P9185-Project3
Effects of DAR-0100A in individuals on cognitive deficits with schizophrenia

Yanran Li YL5465@CUMC.COLUMBIA.EDU
Department of Biostatistics
Columbia University

1 Introduction

In recent years, the study of menopause has gained considerable attention in the field of epidemiology and public health due to its significant implications on women's health and well-being. Menopause marks the end of a woman's reproductive years and is associated with various physiological changes and health risks. Understanding the factors that influence the timing of menopause can provide valuable insights into women's health and aid in the development of interventions to improve health outcomes during the menopausal transition and beyond.

This study leverages a unique dataset comprising longitudinal information on 380 women who had not undergone a hysterectomy or experienced menopause at the time of recruitment. The dataset allows for a comprehensive investigation into the timing of menopause, taking into account a range of demographic and educational factors. By employing statistical models and survival analysis techniques, we aim to explore how race, education, and age at intake contribute to variations in menopause timing. Specifically, our analysis distinguishes between the time to menopause (defined as the age at menopause minus the age at intake) and the age at menopause itself, allowing for a nuanced understanding of the factors influencing these distinct but related outcomes.

Given the complexity of factors influencing menopause, this study contributes to the existing body of knowledge by providing empirical evidence on the role of socioeconomic and demographic factors in determining menopause timing. Furthermore, by examining the survival distributions for different racial groups and the impact of education, this research offers insights into the interplay between these variables and menopause age, contributing to a more comprehensive understanding of menopausal transitions among diverse populations. Through rigorous statistical analysis, this report not only addresses fundamental questions about menopause timing but also sheds light on broader issues of health equity and access to care for women approaching menopause.

This report aimed to utilize survival analysis on menopause age and investigate the factors influencing the outcome.

2 Method

2.1 Data

The dataset analyzed in this report was survival data with event outcomes and censored data. Out of 380, there were 75 women experienced menopause during the follow-up time. Other 305 individuals were censored. Stratified by the event data and censored data, the summary statistics can be found in Table 1, where we can find the difference between menopause age and intake age between the two groups. Among the demographic information, race and education proportions vary in the event data and censored data, which indicates some associations with the survival outcomes.

2.1.1 EDA

Figure 1 visualizes the menopause occurrence against age, where each subject begins observation at varying entry times, leading to a left truncation issue as the data only captures events occurring after this entry point. Additionally, the data is subject to right censoring; for each subject, we either observe the actual age at menopause or a censored observation time if menopause has not occurred by the end of the study period. To enhance the meaningfulness of the analysis, focusing on actual age rather than time since study entry may yield results that are more readily understandable and clinically relevant.

Seeing from Figure 2, each subject commenced the study simultaneously, ensuring uniform starting points and eliminating issues with left truncation. The dataset includes both observed menopause times and instances where menopause was not observed within the study period, indicating the presence of right censoring.

2.1.2 Left Truncation

Left truncation is a statistical phenomenon that occurs when studying data from a population where not all individuals are observed from a starting point. This concept involves excluding individuals who have already experienced Menopause prior to the study's commencement, which can lead to biased estimates if not handled correctly. In this study, all the individuals recruited had no experience in menopause, so the estimates without adjusting for the left truncation are biased. To deal with it, we use the intake age to account for the left truncation time.

2.2 Estimating Survival Function

2.2.1 Parametric: Accelerated Failure Time Model

The Accelerated Failure Time (AFT) model is a parametric approach in survival analysis that examines the impact of explanatory variables on survival time. Specifically, for the i-th subject, the survival time T_i is modeled through the log-linear relationship $\log(T_i) = X_i\beta + \varepsilon_i$, where β is a coefficient vector quantifying the effects of covariates, and ε_i represents a random disturbance assumed to follow a parametric distribution. This error term characterizes the log of the unobserved baseline survival time, $\log(T_{0i})$. Alternatively, the model can be expressed in its exponential form without the logarithm as $T_i = e^{X_i\beta}T_{0i}$,

linking the survival time directly to the covariates, coefficients, and the baseline survival time.

