BIOSTAT M215 Final Project Report

(Group 2)

Xinyang Li, Andy Liu, Yuetong Lyu

Table of Contents

1. Introduction

Scleroderma (also called systemic sclerosis) is a rare disease that involves connective-tissue disorder and hardening the skin. In some people, scleroderma is only characterized by affecting the skin. But in many other people, it can be characterized by harming internal organs that involve the lungs, heart, kidneys, and gastrointestinal tract. About 40% of patients with scleroderma have interstitial lung disease (Tashkin et al., 2006).

For this project, we are interested in finding which variables may be used as covariates in evaluating the effectiveness of oral cyclophosphamide (CYC) versus placebo in the treatment of scleroderma lung disease (ILD-SSc). In the example of the Scleroderma Lung Study, the experiment is longitudinal measured and the time to some event of interest is recorded during follow-up (Elashoff and etc. 2008). The Scleroderma Lung Study is a double-blinded and randomized clinical trial to check whether the oral cyclophosphamide (CYC) has an effect on treating lung disease due to scleroderma. The patients in this study randomly received oral CYC (<= 2 mg per kilogram of body weight per day) or placebo for one year and followed up for one additional year. The percentage of forced vital capacity (FVC) was measured every three months. A treatment failure occurs when the FVC of a patient becomes greater than 15% after three months of treatment (Tashkin et al., 2006). There were 158 participants in this study. Of the population, 145 of them completed at least 6 months of treatment. Three of the participants got a survival time zero on a scale of 0 - 24 months.

2. Methods

a. Data

This is a competing risk dataset from a scleroderma study. The study enrolled 158 patients with scleroderma-related interstitial lung disease, who were randomized to receive either CYC (2 mg/kg; **TXGROUP** = A; 79 patients) or identical-appearing placebo (**TXGROUP** = B; 79 patients) for 24 months treatment. The patients may drop out or die before completing the study, and the average number of visits per patient is 7.3. The forced vital capacity (FVC % predicted) was measured every 3 months from the baseline. **FVC0** is the baseline value of FVC, and **MAXFIB** is the baseline value of maximum lung fibrosis. The variable **surv** is the survival time, from 0 months to 24 months. The **failure_type** is the code of competing risks, of which 0 is censored, 1 is treatment failure or death, and 2 is the informative dropout. Note that both **failure_type** = 0 and 2 are censored data.Our main objective of this dataset is to measure whether oral CYC could decrease the risk of treatment failure or death.

We also made a first-glanced variable selection and created some useful dummy variables. Since except for the baseline FVC (**FVC0**), all the other FVC are time-dependent and correlated, we can only include **FVC0** as a predictor in our model. Therefore the variables **FVC3** to **FVC24** will not be selected as predictors. Also, we created dummy variables for **MAXFIB**. Since the value of **MAXFIB** is from 0 - 6, with the absence of 5, we need to create 5 dummy variables **Z1**, **Z2**, **Z3**, **Z4**, and **Z6**, which represent when **MAXFIB** is 1,2,3,4, and 6. If all these variables equal to 0, it means the **MAXFIB** is 0. The variables **d1** and **d2** are also created to indicate the failure type, which **d1** is the Treatment Failure or Death indicator, and **d2** is the informative dropout indicator. These two indicators will be used to compute the cause-specific hazard for each failure type in the part of the competing risk.

b. Statistical Analysis

The data is right-censored. We first perform a preliminary analysis of our data by estimating and visualizing the survival probabilities as time increases (***Table 1*** and ***Figure 1***). As we can see, the survival probability decreases as time increases, and at the end of 24 months, the survival probability is about 0.6. We also visualize the survival functions by Kaplan-Meier Estimation stratified by the treatment and placebo groups (***Figure 2***). From this figure, we can observe that the survival probability for each treatment group does not have a big difference. However, the survival probability for the CYC group is a little bit higher than that for the placebo group, starting from about 7 months. ***Figure 4*** and ***Figure 5*** show the estimated cumulative hazard function as time grows, measured by two approaches: Nelson-Aalen and Kaplan-Meier estimations. From ***Figure 6***, which is the comparison between the above two methods, we found that the two estimations got very similar estimated cumulative hazard functions.

To control for the variable **MAXFIB**, we want to see the survival probabilities for each value of **MAXFIB** (note that the values of baseline for maximum fibrosis are 0, 1, 2, 3, 4, 6). ***Figure 7*** shows the plots for each **MAXFIB** score. The plot for **MAXFIB** = 6 is hard to view since there is only one observation for the patient with ID 156 and the FVC score is only available at baseline. We can treat this observation as an outlier and examine the **MAXFIB** score except 6. From ***Figure 7***, we can see that for the **MAXFIB** score equals 2 and 4, the survival probability produced by Kaplan-Meier estimation decreases rapidly with the growth of time. And the survival probability at the **MAXFIB** = 0 remains the same from about 7 months. Finally, we did not detect a significant difference between the survival probabilities at other **MAXFIB** scores.

The survival model for time to failures is examined in this analysis. Before solving the competing risk problem, we treat the informative dropout as censored when building Cox's Proportional Hazard model. Though the number of variables in the dataset is small, we still employ a variable selection strategy. We employ the forward and backward selection based on the Akaike information criterion (AIC) to select important covariates contributing to survival.

