Survival Analysis of Leukemia Relapse

David

2025-04-08

Leukemia is a blood cancer that develops in the bone marrow or lymphatic system. The condition results from genetic mutations in the blood cell DNA, mostly white blood cells. These mutations cause the cells to grow and divide uncontrollably, leading to an accumulation of abnormal cells in the bone marrow and the blood, which eventually crowds out healthy cells. There are several risk factors associated with Leukemia e.g. family history of Leukemia, genetic conditions like down syndrome, exposure to chemicals like benzene, gasoline and other industrial chemicals, exposure to tobacco smoke, exposure to cancer radiations etc. Leukemia can broadly be classified into chronic Leukemia and acute Leukemia, where chronic Leukemia affects mature blood cells and develops gradually. Acute Leukemia on the other hand, affects immature blood cells and tends to be aggressive (MOFFITT Cancer Center, n.d.).

This study uses the Leukemia Remission dataset, which is one of the cancer-related datasets found in the **OncoDataSets** R package. The study investigates the effectiveness of maintenance therapy (6-mercaptopurine) on Leukemia remission, and how covariates like sex and white blood cell count impact survival.

```
# Load packages
suppressMessages(
  {
    library(tidyverse)
    library(survival)
    library(survivalAnalysis)
    library(survminer)
    library(OncoDataSets) # Contains the Leukemia Remission data
  }
)

# View the first few observations of Leukemia Remission dataset
head(LeukemiaRemission_df)
```

```
##      sex   wbc  time    event    grp
##    <int> <num> <int>   <fctr> <fctr>
## 1:     1  1.45    35 censored   6-MP
## 2:     1  1.47    34 censored   6-MP
## 3:     1  2.20    32 censored   6-MP
## 4:     1  2.53    32 censored   6-MP
## 5:     1  1.78    25 censored   6-MP
## 6:     1  2.57    23  Relapse   6-MP
```

The dataset *LeukemiaRemission_df*, is a data frame containing data on the duration of remission for acute leukemia. Patients were randomly assigned to maintenance therapy

with an active antileukemic compound called 6-mercaptopurine (6-MP), or a placebo. The dataset contains the variables sex, white blood cell count (WBC), time to relapse, event status, and treatment group the patients were assigned to.

```
# View the structure of the dataset
str(LeukemiaRemission_df)
```

```
## Classes 'data.table' and 'data.frame':   42 obs. of  5 variables:
##  $ sex  : int  1 1 1 1 1 1 1 1 0 0 ...
##  $ wbc  : num  1.45 1.47 2.2 2.53 1.78 2.57 2.32 2.01 2.05 2.16 ...
##  $ time : int  35 34 32 32 25 23 22 20 19 17 ...
##  $ event: Factor w/ 2 levels "Relapse","censored": 2 2 2 2 2 1 1 2 2 2 ...
##  $ grp  : Factor w/ 2 levels "6-MP","Placebo": 1 1 1 1 1 1 1 1 1 1 ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

The data has 42 observations of 5 variables. The variables sex, time to relapse and WBC count are numeric, while event status and treatment groups are factors with two levels each.

```
## Assess for data quality issues

# Check for missing values
map_dbl(LeukemiaRemission_df, ~sum(is.na(.)))
```

```
##   sex   wbc  time event   grp
##     0     0     0     0     0
```

```
# Check for duplicated observations
sum(duplicated(LeukemiaRemission_df))
```

```
## [1] 0
```

There are no missing values in the data, as well as duplicated observations.

```
# Factor the variable sex
LeukemiaRemission_df[["sex"]] <- factor(LeukemiaRemission_df[["sex"]],
                                        labels = c("female", "male"),
                                        levels = c(0,1))