2.2.2 Nonparametric: Kaplan-Meier Estimator

The nonparametric estimation of the survival function is based on the idea of conditional probability. The survival function $S(t_k)$ represents the probability that a subject will survive beyond a certain time point t_k. This can be calculated as the product of the probabilities of survival just beyond each observed event time up until t_k. Formally, the survival function is defined as:

$$\begin{aligned} S(t_k) &= P(T > t_k) = P(T > t_k | T \geq t_k) P(T \geq t_k) \\ &= \prod_{i=0}^{k} P(T > t_i | T \geq t_i) = \prod_{i=0}^{k} [1 - P(T = t_i | T \geq t_i)] \end{aligned} \tag{1}$$

where $P(T = t_i | T \geq t_i)$, the probability of failure at time t_i given survival up to t_i, can be estimated by the ratio $\frac{d_i}{n_i}$. In this ratio, d_i is the number of events (failures) at time t_i and n_i is the number of subjects at risk at that time. This calculation incorporates the concept of left truncation, where individuals enter the risk set at a time not earlier than their intake age.

2.2.3 Common statistic: median survival time

The Median Survival Time refers to the time point at which the survival probability is 50%. This is the time by which half of the population is expected to have experienced the event of interest. There are two primary methods for estimating the median survival time. The parametric approach involves specifying a distribution for the survival times T, and then employing maximum likelihood methods to estimate the parameters of that distribution. On the other hand, the nonparametric approach relies on empirical estimates, which do not assume any particular underlying distribution for the survival times. Both methods aim to provide an estimate for the point in time at which half the subjects have had the event occur.

2.3 Proportional Hazard Model

Proportional Hazard Model relies on several key assumptions. It is predicated on the correct specification of the underlying hazard functions, namely the hazard rate h, cumulative hazard H, and survival function S. The model posits a linear relationship between the covariates and the logarithm of the hazard rate, and it assumes that the effect of the covariates is constant over time, which is to say, time-independent. In the absence of interaction terms, the model asserts that the covariates have an additive effect on the log hazard. The focus of the model is on the hazard functions, formulated as $h_1(t|X_i) = h_0(t)e^{X_i\beta}$, where $h_0(t)$ represents the baseline hazard at time t, X_i stands for the covariates of the i-th individual, and β denotes the coefficient vector. Accounting for left truncation and presuming quasi-independence, the hazard function is expressed as $h_1(t|X_i, \text{intake\_age}) = h_1(t|X_i)$, indicating that the hazard for individuals who join the study later is equivalent to that of those who were included from the beginning.

2.4 Compare Survival Function: Log-Rank Test

The log-rank test is a non-parametric hypothesis test used to compare the survival distributions of two or more groups. It is a commonly used statistical method to evaluate differences in survival rates between treatment groups or populations exposed to different risk factors. The test compares the observed and expected number of events in each group over time, using a test statistic that follows a chi-squared distribution under the null hypothesis of equal survival distributions. The hypothesis test is as follows:

$$H_0 : S_1(t) = S_2(t) = \ldots = S_k(t)$$
$$H_1 : \text{At least one } S_i(t) \neq S_j(t)$$

where we want to compare k survival distributions.

Take 2-Group sample: Under the null hypothesis H_0, the survival functions $S_A(t)$ and $S_B(t)$ are considered equal across all time points t, whereas the alternative hypothesis H_1 suggests that the survival functions $S_A(t)$ and $S_B(t)$ are not equal for at least some time point t. The test statistic is given by:

$$Z = \frac{\sum_{i=1}^{k} \{d_{A,i} - E[d_{A,i}]\}}{\sqrt{\sum_{i=1}^{k} V(d_{A,i})}}$$

where $d_{A,i}$ is the number of events at time t_i for Group A, $E[d_{A,i}]$ is the expected number of events, and $V(d_{A,i})$ is the variance of the number of events. The test statistic Z follows a standard normal distribution under the null hypothesis.

