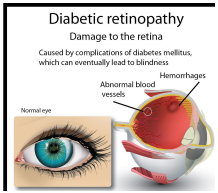
Survival Analysis for Diabetic Retinopathy

University of Calgary
Department of Mathematics and Statistics
Stat 633/533
Final Project
Adeepa Gustinna Wadu(30123735)
Rukshana Rashid Binti(30179306)
Lucy Agyei(30189582)

Outline

Introduction

- Diabetic retinopathy (also known as diabetic eye disease), is a medical condition in which damage occurs to the retina due to diabetes mellitus. It is a leading cause of vision loss and blindness in people who have diabetes.
- The International Diabetes Federation (IDF) estimated the global population with diabetes mellitus (DM) to be 463 million in 2019 and 700 million in 2045. Diabetic retinopathy remains a common complication of DM and a leading cause of preventable blindness in the adult working population.



Objective

- The aim of our study is to analyse the association between different factors of diabetic retinopathy patients using different survival analysis tools.

Dataset

- The dataset **diabetic**, partial results from a trial of laser coagulation for the treatment of diabetic retinopathy in **survival** package
- 197 patients were a 50 percent random sample of the patients with "high-risk" diabetic retinopathy
- Each patient had one eye randomized to laser treatment and the other eye received no treatment.
- For each eye, the event of interest was the time from initiation of treatment to the time when visual acuity dropped below 5/200
- Survival times in this dataset are therefore the actual time to blindness in months.
- Censoring was caused by death, dropout, or end of the study.

Dataset

Variables of the dataset

Variable Name	Description
id	Subject id
laser	Laser type: xenon or argon
age	Age at diagnosis
eye	factor with levels of left right
trt	treatment: 0 = no treatment, 1= laser
risk	risk group of 6-12
time	time to event or last follow-up
status	status of 0= censored or 1 = visual loss

Table 1: Data Set

Literature Review

- V. Sujatha has analysis simillar dataset but haven't performed veariable selection, assumption checking and model diagnostic
- A. Bora et al. performed analysis using deep learning but have't permormed using cox ph model

Kaplain-Meier Estimate

A Kaplain-Meier estimate model assumes that;

$$\hat{S}(t) = \prod_{t_j < t}(1 - d_j/Y_j)$$

where $t_i \leq t$

- t_j: time at which event occurs
- d_j: number of events at t_j
- Y_j: number of individuals alive and at risk of event before t_j

Nelson-Aalen Estimator

The Nelson-Aalen estimator is given by;

$$\hat{H}(t) = \sum_{t_j<t}(d_j/Y_j)$$

When $t_i \leq t$, and

$$\hat{H}(t) = 0$$

When $t_i > t$ The variance is estimated by

$$\sigma^2_H = \sum_{t_j<t} d_j/(Y_j)^2$$

- t_j: time at which event occurs
- d_j: number of events at t_j
- Y_j: number of individuals alive and at risk of event before t_j

Confidence Interval for Median

$100(1-\alpha)\%$ Linear Confidence Interval

$$-Z_{1-\alpha/2} \leq \frac{\hat{S}(t)-(1-p)}{\hat{Var}^{1/2}[\hat{S}(t)]} \leq Z_{1-\alpha/2}$$

$100(1-\alpha)\%$ Log-Transformed Confidence Interval

$$-Z_{1-\alpha/2} \leq \frac{[ln[-ln(\hat{S}(t))]-ln[-ln(1-p)]][\hat{S}(t)ln(\hat{S}(t))]}{\hat{Var}^{1/2}[\hat{S}(t)]} \leq Z_{1-\alpha/2}$$

$100(1-\alpha)\%$ Arc-Sin Transformed Confidence Interval

$$-Z_{1-\alpha/2} \leq \frac{2[arcsin[\hat{S}^{1/2}(t)]-arcsin[(1-p]^{1/2}][\hat{S}(t)(1-\hat{S}(t)]^{1/2}}{\hat{Var}^{1/2}[\hat{S}(t)]} \leq Z_{1-\alpha/2}$$

Two Sample Test

Hazard groups

- $h_1(t)$: hazard function for no treatment
- $h_2(t)$: hazard function for laser treatment