# Generate Summary Statistics for each and every variable
summary(LeukemiaRemission_df)
```

```
##      sex          wbc             time           event         grp
##  female:22   Min.   :1.450   Min.   : 1.00   Relapse :30   6-MP   :21
##  male  :20   1st Qu.:2.303   1st Qu.: 6.00   censored:12   Placebo:21
##              Median :2.800   Median :10.50
##              Mean   :2.930   Mean   :12.88
##              3rd Qu.:3.490   3rd Qu.:18.50
##              Max.   :5.000   Max.   :35.00
```

Female patients were 22 and male patients were 20. The minimum, median, mean and maximum white blood cell counts were 1.45, 2.80, 2.93 and 5.00 respectively. The median time to relapse was ten and a half weeks. Half of the patients received the maintenance treatment (6-mercaptopurine) while another half received placebo. 30 patients had a relapse while 12 patients were censored.

Kaplan Meier Model

Kaplan Meier model is a non-parametric model used to analyze time-to-event data. It has no parameter and makes no assumption about the data distribution.

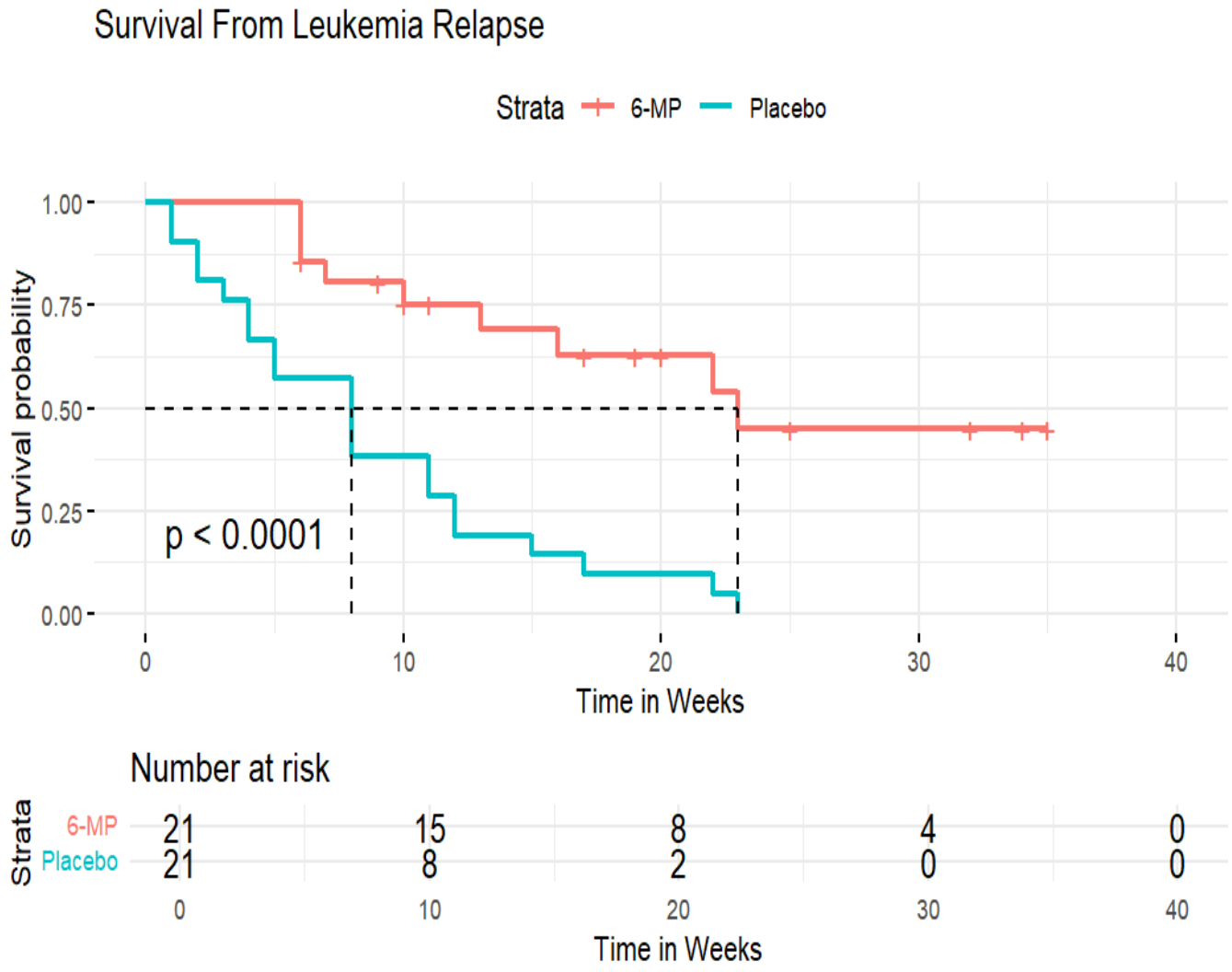
I'll fit two models i.e. one for comparing relapse between treatment groups and another for comparing relapse between males and females.