3 Result

3.1 Menopause Time

Assuming that the distribution of menopause times for these subjects is approximately exponential, the survival curves for the median menopause time for all subjects disregarding covariates are shown in Figure 3. From Figure 3, the red line is fitted exponential curve, whose λ is 0.05680. The Black line is fitted KM curve, and both 95% CIs are shown in pictures. A Kaplan-Meier survival curve with an accompanying risk table is shown in Figure 4, where the black step function represents the estimated survival probability over time, and the shaded gray area represents the 95% confidence interval. Comparing the differences between AFT model and the KM estimator: AFT model assumes that the underlying survival times follow an exponential distribution, which tends to produce a more rigid survival curve compared to the KM estimator, which makes fewer assumptions about the underlying distribution. Due to the parametric nature of the AFT model, it allows for extrapolation of the median survival time, something that is not directly possible with the non-parametric KM estimator. When comparing confidence intervals (CIs) for survival probabilities at various time points, it is noted that the 95% CIs of the KM estimator often have significant overlaps, reflecting the variability and uncertainty in the survival estimates.

After fitting proportional hazards model as a function of race, education, and intake_age, we found: Holding all other variables constant, hazard for experiencing menopause at time t for NHB participants is approximately 2.5 times of that for NHW participants. And hazard for experiencing menopause at time t for College Graduates is approximately 0.4 times of that for Post-graduates. Details are shown in Table 2. We also conducted Schoenfeld residuals test to evaluate the proportional hazards assumption of a Cox proportional hazards regression model. When we apply this test to the Cox model fitted on the dataset, which predicts time to menopause based on race, education, and intake age, the resulting p-values for the individual tests for race and education are 0.7821 and 0.713, respectively, and 0.5039 for intake age. These p-values suggest that there is no significant evidence against the proportional hazards assumption for these covariates, as all are well above the conventional threshold of 0.05. Furthermore, the global test for the model yields a p-value of 0.9217, reinforcing the conclusion that the proportional hazards assumption holds across all covariates in the model. This ensured that the covariates have a multiplicative effect on the hazard that is constant over time, a foundational assumption of the Cox model.

3.2 Menopause Age

We conducted nonparametric estimate for the survival function $S(t) = \int_0^t f(x)\,dx$ of *menopause age* and the estimated Kaplan and Meier Curve are shown in Figure 6. The estimated median survival time is 55. 95% with CI (53.9%, 56.3%). We then fitted a proportional hazard model with race as the only explanatory variable and conducted a log-rank test. The Kaplan-Meier survival curve, stratified by race was shown in Figure 7, where each line represents the survival probability for a different racial group. The log-rank test statistic is 6.67 with p value 0.04. Therefore, we are confident to reject the null hypothesis and conclude that the survival distributions for the three race groups are not equivalent.

Adjusting for education, only NHW patients have a significant difference in hazard rate compared with the NHB patients. Seeing from Table 3, the Menopause rate among NHW patients is 0.3996 times that of NHB patients throughout the study period (adjusting for education) and this could be as little as 0.208 times or as much as 0.770 times with 95% confidence. Adjusting for education, a 95% confidence interval estimates for the relative risk of *Menopause Age* for a NHB Patient with an Other Ethnicity patient is [0.141, 1.018].

Based on the regression model for *Menopause Age* as a function of race and education, Figure 8 shows an estimate of the baseline survival function for White non-Hispanic patients with Post-graduate education.

Figure 9 shares the result of testing proportional hazards assumption in our Cox regression models. The residuals scatter for both Race & Education randomly around 0, suggesting that the proportional hazards assumption holds. Also, the p-values are above the common alpha level (e.g., 0.05), indicating no statistical evidence to reject the proportional hazards assumption for these covariates. Thus, the proportional hazards model assumptions are satisfied based on the Schoenfeld residuals test.

4 Conclusion & Discussion

This report focuses on the survival analysis for Menopause and the association between race and education. We found white non-Hispanic patients have less hazard rate compared with

other races while adjusting for education. On the other hand, the post-graduate patients have a higher hazard rate compared with other education levels while adjusting for race.

