# Predicting Time to Death for Acute Myeloid Leukemia Using Survival Analysis

Written by Sierra Rossman

## Part 1: Getting Set Up

Before we dive in, we'll need a few R packages to make this analysis smooth:

```
library(dplyr)    #For data manipulation

library(survival) #The workhorse for survival analysis in R

library(survminer)#To make our survival plots look pretty

library(gtsummary)#For nice summary tables
```

The `` `survival` `` package comes pre-installed with R and is essential for this kind of analysis. If you want better-looking graphs, though, `` `survminer` `` is your best friend!

## Part 2: The Data

We're working with a dataset from cBioPortal for cancer genomics that includes genomic and clinical data from 942 acute myeloid leukemia (AML) samples from Oregon Health and Science University's "Beat AML" program. This data spans over 10 years and represents 805 patients. While the dataset originally had 67 variables, we'll keep it simple and focus on 8.

**AgeAtDiagnosis:** How old the patient was when they were diagnosed.

**NumCumulativeTreatments:** Number of treatments they received.

**MutationCount:** Number of gene mutations.

**OverallSurvivalMonths:** How long they survived post-diagnosis.

**OverallSurvivalStatus:** Whether they survived by the study's end or not.

**Sex:** Male or female.

**SpecimenType:** How/where the sample was collected.

```
data <- read.csv("aml_ohsu_2022_clinical_data.csv")
```

## Part 3: What Is Survival Analysis?

In general, survival analysis analyzes time-to-event data and is all about answering questions like:

- How long do AML patients typically survive after diagnosis?
- Does survival time differ between men and women?
- Does the type of sample collection impact survival time?

When you have data from a medical study, particularly when you're measuring survival, it is likely that not all of the patients will die by the end of the study. Or, maybe you have patients who drop out of the study for reasons other than death. In both of these situations, you won't have a time the event (death) takes place since it never occurred. This is termed "censoring" – meaning we know that the patient survived up to the loss of follow up (end of the study, time of drop out, etc.), but you don't know anything about their survival after that. The good news is survival analysis can handle this!

During this analysis, we'll look at three methods:

### 1. The Kaplan-Meier Curve

A graph that illustrates the cumulative survival probability over time and is great for visualizing survival differences between groups.

### 2. The Cox Proportional-Hazards Regression Model

This regression model is useful for assessing how variables impact survival and can do so for multiple variables.

### 3. The Log Rank Test

This test compares the survival experience between two or more independent groups.

## Part 4: Survival Curves

Let's start by creating a survival object that combines `OverallSurvivalMonths` and `OverallSurvivalStatus` called "s." This object shows us how long the patient was tracked (in months) and whether the event (death) occurred, or the sample was censored (marked with a `+`).

For example, we can see that the first patient was tracked for about 8.84 months and was not censored.

```r
s <- Surv(data$OverallSurvivalMonths, data$OverallSurvivalStatus)
head(s)
```

```
[1]  8.8438356 11.0136986  1.7424658  0.5589041 13.2821918 22.6520548
```

Now, we can create a basic survival curve that doesn't consider any groupings. To see what is happening with our survival curve, we can group the times and event occurs or a sample is censored by every 24 months (2 years). We can see the time interval, how many patients are still alive, how many patients died during the interval, and the rate of survival in this table.

```
sfit <- survfit(Surv(OverallSurvivalMonths, OverallSurvivalStatus)~1, data =
data)

summary(sfit, times = seq(0,200,24))

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    0    910       0    1.000  0.0000       1.0000        1.000
   24    267     475    0.410  0.0178       0.3768        0.447
   48    125      63    0.295  0.0180       0.2616        0.332
   72     48      20    0.240  0.0185       0.2066        0.279
   96     13       5    0.203  0.0221       0.1636        0.251
  120      3       2    0.130  0.0450       0.0663        0.256
  144      1       0    0.130  0.0450       0.0663        0.256
  168      1       0    0.130  0.0450       0.0663        0.256
```

Now, let's see how sex impacts the survival of a patient!

To do this, we replace the default intercept with the `Sex` variable.