Hypthesis

$H_0 : h_1(t) = h_2(t)$ for all t $\leq \tau$

$H_a : h_1(t) \neq h_2(t)$ for some $\leq \tau$

Log-Rank Test

- Statistic (U_L): $U_L = \sum_{j=1}^{n_D}[D_{1j} - D_j Y_{1j}/Y_j]$
- Variance (V_L):
  $V_L = \sum_{j=1}^{n_D} D_j \frac{Y_{1j}}{Y_j} \frac{Y_{2j}}{Y_j} \frac{Y_j - D_j}{Y_j - 1}$
- Test statistics: $(Z) = U_L/\sqrt{V_L} \sim N(0,1)$

Peto-Prentice's Wilcoxon

- Statistic (U_L): $U_P W = \sum_{j=1}^{n_D} [D_{1j} - D_j W_{1j}/W_j]$
- Variance (V_L):
  $V_P W = \sum_{j=1}^{n_D} D_j \frac{W_{1j}}{W_j} \frac{W_{2j}}{Wj} \frac{W_j - D_j}{W_j - 1}$
- where $W_j = \hat{S}_(KM)$
  Test statistics: $(Z) = U_P W / \sqrt{V_P W} \sim N(0, 1)$

Cox Proportional Hazards Model

A Cox proportional hazards model assumes that;

$$h(t|z) = h_0(t)\exp(\beta^\top Z)$$

where $h_0(t)$ is the baseline hazard function, $Z = (z_1, , z_p)^\top$ is a p x 1 vector of covariates and $\beta = (\beta_1, , \beta_p)^\top$ is a p x 1 vector of regression coefficient. $\exp(z^\top\beta)$ is called the Cox multiplier or proportionality constant or relative risk.

The partial likelihood function is given as:

$$L(\beta) = \prod_{i=1}^{D} \frac{C_i^*}{\sum_{j\in\mathcal{R}} C_j}$$

Where;

$$C_i^* = exp[\sum_{k=1}^{p} \beta_k Z_{ik}]$$

COX PH Assumption

- Relationship between the log hazard and each covariate is linear
- Use of time-dependent covariates to test the adequacy of the proportional hazards assumption.
- Time dependent model

$$h(t|z) = h_0(t)\exp(\beta^\top Z + \gamma g(t))$$

- In addition, testing whether or not γ is significantly from zero allows us the opportunity to evaluate the proportional hazards assumption

$$log\frac{(h(t|z_1)}{(h(t|z_2)} = (Z_1 - Z_0)(\beta + \gamma g(t))$$

Cox-Snell Residuals

- To check the overall fit of the model, we use Cox- Snell residual, which is given by;

$$r_j = \hat{H}_0(T_j)\exp(\beta^\top Z_j)$$

- where $Z_j = (z_1, , z_p)^\top$ is a p x 1 vector of fixed-time covariates and r_j's are censored samples from the exponential distribution given that the assumed Cox model holds and the estimates of $H_0(t)$ and β are close to its true values.

- Plot NA estimates of residuals versus residual . If the Cox model provides a good fit of the data, we expect a straight line through the origin with slope 1

Martingale residual

- Used to determine the function form of a covariates
- If all covariates are time independent, with the usual RC (right-censoring), the martingale residual reduces to

$$\hat{M}_j = \delta_j - \hat{H}_0(T_j)\exp(\beta^\top Z_j)$$

- Draw the scatter plot of Mj versus the value of Z1 for the jth observation, and overlay it with a LOWESS smoothed curve
- The smoothed curve suggests f . For e.g., if the plot is linear, use Z

Model Building

- There are only fixed-time covariates in the dataset
- To identify possible confounders Used the wald test to perform local test adjusted to treatment type
- Select the variable with smallest p-value and include in to the model
- When all local tests are not significant end the selection process
- Finally Cox PH assumption and diagnostic tests were performed

Kaplan-Meier Survival Rates

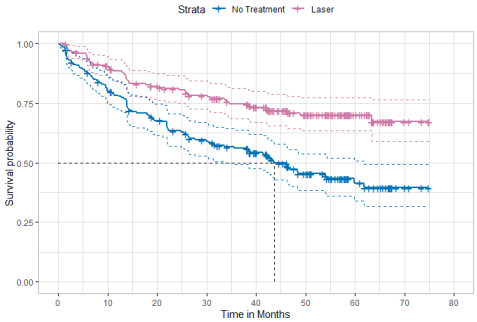


Figure 1: Kaplan-Meier Survival Curve

- survival rate of patients who had treatment "laser" is higher than patients who had no treatment.