A model is fit to assess the effect of the potential covariates on survival for lung disease due to scleroderma. We fit Cox's proportional hazards models using time to failures as our events. Efron's likelihood is used to handle the ties. The model for the three potential covariates are assessed independently in separate Cox regressions, and then the significant ones are combined in a final Cox's proportional hazards model.

The Cox-Snell residual plot is used to assess the appropriateness of the model. We can draw the Cox-Snell residual plot by estimating the Nelson-Aalen cumulative hazard function with the residual to be the x-axis. If the Nelson-Aalen estimator fits well with a 45-degree increasing straight line, the Cox's proportional hazard model we fitted was appropriate.

Here the data fails due to two causes (**failure_type** = 1 and 2). Therefore, it is a typical competing risks problem. In this kind of problem, we are not interested in the hazard rate, but in the probabilities, which can summarize the likelihood of the occurrence of a competing risk (Klein and etc. 2006).

3. Results

a. Variable Selection

We perform a variable selection on all 3 variables (**FVC0, TXGROUP, MAXFIB**) even though our dataset only contains a limited number of variables and the dataset is kind of small (with a sample size of 158).

The starting point of the variable selection process for this problem is to perform the global test of the hypothesis of no difference in the survival of the two treatment groups. The next step is to perform forward and backward selection on the complete model based on the Akaike information criterion (AIC). AIC examines the likelihood and the number of parameters included in the model. It attempts to balance the need for a model that fits the data very well to that of having a simple model with few parameters. AIC will decrease as variables are added to the model. When it increases, it means that the added variables are unnecessary. Therefore, as long as AIC increases at any step in this process, then we stop and base our inference about the primary hypothesis on the last model (Klein and etc. 2006).

The resulting statistics of performing the forward and backward variable selection are summarized in ***Table 2***. **TXGROUP** and **MAXFIB** are selected in the final model. And the final model resulting statistics are presented in ***Table 8***.

b. Model Results

The statistics for fitting Cox's Proportional Hazard model to all variables in the dataset are included in ***Table 3***. We still include all 3 variables in our model given we only have a small number of variables. The result shows that none of the variables are significant. The

interpretations of the estimated coefficients are the following. The exponential of the coefficient of the baseline FVC is 0.990. It means that every one percent increase in the baseline FVC will result in a 1 percent decrease in the hazard rate, controlling for the other variables. Similarly, the exponential of the coefficient of the **TXGROUP** of 0.518 is the risk of failure if the individual received the treatment relative to the risk of failure should the individual have received the placebo. The exponentials of the coefficients of **Z1**, **Z2**, **Z3**, and **Z4** are 0.740, 2.008, 0.603, 1.540 respectively. It means that the risk of failures for individuals who have 1 and 3 baseline maximum fibrosis is lower than the risk for individuals who have 0 baseline maximum fibrosis.

The Cox-Snell Residual plot is presented in ***Figure 3***. As we recall, if Cox's Proportional Hazard model fits the data, then the plot should follow a 45-degree line. However, our plot does not quite follow the 45-degree line. This might be a result of the small sample size and the lack of predictors.

The competing risks were assessed by two methods. One is through the Cause-Specific Cumulative Hazard, and another is through the Cumulative Incidence Functions (CIF). For the first method, ***Table 4*** and ***Table 5*** show the summary for fitting survival models based on variables **d1** and **d2**. From ***Table 4***, we can see that there is one type 1 failure (**d1** = 1 for encountering a treatment failure or death, and **d1** = 0 for not) at time 4 (months). Similarly, there is one type 1 failure at time 6, and four type 1 failure at time 7, and so on. From Table 5, we can see that 8 patients encountered a type 2 failure (**d2** = 1 for encountering an informative dropout, and **d2** = 0 for not) at time 12, and there are 14 patients with a type 2 failure at time 24. Besides, the two tables contain the survival probabilities, standard errors, and 95% confidence intervals at different time points on the curves.

For the second method, we performed a CIF analysis. We are interested in finding summary curves that can tell us how the likelihood of failure events change over time in months. The CIF can be computed directly from the joint density function of the potential failure times or from the cause-specific hazard rates (Klein and etc. 2006). In R, we can fit models for the estimated CIFs and then get the estimated probability at a specific time using the timepoints function in the "cmprsk" package. ***Figures 8*** and ***9*** show the probability curves for failure events 1 and 2 throughout the experiment (24 months) comparing to the net probability. From the two plots, we can see that the probability curves for two competing risks fit well with their net probabilities, however, the failure event of informative dropout fits better. Finally, ***Figure 10*** shows the cumulative incidence plots for the two competing risks and the cumulative incidence plot for the sum of the two competing risks. Also, the distance between y-axis 1 and the solid black line represents the disease-free survival probability. From this graph, we can conclude that the failure event of informative dropout (**failure_type** = 2) has a higher probability than the other failure event (treatment failure or death).

We also fitted the competing risks regression model using the crr function in the "cmprsk" package in R. This model is different from Cox's Proportional Hazard model we fitted before since the coefficients in this model are the effects of the covariate on the sub-distribution of a specific failure type. The results, which contain the estimated regression coefficients, standard errors, and the two-sided p-values for each coefficient, are presented in ***Table 7***. Here, at failure code 1, the estimated value for **FVC0** coefficient is -0.004197, meaning that the effect of the covariate on the sub-distribution of the first competing risk is -0.004197. Similar explanations may apply to all other coefficients. By comparing the estimated parameters of

coefficients for the two competing risks, we conclude that the **FVC0** does not differ too much by the two risks, but the coefficients **TXGROUP** and **MAXFIB** do differ in the two failure types.

