P9185 Statistical Practices and Research for Interdisciplinary Sciences (SPRIS)

Project III Report

A study of Menopause Time

Jungang Zou, jz3183

Department of Biostatistics,

Mailman School of Public Health, Columbia University

April 18, 2023

Contents

1 Introduction

Menopause, the cessation of a woman's reproductive years, is a natural biological process that typically occurs between the ages of 45 and 55 but can happen earlier or later. Defined as the absence of menstrual periods for 12 consecutive months, menopause can vary among women, with the average age in the United States being approximately 51 years old. The timing of menopause is influenced by various factors, including genetics. Women often experience menopause around the same age as their relatives, indicating a familial tendency. Moreover, external factors such as smoking, chemotherapy or radiation therapy, autoimmune disorders, and surgeries like hysterectomies can also impact the age at which menopause occurs. A survival study was designed to investigate when patients experienced menopause. Totally, 380 women who have not had a hysterectomy and had not experienced menopause before intake were recruited. Other demographic information such as intake age, race, and education was collected. The follow-up for individuals ends when they experienced menopause or were censored due to either drop-out or study ending.

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

2 Exploratory Data Analysis

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 Figure 1. From the summary statistics, we can find the difference between menopause age and intake age in the two groups. Among the demographic information, race and education proportions vary in the event data and censored data, which indicates some associations for the survival outcomes.

3 Methods

3.1 Notation

Consider a dataset consisting of n individuals. Let $MenopauseA_i$ denote the age when the event happened or was censored and $IntakeA_i$ as the age when the individual took part in the study. We created a new variable $MenopauseT_i = MenopauseA_i - IntakeA_i$, indicating the duration of time in the study at which the patient experienced menopause. For the survival outcome, we use $\delta_i = 1$ to denote the event that happens for i-th individual and censored outcome if $\delta_i = 0$. So the survival outcome for i-th individual is the pair $(MenopauseT_i, \delta_i)$ if we care about the $MenopauseT$. On the other hand, if we care about the absolute age when the individual has menopause, we will set the survival outcome as $(IntakeA_i, MenopauseT_i, \delta_i)$, where $IntakeA_i$ will account for the left-truncation problem. We use $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)$ to denote covariates and $S(t)$ for survival function. In survival analysis, sometimes we are concerned about the hazard function $h(t)$. The relationship between $S(t)$ and $h(t)$ is:

$$S(t) = e^{-\int_0^t h(u)du} \tag{1}$$

3.2 Left truncation

Left truncation is a statistical phenomenon that occurs when studying data from a population in which not all individuals are observed from a starting point. This concept involves excluding individuals who have already experienced the event of interest 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 $IntakeA_i$ to account for the left truncation time.

3.3 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:

$$\begin{aligned} H_0 &: S_1(t) = S_2(t) = ... = S_k(t) \\ H_1 &: \ At\ least\ one\ S_i(t) \neq S_j(t) \end{aligned} \tag{2}$$

where we want to compare k survival distributions. More details can be found in literature [1–3].

3.4 PH assumption and Cox regression

The Proportional Hazards (PH) model allows for the estimation of the hazard rate or risk of an event occurring over time. It assumes that the hazard rates for different groups being compared are proportional, meaning that the hazard rate ratio is constant over time. Cox regression, is one of the PH models, which is used to analyze survival data in medical research. It is a type of multivariate regression model that allows for the assessment of the impact of multiple predictors or risk factors on the hazard rate or risk of an event occurring over time: $h_i(t) = h_0(t)e^{\beta_1 x_{i1}+\beta_2 x_{i2}+\ldots+\beta_p x_{ip}} = h_0(t)e^{X_i\beta}$, where $h_0(t)$ is the non-parametric baseline hazard function.

The Cox regression model is based on the proportional hazards assumption, which states that the ratio of hazard rates between two groups is constant over time: $\frac{h_i(t)}{h_j(t)} = e^{X_i\beta - X_j\beta} = e^{(X_i - X_j)\beta}$. Since $(X_i - X_j)$ is given, the ratio of hazard rates is a constant value. The null hypothesis is that the hazard ratio for a covariate is constant over time. The alternative hypothesis is that the hazard ratio is not constant over time, meaning that the effect of the covariate on the hazard rate varies over time.

The survival model we used in this study is in Appendix.1.