Kaplan-Meier Survival Rates

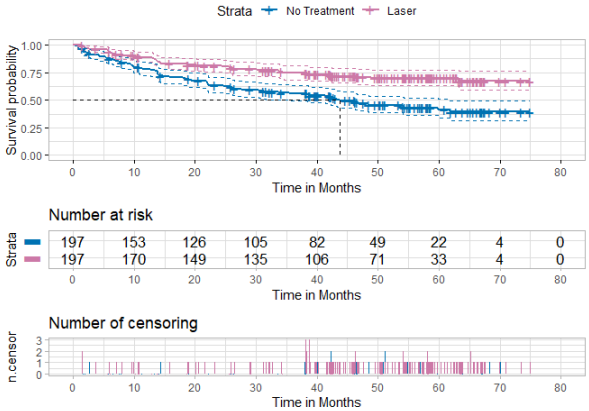


Figure 2: Survival Curve with censoing information

Nalson-Aelon Estimate

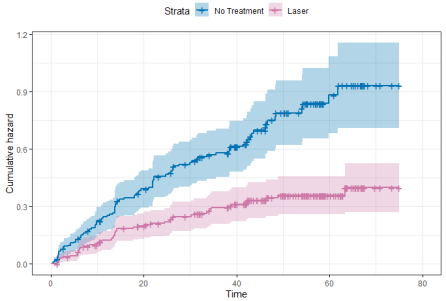


Figure 3: Cumulative Hazard Curve

- The cumulative hazard rate of patients who had no treatment is higher than the patients who had treatment "laser".

Confidence interval for Median

treatment	n	events	median	.95LCL	.95UCL
no treatment	197	101	43.7	33.6	59.8
laser	197	54	NA	NA	NA

Table 2: 95% Linear Confidence Interval for median for treatment group (no treatment, laser treatment)

treatment	n	events	median	.95LCL	.95UCL
no treatment	197	101	43.7	31.6	59.8
laser	197	54	NA	NA	NA

Table 3: 95% Log Transformed Confidence Interval for median for treatment group (no treatment, laser treatment)

Confidence interval for Median

treatment	n	events	median	.95LCL	.95UCL
no treatment	197	101	43.7	33.6	59.8
laser	197	54	NA	NA	NA

Table 4: 95% Arc-Sin Confidence Interval for median for treatment group (no treatment, laser treatment)

From the above tables, we find median for patients who had **no treatment** and for patients who had **laser** treatment we didn't find median. Also, 95% confidence intervals are found for no treatment group.

Two Sample Test

We will perform two sample test for treatment groups (no treatment, laser) using Log-rank test, Peto-Pentrice's Wilcoxon test. Here,

- $h_1(t)$: hazard function for no treatment
- $h_2(t)$: hazard function for laser treatment

Hypthesis:

$H_0 : h_1(t) = h_2(t)$ for all t $\leq \tau$

$H_a : h_1(t) \neq h_2(t)$ for some $\leq \tau$

Two Sample Test

treatment	N	Observed	Expected	$(O-E)^2/E$	$(O-E)^2/V$
trt=0 (no treatment)	197	101	71.8	11.9	22.2
trt=1 (laser)	197	54	83.2	10.3	22.2

Table 5: Log-rank test for patients who had no treatment and who had laser treatment

Chisq= 22.2 on 1 degrees of freedom, p= 2e-06

treatment	N	Observed	Expected	$(O-E)^2/E$	$(O-E)^2/V$
trt=0 (no treatment)	197	80.3	57.6	8.95	20.7
trt=1 (laser)	197	43.1	65.8	7.84	20.7

Table 6: Peto-Pentrice's Wilcoxon test for patients who had laser and no treatment

Two Sample Test

From the tables, we found that,

Log-rank Test

- test statistic=22.2
- p-value=2e-06

As p-value $< \alpha$=0.05, so we can reject H_O at $\alpha = 0.05$ level of significance. That means, the hazard rate of patients who had no treatment and who had laser are not same.

Two Sample Test

Peto-Pentrice's Wilcoxon Test

- test statistic=20.7
- p-value=6e-06

As p-value $< \alpha$=0.05, so we can reject H_O at $\alpha = 0.05$ level of significance. That means, the hazard rate of patients who had no treatment and who had laser are not same.