4. Conclusion

In this project, we conduct variable selection based on Cox's Proportional Hazard model and we handle the informative censoring by treating the informative censored event as a competing risk. From the above analysis, we conclude that oral CYC has a significant effect on decreasing the risk of treatment failure or death for lung disease due to scleroderma.

There are some limitations in our project given the small sample size. The Cox-Snell Residual plot raises concerns that Cox's Proportional Hazard model may not be a good fit for the data. In addition, we do not have enough data about individuals with 6 baseline maximum lung fibrosis given we only have one individual with **MAXFIB** equals 6 in our dataset. These issues may be addressed by further study.

5. References

(1) Tashkin DP, Elashoff RM, Clements PJ, et al. Cyclophosphamide versus placebo in scleroderma lung disease. The New England Journal ofMedicine 2006;354:2655–2666.

(2) Balbir-Gurman A, Yigla M, Guralnik L, Hardak E, Solomonov A, Rozin AP, Toledano K, Dagan A, Bishara R, Markovits D, Nahir MA, Braun-Moscovici Y. Long-term follow-up of patients with scleroderma interstitial lung disease treated with intravenous cyclophosphamide pulse therapy: a single-center experience. Isr Med Assoc J. 2015 Mar;17(3):150-6. PMID: 25946765.

(3) Klein, John P., and Melvin L. Moeschberger. *Survival analysis: techniques for censored and truncated data*. Springer Science & Business Media, 2006

6. Appendix

a. Tables

Table 1: Estimated Survival Probabilities

```
           coef exp(coef) se(coef)      z      p
TXGROUP -0.6025    0.5474   0.5405 -1.115 0.2649
Z2       0.9609    2.6140   0.5309  1.810 0.0703

Likelihood ratio test=3.71  on 2 df, p=0.1565
n= 153, number of events= 15
```

Table 2: Results of the Forward and Backward Variable Selection

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
4	149	1	0.993	0.00669	0.980	1.000
6	147	1	0.987	0.00946	0.968	1.000
7	145	4	0.959	0.01627	0.928	0.992
9	140	3	0.939	0.01978	0.901	0.978
10	136	1	0.932	0.02081	0.892	0.974
12	132	2	0.918	0.02276	0.874	0.963
16	117	2	0.902	0.02493	0.854	0.952
17	115	1	0.894	0.02592	0.845	0.946
20	110	1	0.886	0.02692	0.835	0.940

Table 3: Results of Cox's Proportional Hazard model

```
  n= 153, number of events= 15

            coef exp(coef) se(coef)      z
FVC0    -0.01017   0.98988  0.02285 -0.445
TXGROUP -0.65733   0.51823  0.54314 -1.210
Z1      -0.30066   0.74033  1.15952 -0.259
Z2       0.69695   2.00761  1.08220  0.644
Z3      -0.50566   0.60311  1.17513 -0.430
Z4       0.43151   1.53959  1.43133  0.301
        Pr(>|z|)
FVC0       0.656
TXGROUP    0.226
Z1         0.795
Z2         0.520
Z3         0.667
Z4         0.763

        exp(coef) exp(-coef) lower .95 upper .95
FVC0       0.9899     1.0102   0.94652     1.035
TXGROUP    0.5182     1.9296   0.17873     1.503
Z1         0.7403     1.3508   0.07628     7.185
Z2         2.0076     0.4981   0.24072    16.744
Z3         0.6031     1.6581   0.06027     6.035
Z4         1.5396     0.6495   0.09312    25.454

Concordance= 0.673  (se = 0.063 )
Likelihood ratio test= 4.49  on 6 df,   p=0.6
Wald test            = 4.5  on 6 df,   p=0.6
Score (logrank) test = 4.69  on 6 df,   p=0.6
```

Table 4: Summary of Intercept-only Survival Model on **d1**

```
> summary(fit.death)
Call: survfit(formula = Surv(surv, d1) ~ 1, data = mydata)

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    4    149       1    0.993 0.00669        0.980        1.000
    6    147       1    0.987 0.00946        0.968        1.000
    7    145       4    0.959 0.01627        0.928        0.992
    9    140       3    0.939 0.01978        0.901        0.978
   10    136       1    0.932 0.02081        0.892        0.974
   12    132       2    0.918 0.02276        0.874        0.963
   16    117       2    0.902 0.02493        0.854        0.952
   17    115       1    0.894 0.02592        0.845        0.946
   20    110       1    0.886 0.02692        0.835        0.940
```

Table 5: Summary of Intercept-only Survival Model on **d2**

```
> summary(fit.dropout)
Call: survfit(formula = Surv(surv, d2) ~ 1, data = mydata)

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    1    155       1    0.994 0.00643        0.981        1.000
    3    152       2    0.980 0.01116        0.959        1.000
    5    148       1    0.974 0.01290        0.949        0.999
    6    147       1    0.967 0.01442        0.939        0.996
    8    141       1    0.960 0.01586        0.930        0.992
    9    140       1    0.954 0.01717        0.920        0.988
   10    136       1    0.946 0.01842        0.911        0.983
   11    134       2    0.932 0.02068        0.893        0.974
   12    132       8    0.876 0.02743        0.824        0.931
   15    119       2    0.861 0.02887        0.806        0.920
   18    114       2    0.846 0.03028        0.789        0.908
   21    109       1    0.838 0.03098        0.780        0.901
   24    107      14    0.729 0.03836        0.657        0.808
```