```
# Fit a Kaplan Meier model with treatment group as the covariate
KM1 <- survfit(Surv(time, event == "Relapse") ~ grp, data =
LeukemiaRemission_df)
# Model summary
summary(KM1)
```

```
## Call: survfit(formula = Surv(time, event == "Relapse") ~ grp, data =
LeukemiaRemission_df)
##
##                 grp=6-MP
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     6     21       3    0.857  0.0764        0.720        1.000
##     7     17       1    0.807  0.0869        0.653        0.996
##    10     15       1    0.753  0.0963        0.586        0.968
##    13     12       1    0.690  0.1068        0.510        0.935
##    16     11       1    0.627  0.1141        0.439        0.896
##    22      7       1    0.538  0.1282        0.337        0.858
##    23      6       1    0.448  0.1346        0.249        0.807
##
##                 grp=Placebo
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     1     21       2   0.9048  0.0641      0.78754        1.000
##     2     19       2   0.8095  0.0857      0.65785        0.996
##     3     17       1   0.7619  0.0929      0.59988        0.968
##     4     16       2   0.6667  0.1029      0.49268        0.902
##     5     14       2   0.5714  0.1080      0.39455        0.828
##     8     12       4   0.3810  0.1060      0.22085        0.657
##    11      8       2   0.2857  0.0986      0.14529        0.562
##    12      6       2   0.1905  0.0857      0.07887        0.460
##    15      4       1   0.1429  0.0764      0.05011        0.407
##    17      3       1   0.0952  0.0641      0.02549        0.356
##    22      2       1   0.0476  0.0465      0.00703        0.322
##    23      1       1   0.0000     NaN           NA           NA
```

The number at risk decreases as we move in time. Survival probability also decreases with time.