Appendix A. Figures and Tables

Table 1: Summary statistics for each variable stratified by censored and event outcomes

	Censored (N=305)	**Observed menopause** (N=75)	**Overall** (N=380)
Menopause age			
Mean (SD)	51.2 (2.43)	51.9 (2.57)	51.3 (2.47)
Median [Min, Max]	50.8 [44.7, 64.3]	51.9 [47.3, 56.7]	50.9 [44.7, 64.3]
Intake age			
Mean (SD)	47.4 (2.44)	49.7 (2.62)	47.9 (2.64)
Median [Min, Max]	46.9 [44.4, 59.6]	49.4 [45.6, 55.8]	47.2 [44.4, 59.6]
Menopause time			
Mean (SD)	3.80 (1.38)	2.15 (1.14)	3.47 (1.48)
Median [Min, Max]	4.38 [0.0164, 5.54]	1.94 [0.0329, 4.04]	4.04 [0.0164, 5.54]
Race			
NHW	248 (81.3%)	56 (74.7%)	304 (80.0%)
NHB	24 (7.9%)	13 (17.3%)	37 (9.7%)
Others	33 (10.8%)	6 (8.0%)	39 (10.3%)
Education			
Post-graduate	132 (43.3%)	35 (46.7%)	167 (43.9%)
College Graduate	81 (26.6%)	15 (20.0%)	96 (25.3%)
Some College	54 (17.7%)	16 (21.3%)	70 (18.4%)
High School Education(less)	38 (12.5%)	9 (12.0%)	47 (12.4%)

Table 2: Fitted proportional hazards model results.

Variable	n	Event	HR [CI]	p
Race	380	75		0.040
NHW			Reference	
NHB			2.463 [1.287-4.716]	0.006
Other			1.017 [0.435-2.377]	0.970
Education	380	75		0.022
Post-graduate			Reference	
College Graduate			0.412 [0.212-0.801]	0.009
Some College			1.055 [0.577-1.928]	0.862
High School Education(less)			0.555 [0.253-1.218]	0.142
Intake_age	380	75	1.364 [1.270-1.465]	<0.001

