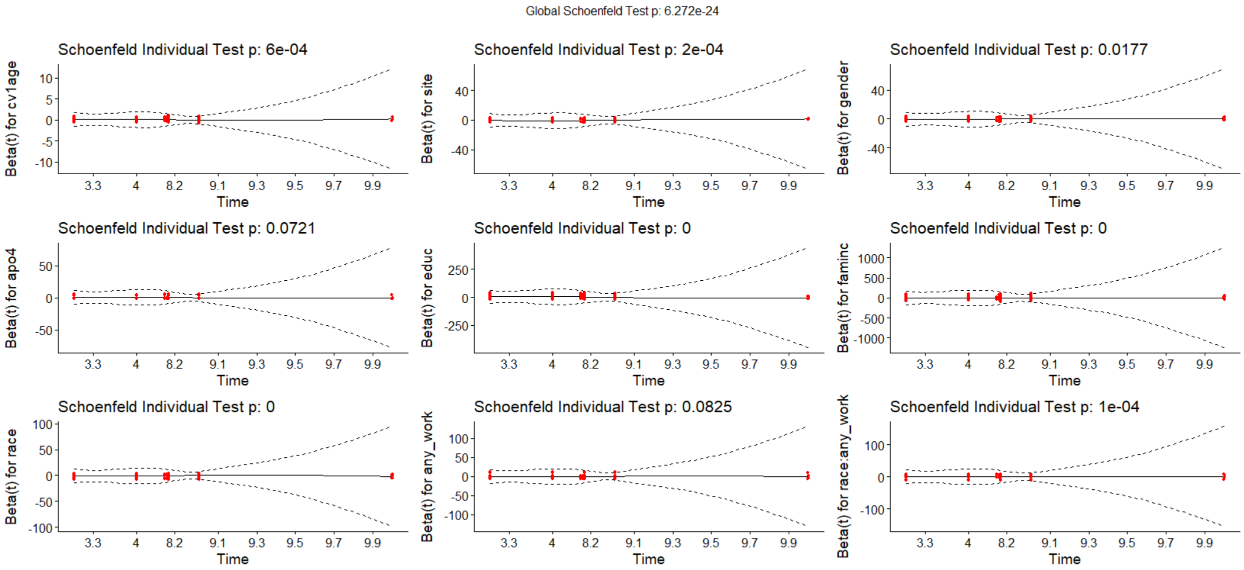

Figure 1: Proportional Hazards Assumption Plots

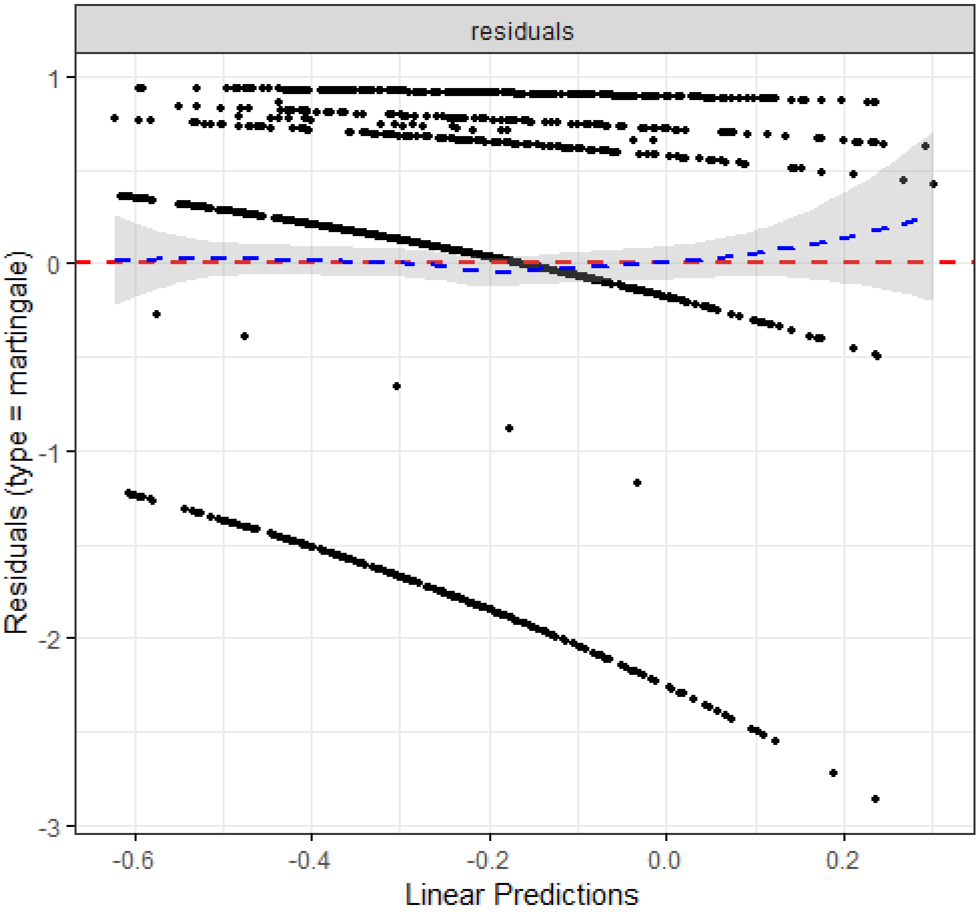

Figure 2: Residuals Plot-Linearity Assumption