```
# Plot Survival Curves
ggsurvplot(KM1, data = LeukemiaRemission_df,
           risk.table = TRUE, pval = TRUE,
           surv.median.line = "hv",
           legend.labs = c("6-MP", "Placebo"),
           risk.table.y.text.col = TRUE,
           risk.table.y.text = TRUE,
           xlab = "Time in Weeks",
           title = "Survival From Leukemia Relapse",
           ggtheme = theme_minimal())
```

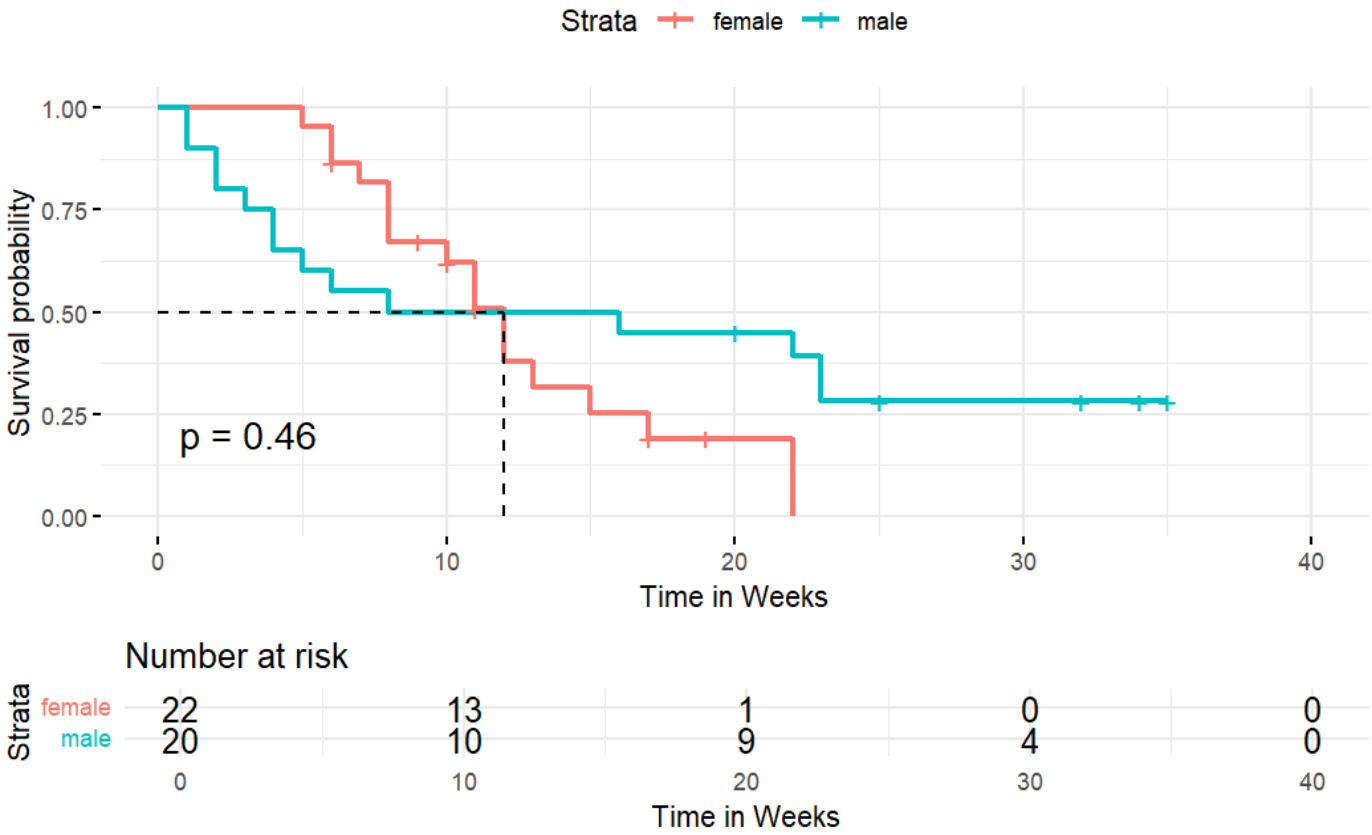


The 6-MP treatment was effective ($p < .05$). Patients who were under the maintenance therapy had higher survival rates. The median survival time for patients who received the 6-MP treatment is 23 weeks, while that of patients who received placebo is 9 weeks.

```
# Fit another Kaplan Meier model with sex as the covariate
KM2 <- survfit(Surv(time, event == "Relapse") ~ sex,
               data = LeukemiaRemission_df)
# Model summary
summary(KM2)
```

```
## Call: survfit(formula = Surv(time, event == "Relapse") ~ sex, data =
LeukemiaRemission_df)
##
##                 sex=female
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     5     22       1    0.955  0.0444       0.8714        1.000
##     6     21       2    0.864  0.0732       0.7315        1.000
##     7     18       1    0.816  0.0834       0.6676        0.997
##     8     17       3    0.672  0.1020       0.4988        0.905
##    10     13       1    0.620  0.1064       0.4429        0.868
##    11     11       2    0.507  0.1131       0.3278        0.785
##    12      8       2    0.380  0.1150       0.2104        0.688
##    13      6       1    0.317  0.1119       0.1587        0.633
##    15      5       1    0.254  0.1060       0.1118        0.575
##    17      4       1    0.190  0.0966       0.0703        0.515
##    22      1       1    0.000     NaN           NA           NA
##
##                 sex=male
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     1     20       2    0.900  0.0671        0.778        1.000
##     2     18       2    0.800  0.0894        0.643        0.996
##     3     16       1    0.750  0.0968        0.582        0.966
##     4     15       2    0.650  0.1067        0.471        0.897
##     5     13       1    0.600  0.1095        0.420        0.858
##     6     12       1    0.550  0.1112        0.370        0.818
##     8     11       1    0.500  0.1118        0.323        0.775
##    16     10       1    0.450  0.1112        0.277        0.731
##    22      8       1    0.394  0.1106        0.227        0.683
##    23      7       2    0.281  0.1038        0.136        0.580
```