Figure 1: Menopause v.s. Age

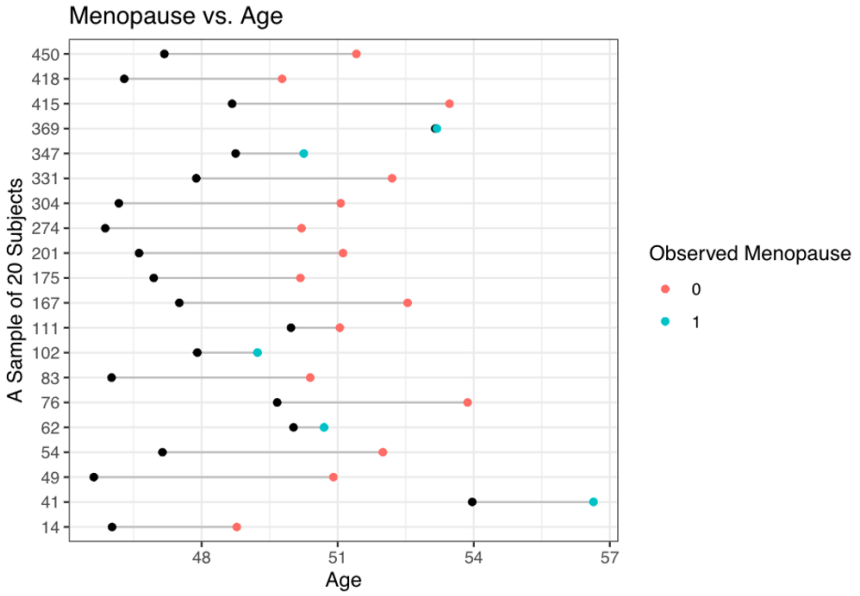

Figure 2: Menopause v.s. Duration time in Study

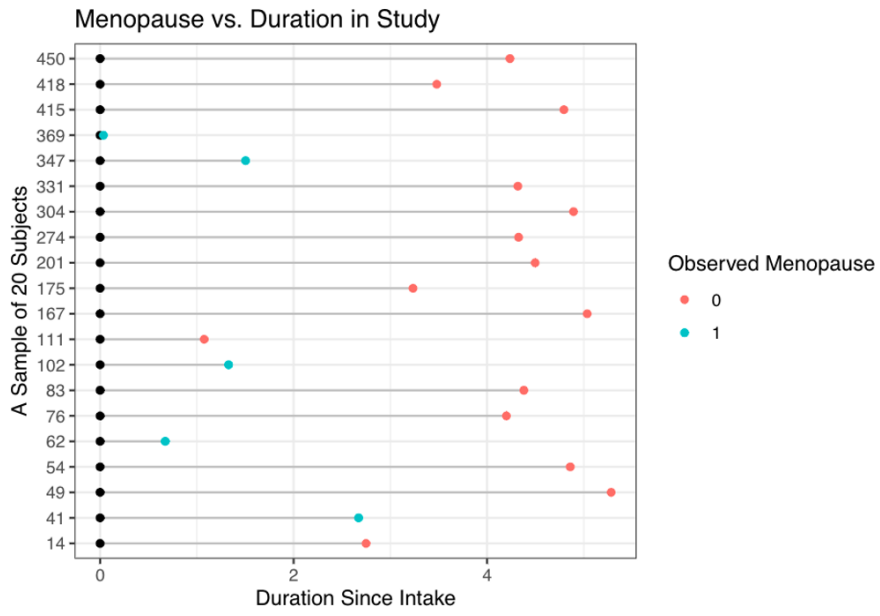

Figure 3: Estimated Survival Function for Time to Menopause

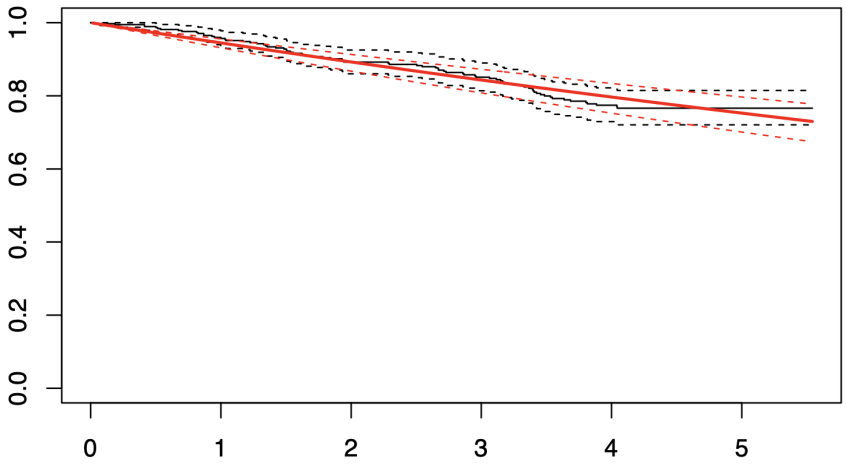

Figure 4: Kaplan-Meier Estimate for Time to Menopause