Table 6: Survival Differences Testing Results

```
Call:
survdiff(formula = Surv(surv, censored) ~ MAXFIB, data = mydata[!is.na(mydata$MAXFIB),
    ], subset = (TXGROUP == "A"))

          N Observed Expected (O-E)^2/E (O-E)^2/V
MAXFIB=0  5        3     1.82     0.767     0.889
MAXFIB=1 18        4     6.32     0.850     1.196
MAXFIB=2 28       11     9.73     0.165     0.277
MAXFIB=3 19        6     7.31     0.236     0.349
MAXFIB=4  5        3     1.82     0.769     0.883

 Chisq= 3  on 4 degrees of freedom, p= 0.6
Call:
survdiff(formula = Surv(surv, censored) ~ MAXFIB, data = mydata[!is.na(mydata$MAXFIB),
    ], subset = (TXGROUP == "B"))

          N Observed Expected (O-E)^2/E (O-E)^2/V
MAXFIB=0  6        1   2.1372     0.605     0.684
MAXFIB=1 22        5   8.2713     1.294     1.999
MAXFIB=2 19        7   4.4672     1.436     1.806
MAXFIB=3 27        8   9.3947     0.207     0.342
MAXFIB=4  3        3   0.7163     7.280     7.706
MAXFIB=6  1        1   0.0133    73.013    74.000

 Chisq= 85.3  on 5 degrees of freedom, p= <2e-16
Call:
survdiff(formula = Surv(surv, censored) ~ MAXFIB + strata(TXGROUP),
    data = mydata[!is.na(mydata$MAXFIB), ])

          N Observed Expected (O-E)^2/E (O-E)^2/V
MAXFIB=0 11        4   3.9559  4.93e-04  5.63e-04
MAXFIB=1 40        9  14.5885  2.14e+00  3.17e+00
MAXFIB=2 47       18  14.2004  1.02e+00  1.55e+00
MAXFIB=3 46       14  16.7080  4.39e-01  6.90e-01
MAXFIB=4  8        6   2.5340  4.74e+00  5.32e+00
MAXFIB=6  1        1   0.0133  7.30e+01  7.40e+01

 Chisq= 83.1  on 5 degrees of freedom, p= <2e-16
```

Table 7: Competing Risks Regression Results for Failure code 1 and 2

```
convergence:  TRUE
coefficients:
     FVC0   TXGROUP    MAXFIB
-0.004197 -0.399900 -0.046250
standard errors:
[1] 0.02055 0.52220 0.19960
two-sided p-values:
   FVC0 TXGROUP  MAXFIB
   0.84    0.44    0.82

convergence:  TRUE
coefficients:
     FVC0  TXGROUP   MAXFIB
-0.00389  0.23000  0.35030
standard errors:
[1] 0.01648 0.33230 0.20440
two-sided p-values:
   FVC0 TXGROUP  MAXFIB
  0.810   0.490   0.087
```

Table 8: The Final Table by Variable Selection

```
  n= 153, number of events= 15

           coef exp(coef) se(coef)      z Pr(>|z|)
TXGROUP -0.6430    0.5257   0.5434 -1.183    0.237
Z1      -0.2491    0.7795   1.1548 -0.216    0.829
Z2       0.7489    2.1147   1.0762  0.696    0.487
Z3      -0.4034    0.6680   1.1550 -0.349    0.727
Z4       0.5208    1.6834   1.4221  0.366    0.714

        exp(coef) exp(-coef) lower .95 upper .95
TXGROUP    0.5257     1.9023   0.18121     1.525
Z1         0.7795     1.2828   0.08107     7.495
Z2         2.1147     0.4729   0.25654    17.431
Z3         0.6680     1.4969   0.06945     6.426
Z4         1.6834     0.5940   0.10368    27.333

Concordance= 0.663  (se = 0.062 )
Likelihood ratio test= 4.25  on 5 df,   p=0.5
Wald test            = 4.26  on 5 df,   p=0.5
Score (logrank) test = 4.43  on 5 df,   p=0.5
```