4 Results

Stratified by *race*, we applied a log-rank test to compare the survival distributions in different race groups. The test statistics for the log-rank test is 6.67 with 2 degrees of freedom, and p-value $p = 0.04 < 0.05$. So we conclude the survival distributions among different race groups are distinct.

To make the Cox regression valid, we test the PH assumption. The results are summarized in Appendix.2. We can find the p-values for *race*, *education*, and the overall model are simultaneously larger than 0.1. So we conclude the PH assumption holds for our Cox model.

The estimates for the Cox model are in Figure 2. Adjusting for *education*, only white non-Hispanic patients have a significant difference in hazard rate compared with the black patients: adjusting for *education*, the Menopause rate among white non-Hispanic patients is 0.3996 times that of black patients throughout the study period and this could be as little as 0.7695 times or as much as 0.2075 times with 95 percent confidence. Adjusting for *race*, only the individuals whose education level is post-graduate have a significant difference in hazard rate compared with the individuals who are college graduates: adjusting for *race*, the Menopause rate among post-graduate patients is 1.9277 times that of college-graduate patients and this could be as little as 1.0295 times or as much as 3.6094 times with 95 percent confidence.

5 Conclusion

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*.

Appendices

The appendix includes all supplementary tables, formulas, and figures that are referred to in this report.

.1 Cox regression

The Cox regression in this study is as follows:

$$
\begin{aligned}
h_i(t) &= h_0(t)exp(\eta_i) \\
\eta_i &= \beta_1 \text{I(race=Other Ethnicity)}_i + \beta_2 \text{I(race=White non-Hispanic)}_i \\
&\quad + \beta_3 \text{I(education=High School Education (or less)}_i + \beta_4 \text{I(Post-graduate)}_i \\
&\quad + \beta_5 \text{I(education=Some College)}_i
\end{aligned} \tag{3}
$$

the survival outcome is the pair $(IntakeA_i, MenopauseT_i, \delta_i)$.

.2 PH assumption

PH assumption:

	chisq	df	p
race	1.869	2	0.39
education	0.924	3	0.82
GLOBAL	3.165	5	0.67

Table 1: PH assumption

.3 Figures

	Censored at menopause_age (N=305)	Observed to experience menopause (N=75)	Overall (N=380)
Menopause Age (years)			
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 (years)			
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 - Intake Age (years)			
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			
White non-Hispanic	248 (81.3%)	56 (74.7%)	304 (80.0%)
Black, non-Hispanic	24 (7.9%)	13 (17.3%)	37 (9.7%)
Other Ethnicity	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 (or less)	38 (12.5%)	9 (12.0%)	47 (12.4%)

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

```
  n= 380, number of events= 75

                                        coef exp(coef)  se(coef)      z Pr(>|z|)
raceOther Ethnicity                   -0.970952  0.378722  0.504297 -1.925  0.05418 .
raceWhite non-Hispanic                -0.917290  0.399600  0.334301 -2.744  0.00607 **
educationHigh School Education (or less) -0.005844  0.994173  0.438044 -0.013  0.98936
educationPost-graduate                 0.656311  1.927669  0.320016  2.051  0.04028 *
educationSome College                  0.659416  1.933664  0.371416  1.775  0.07583 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                                      exp(coef) exp(-coef) lower .95 upper .95
raceOther Ethnicity                      0.3787     2.6405    0.1409    1.0176
raceWhite non-Hispanic                   0.3996     2.5025    0.2075    0.7695
educationHigh School Education (or less)    0.9942     1.0059    0.4213    2.3460
educationPost-graduate                   1.9277     0.5188    1.0295    3.6094
educationSome College                    1.9337     0.5172    0.9338    4.0043

Concordance= 0.585  (se = 0.038 )
Likelihood ratio test= 11.98  on 5 df,   p=0.03
Wald test            = 12.63  on 5 df,   p=0.03
Score (logrank) test = 13.05  on 5 df,   p=0.02
```

Figure 2: Estimates for the Cox model

References

[1] Nathan Mantel et al. "Evaluation of survival data and two new rank order statistics arising in its consideration". In: *Cancer Chemother Rep* 50.3 (1966), pp. 163–170.

[2] Richard Peto and Julian Peto. "Asymptotically efficient rank invariant test procedures". In: *Journal of the Royal Statistical Society: Series A (General)* 135.2 (1972), pp. 185–198.

[3] Ross L Prentice. "Linear rank tests with right censored data". In: *Biometrika* 65.1 (1978), pp. 167–179.