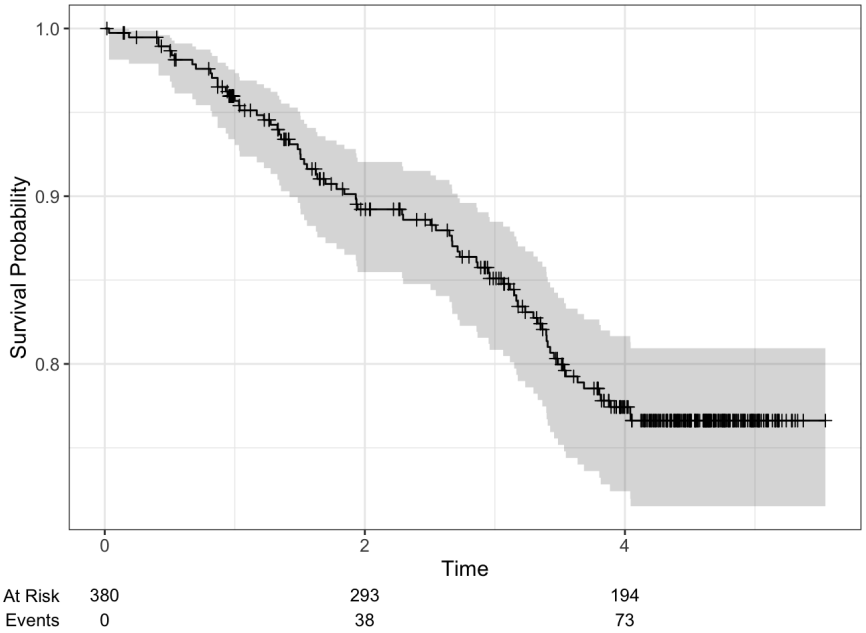

Figure 5: Proportional Hazards Assumption Check for Cox Model

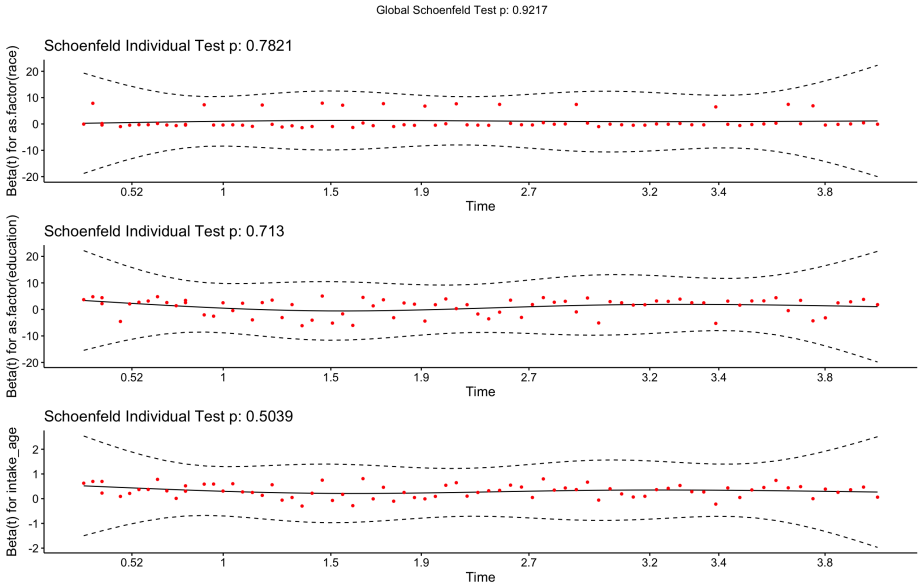

Figure 6: Estimated Kaplan and Meier Curve for Menopause Age.

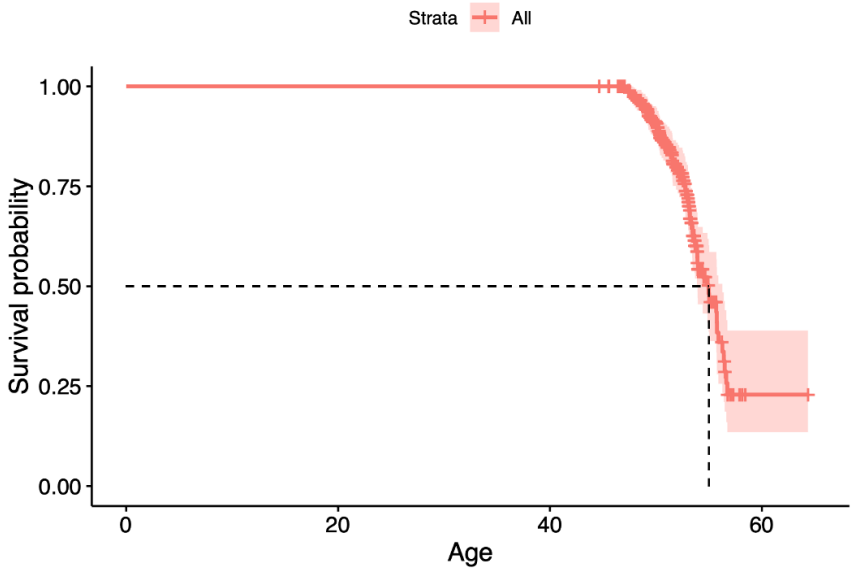

Figure 7: Estimated Kaplan and Meier Curve for .