b. Figures

Figure 1: Estimated Survival Probability vs. Time

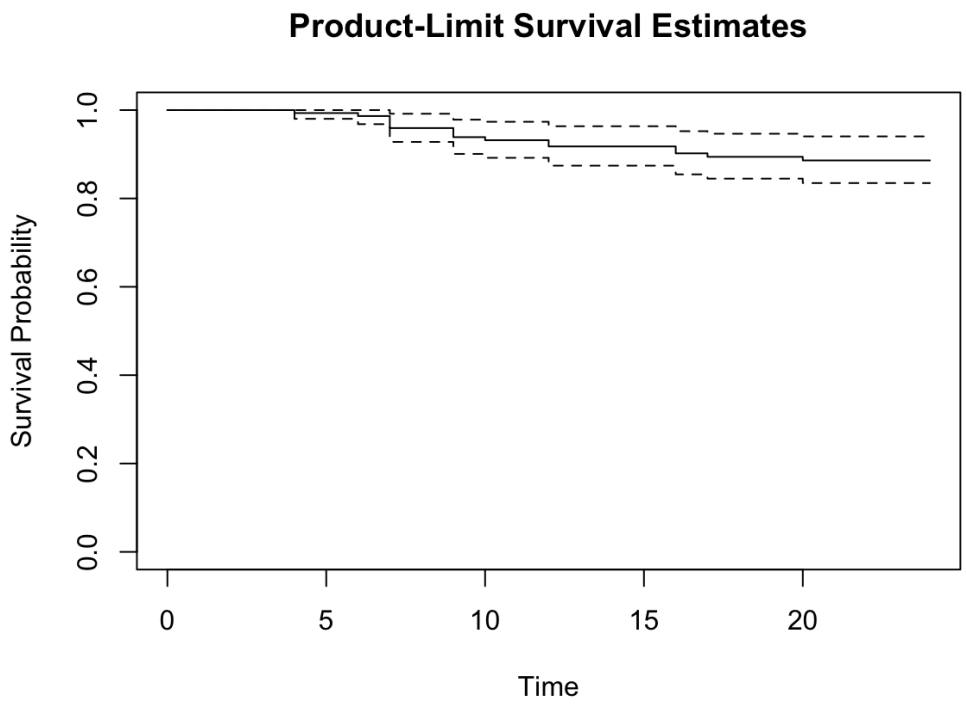

Figure 2: Estimated Survival Function vs. Time stratified by Treatment Group

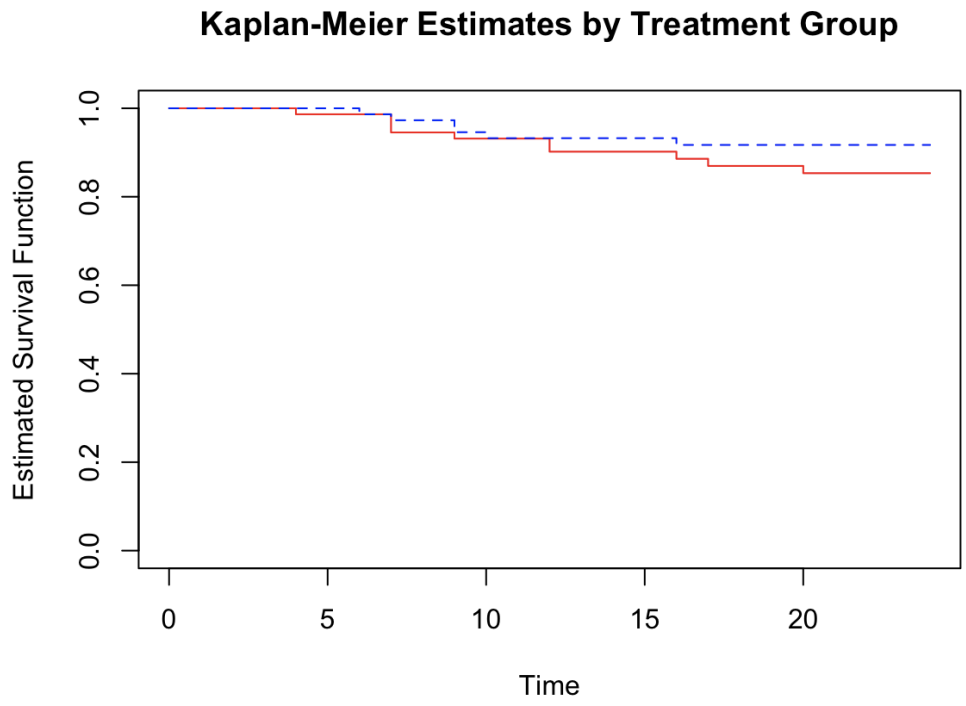

Figure 3: Cox-Snell Residual Plot