```
# Plot Survival Curves to compare the survival rates between males and
females
ggsurvplot(KM2, data = LeukemiaRemission_df,
           risk.table = TRUE, pval = TRUE,
           surv.median.line = "hv",
           legend.labs = c("female", "male"),
           risk.table.y.text.col = TRUE,
           risk.table.y.text = TRUE,
           xlab = "Time in Weeks",
           ggtheme = theme_minimal())
```



The survival rates between males and females do not differ significantly ($p > .05$). The median survival time for both males and females is about 12 weeks. Before reaching the median survival time, females tend to have higher survival rates than males. After the median survival time (12 weeks), males have higher survival rates than females, even though this is not statistically significant.

Cox Proportional Hazard Model

Cox PH model is a semi-parametric model used to model time-to-event data. This model is useful for investigating the association between the survival time and one or more predictor variables (covariates). It provides a way of examining how different covariates influence the hazard rate at a particular point in time. The Cox model is expressed by the hazard function **{h(t)}**, which is given by;

- $h(t) = h0(t)*\exp(\beta 1x1+\beta 2x2+...+\beta pxp)$
- The term **h0** is called the baseline hazard. It corresponds to the value of the hazard when all the covariates (xi's) are at/equal to zero for numeric features (the quantity exp(0) equals 1), or at reference levels for categorical features.

The hazard ratios are given by **exp(βi)**. To summarize the hazard ratios;

- HR = 1: No difference in the hazard rates between the groups being compared, or increase/decrease in the values of a numeric covariate does not affect the hazard rate.
- HR < 1: Reduction in the hazard i.e. one group has a lower hazard rate than the other group, or increase in the values of a numeric covariate decreases the hazard rate.
- HR > 1: Increase in hazard i.e. one group has a higher hazard rate than the other group, or increase in the values of a numeric covariate increases the hazard rate and in-turn reduces the length of survival.

The Cox PH model is popular because it is robust, and the results from using the Cox model will closely approximate the results from the correct parametric model.

Advantages of Cox PH Model

- Flexibility - Since the Cox PH model is a semi-parametric model, it allows for flexibility in modeling the baseline hazard function. This makes it suitable for a wide range of survival data, including those where the hazard function is not easily described by a parametric distribution.
- The Cox PH model can handle censored data effectively.
- The Cox PH model provides easily interpretable hazard ratios, which quantify the effect of covariates on the hazard of experiencing an event.
- The Cox PH model can handle time-dependent covariates, allowing researchers to account for changing variables over the course of a study.

Disadvantages of the Cox PH Model

- The Cox PH model does not provide a direct estimate of the baseline hazard function.
- Cox PH model can be sensitive to outliers in the data.
- Multicollinearity - If covariates are highly correlated, the Cox PH model can suffer from multicollinearity issues, making it challenging to estimate the independent effects of individual variables accurately.
- Proportional Hazards Assumption - The validity of the Cox PH model relies on the assumption that the hazard ratios (relative risks) remain constant over time. This assumption may not hold in all cases, and when it's violated, it can lead to biased or incorrect results.

Before fitting a Cox PH model, I'll reverse the order in which the levels of treatment group appear so that the new order begins with Placebo. This will make it easy to interpret the results of the Cox PH model.

```
# Reverse the levels of treatment group to begin with Placebo
LeukemiaRemission_df <- LeukemiaRemission_df |>
```

```
  mutate(group = factor(grp, labels = rev(c("6-MP", "Placebo")),
                        levels = rev(levels(LeukemiaRemission_df$grp))))

# Fit a Cox PH model, use treatment group, white blood cell count and sex as
the covariates
Cox_model <- coxph(formula = Surv(time, event == "Relapse") ~ group + wbc +
sex,
                   data = LeukemiaRemission_df)
# Model summary
summary(Cox_model)
```