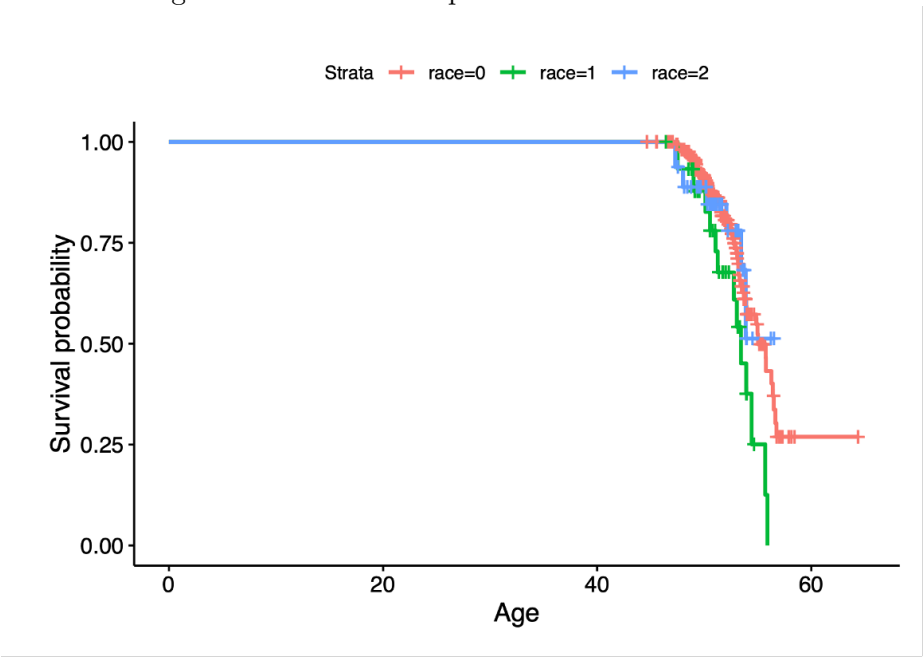

Figure 8: An estimate of the baseline survival function for NHW patients with Post-graduate education.

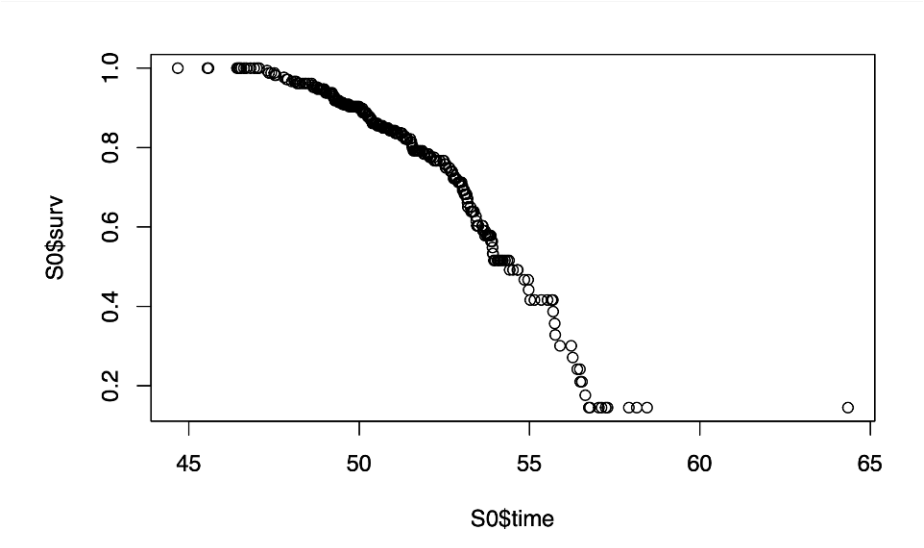

Table 3: Hazard ratios (HR) with CI and p-values, stratified by race and education.

Variable	n	Event	HR [CI]	p
Race	380	75		0.035
NHB			Reference	
NHW			0.400 [0.208-0.769]	0.006
Others			0.379 [0.141-1.018]	0.054
Education	380	75		0.086
Post-graduate			Reference	
College Graduate			0.519 [0.277-0.971]	0.040
Some College			1.003 [0.548-1.836]	0.992
High School Education(less)			0.516 [0.232-1.147]	0.104

Figure 9: Estimated Kaplan and Meier Curve for .