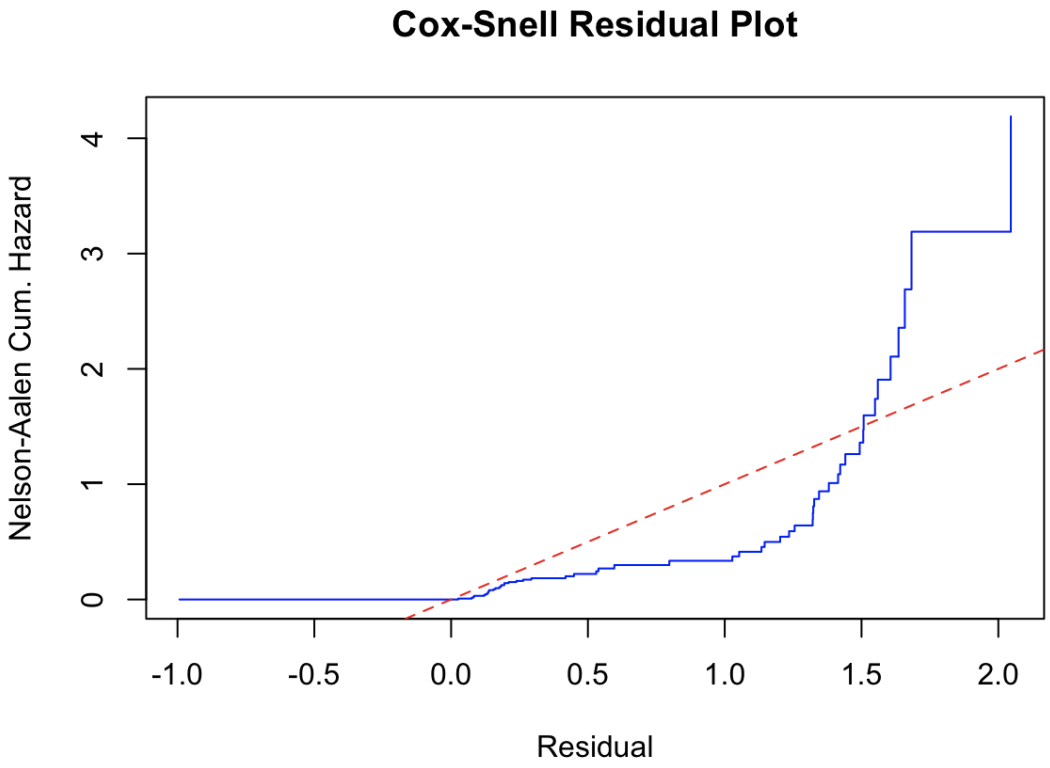

Figure 4: Estimated Cumulative Hazard Function Using Kaplan-Meier Estimation

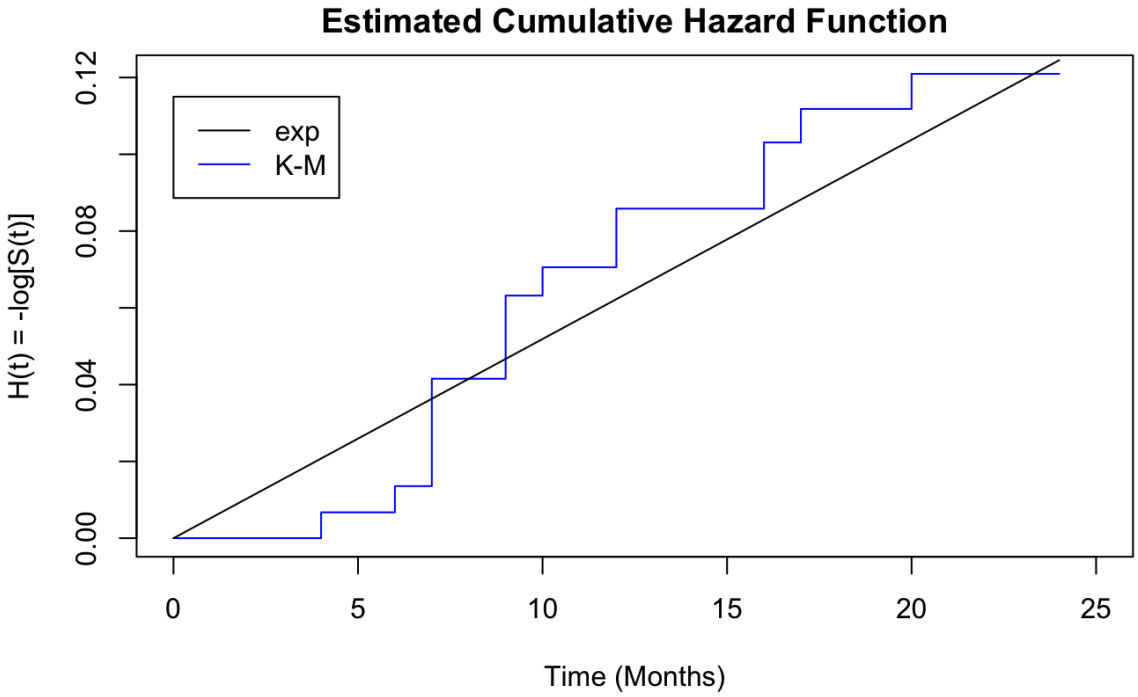

Figure 5: Estimated Cumulative Hazard Function Using Nelson-Aalen Estimation