Log-rank test and Peto-Pentrice's Wilcoxon test both of them rejects H_O

Model Building

Variable	DF	Wald CHi Squared	P-Value	AIC
Laser	1	0.10913	0.74114	1717.7
Age	1	0.54014	0.4623759	1717.2
Eye	1	4.81285	0.0282483	1712.9
Risk	1	6.82804	0.0089738	1710.8

Table 7: Local test results adjusted to treatment type

- Risk variable was selected first to include to the model

Variable selection

Variable	DF	Wald CHi Squared	P-Value	AIC
Laser	1	0.18536	0.66681	1712.6
Age	1	0.79078	0.373865	1712.0
Eye	1	4.52085	0.033484	1708.3

Table 8: Local test results adjusted to treatment type Risk

- Eye was selected to include to the model

Variable	DF	Wald CHi Squared	P-Value	AIC
Laser	1	0.13584	0.71246	1710.1
Age	1	0.57197	0.44948	1709.7

Table 9: Local test results adjusted to treatment type, Risk Eye

Final Model

- Interactions are not significant

Variable	coef	exp(coef)	se(coef)	z	p
Treatment	-0.818	0.441	0.170	-4.8	0.000001
Risk	0.142	1.153	0.056	2.6	0.01
Eyeright	0.346	1.413	0.163	2.1	0.03

Table 10: Final Model

- Likelihood ratio test=34 on 3 df, p=2.1e-07 n= 394, number of events= 155
- Model is significant

Assumption Checking

Variable	coef	exp(coef)	se(coef)	z	p
Treatment	-0.6853	0.5040	0.4167	-1.64	0.1
$Z1_t$	-0.0364	0.9642	0.1530	-0.24	0.81

Table 11: Time dependent model to check Cox PH assumption

Test	DF	P-Value
Likelihood ratio test	2	0.00001
Wald test	2	0.00002
Logrank test	2	0.00001

- Assumption is satisfied

Model Diagnostics

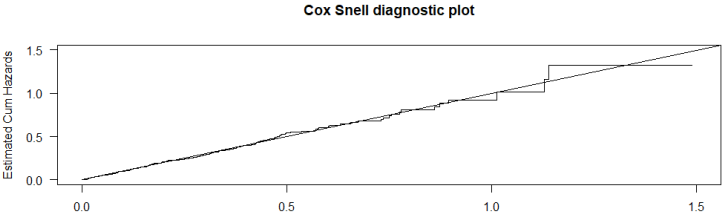


Figure 4: Cox Snell Diagnosis plot

- Model has a significant overall fit

Martingale Residual plot

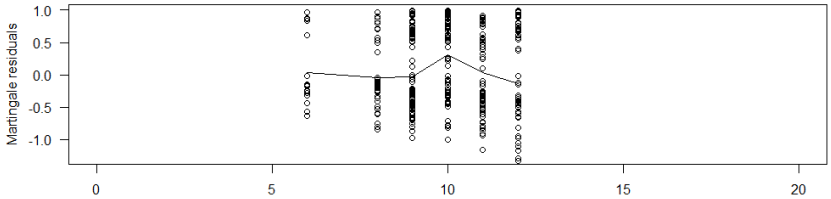


Figure 5: Martingale Residual plot

Future work

The rate of diabetic retinopathy patients are increasing alarmingly. So, by analysing more dataset related to diabetic retinopathy by fitting models like we did here, we can know the causes behind it and will be able to give more acurate prediction which may help finding better treatment for this disease.

References

- J.P. Klein and M.L. Moeschberger, Survival Analysis: Techniques for Censored and Truncated Data, 2nd Edition: Springer.
- V. Sujatha, Parametric and Non-Parametric Approaches on Survival Analysis for Diabetic Retinopathy Data, 2016
- L.T Zhen et al. Global prevalence of diabetic retinopathy and projection of burden through 2045: systematic review and Meta- Analysis
- A. Bora, S. Balasubramanian, B. Babenko et. al. ,Predicting the risk of developing diabetic retinopathy using deep learning
- S. Prinja, N. Gupta,1 and R. Verma, Censoring in Clinical Trials: Review of Survival Analysis Techniques, 2010

Thank You