```
## Call:
## coxph(formula = Surv(time, event == "Relapse") ~ group + wbc +
##     sex, data = LeukemiaRemission_df)
##
##   n= 42, number of events= 30
##
##               coef exp(coef) se(coef)      z Pr(>|z|)
## group6-MP -1.5036    0.2223   0.4615 -3.258  0.00112 **
## wbc        1.6819    5.3760   0.3366  4.997 5.82e-07 ***
## sexmale    0.3147    1.3698   0.4545  0.692  0.48872
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##           exp(coef) exp(-coef) lower .95 upper .95
## group6-MP    0.2223      4.498   0.08998    0.5493
## wbc          5.3760      0.186   2.77944   10.3982
## sexmale      1.3698      0.730   0.56206    3.3384
##
## Concordance= 0.851  (se = 0.041 )
## Likelihood ratio test= 47.19  on 3 df,   p=3e-10
## Wald test            = 33.54  on 3 df,   p=2e-07
## Score (logrank) test = 48.01  on 3 df,   p=2e-10
```

The overall model is significant (p-values from Likelihood ratio test, Wald test and logrank test are all less than 0.05). The sex covariate is however not significant ($p > .05$), while 6-MP treatment and white blood cell count are significant ($p < .05$).

- Holding other covariates constant, the risk of Leukemia relapse for patients who received 6-MP treatment is reduced by a factor of 0.22, or 77.77%.
- Holding other covariates constant, every unit increase in WBC count induces the hazard of relapse by a factor of 5.376. In other words, increased accumulation of abnormal cells in the bone marrow and the blood crowds out healthy cells, and this increases the hazard of Leukemia relapse.

```
# Test for the Proportional Hazards Assumption
cox.zph(Cox_model)
```

```
##        chisq df    p
## group  0.036  1 0.85
```

```
## wbc    0.142  1 0.71
## sex    5.420  1 0.02
## GLOBAL 5.879  3 0.12
```

The global test is not significant (p > .05), implying that the proportional Hazards assumption holds, even though it only holds for the treatment group and white blood cell count covariates. The assumption does not hold for the sex covariate.

```
# Fit another Cox PH model, this time omitting the sex covariate
Cox_Model <- coxph(formula = Surv(time, event == "Relapse") ~ group + wbc,
                   data = LeukemiaRemission_df)
# Model summary
summary(Cox_Model)
```

```
## Call:
## coxph(formula = Surv(time, event == "Relapse") ~ group + wbc,
##     data = LeukemiaRemission_df)
##
##   n= 42, number of events= 30
##
##              coef exp(coef) se(coef)      z Pr(>|z|)
## group6-MP -1.3861    0.2501   0.4248 -3.263   0.0011 **
## wbc        1.6909    5.4243   0.3359  5.034  4.8e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##           exp(coef) exp(-coef) lower .95 upper .95
## group6-MP    0.2501     3.9991    0.1088    0.5749
## wbc          5.4243     0.1844    2.8082   10.4776
##
## Concordance= 0.852  (se = 0.04 )
## Likelihood ratio test= 46.71  on 2 df,   p=7e-11
## Wald test            = 33.6  on 2 df,   p=5e-08
## Score (logrank) test = 46.07  on 2 df,   p=1e-10
```

The overall model is significant (p-values from Likelihood ratio test, Wald test and logrank test are all less than 0.05). The covariates 6-MP treatment and WBC count are also highly significant (p < .05).

- Holding the WBC covariate constant, the risk of Leukemia relapse for patients who received 6-MP treatment is reduced by a factor of 0.25, or 74.99%.
- Holding the treatment group covariate constant, every unit increase in WBC count induces the hazard of Leukemia relapse by a factor of 5.42 (Increased accumulation of abnormal cells in the bone marrow and the blood crowds out healthy cells, thereby increasing the hazard of Leukemia relapse).

References

City of Hope. (2022, May 26). Leukemia causes and risk factors. CancerCenter.com. Retrieved on April 7, 2025, from https://www.cancercenter.com/cancer-types/leukemia/risk-factors

Kassambara, A. (2016, December). Cox proportional hazards model. STHDA. Retrieved on April 7, 2025, from http://www.sthda.com/english/wiki/cox-proportional-hazards-model

Kassambara, A. (2016). survminer: Survival analysis and visualization [GitHub repository]. Retrieved on April 7, 2025, from https://github.com/kassambara/survminer

Kassambara, A. (2016, December). Cox model assumptions. STHDA. Retrieved on April 7, 2025, from http://www.sthda.com/english/wiki/cox-model-assumptions

MOFFITT Cancer Center. (n.d.). Causes of leukemia: Understanding risk factors. Retrieved on April 7, 2025, from https://www.moffitt.org/cancers/leukemia/diagnosis/causes/