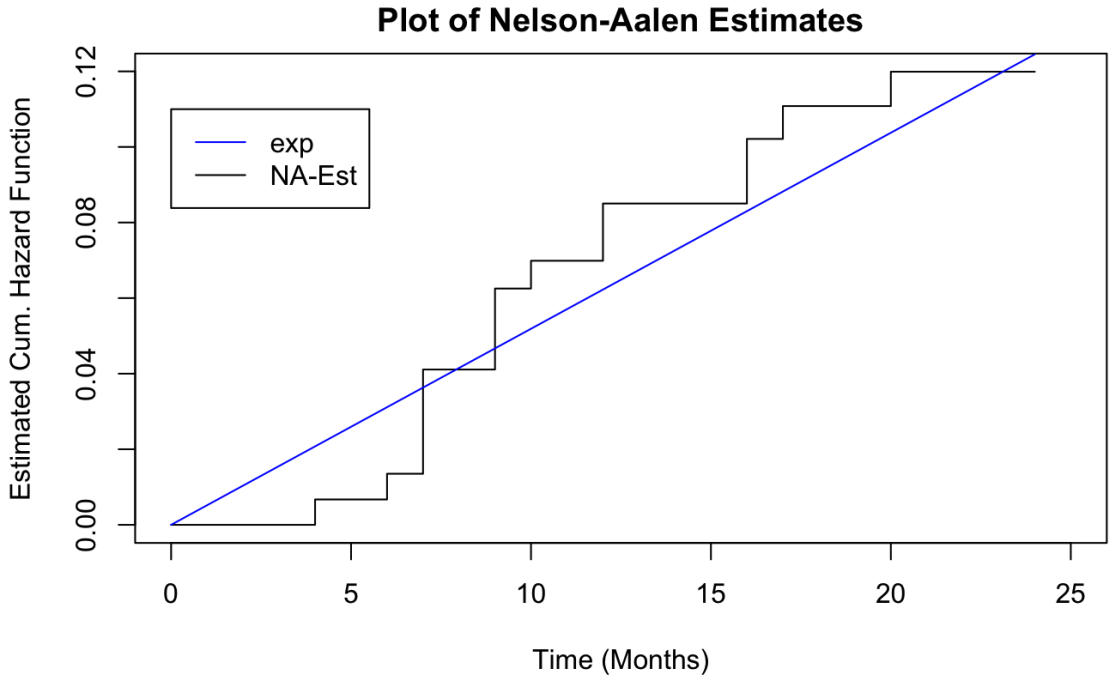

Figure 6: Comparison for Estimated CHF Between K-M and N-A Estimations

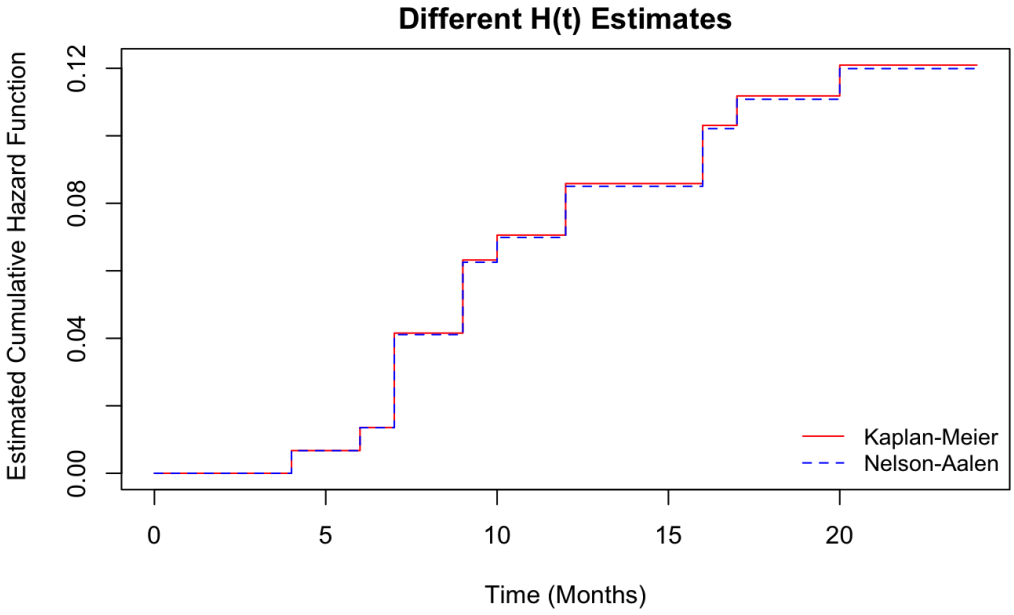

Figure 7: Kaplan-Meier Estimation By MAXFIB Scores

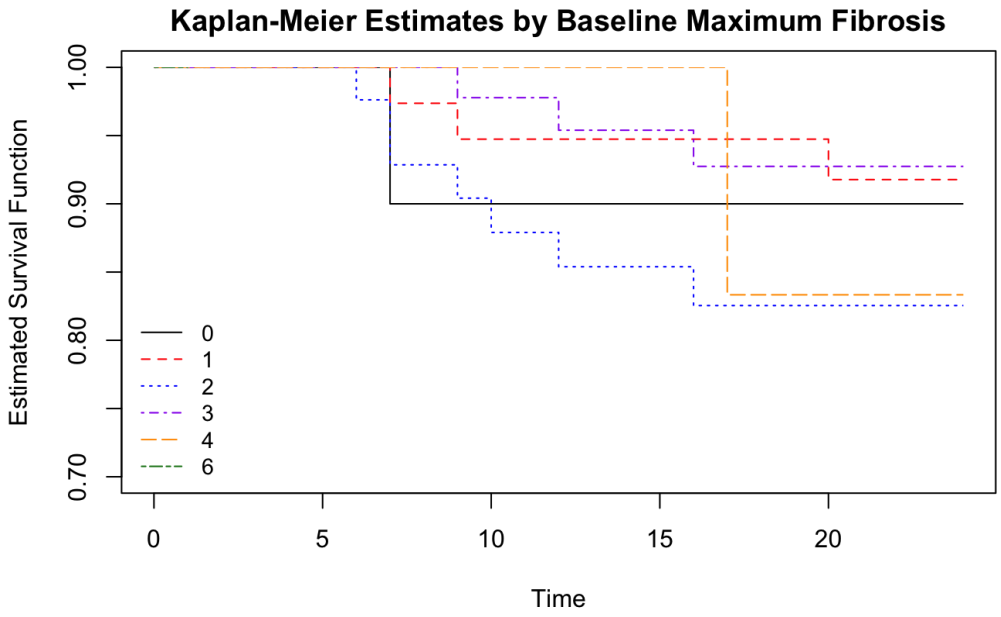

Figure 8: Comparison of Probability of Informative Dropout

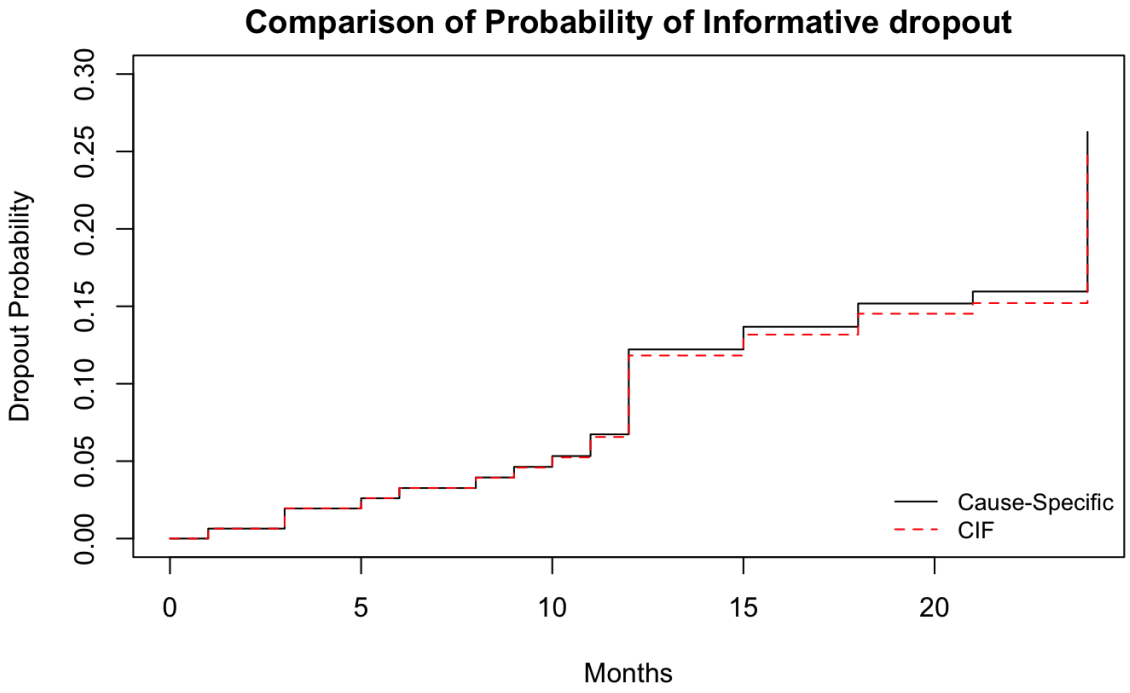

Figure 9: Comparison of Probability of Treatment Failure or Death

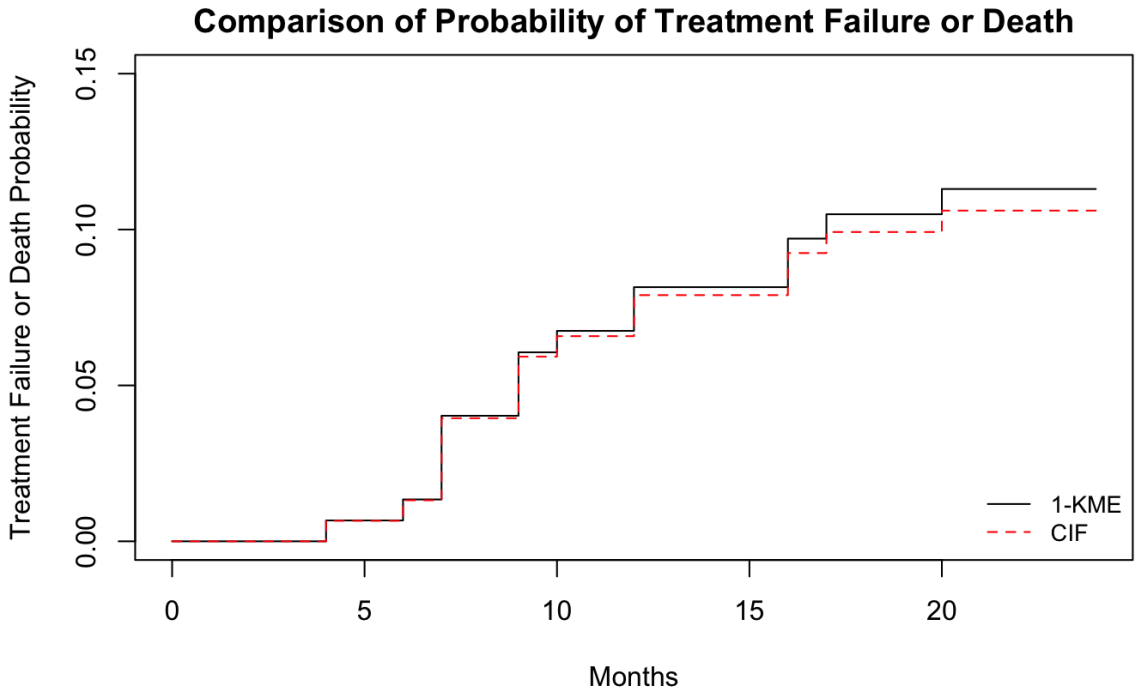

Figure 10: Interaction Between Treatment Failure or Death and Informative Dropout