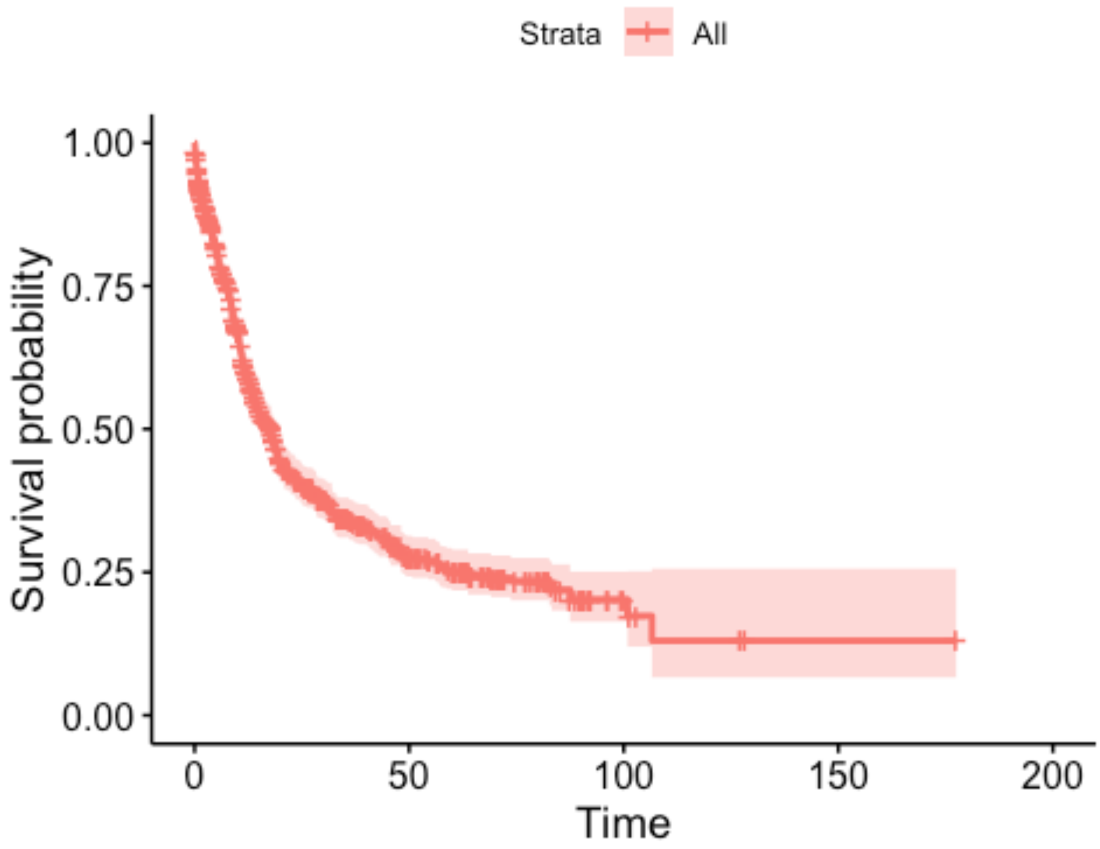
```
sfitSex <- survfit(Surv(OverallSurvivalMonths, OverallSurvivalStatus)~Sex, da
ta = data)
```

### Kaplan-Meier Survival Curve

Now that we've fit two survival curves, let's visualize them using Kaplan-Meier graphs.

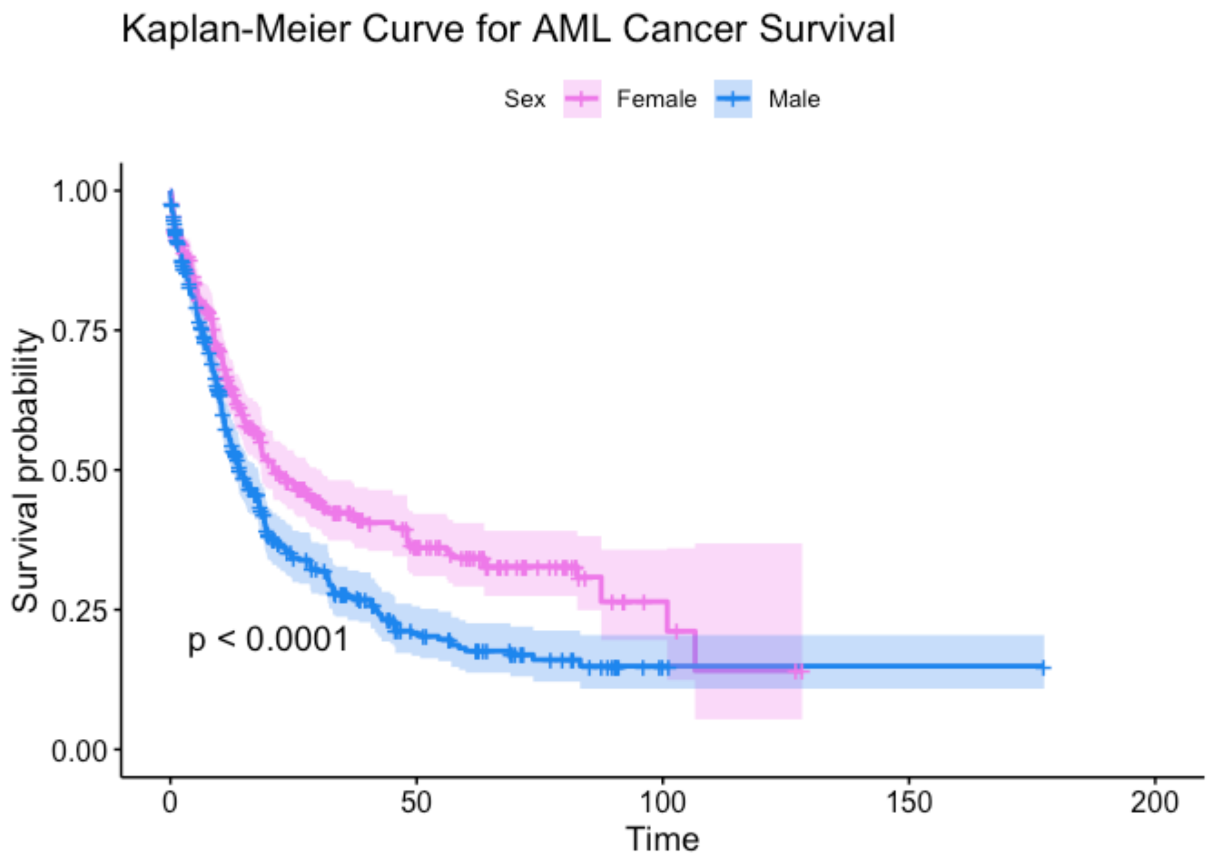
First, we'll visualize the simple curve with no groupings to show the overall survival of patients in the study.

```
ggsurvplot(sfit, data)
```



Now let’s we can visualize the survival curves grouped by sex.

```
ggsurvplot(sfitSex, data, conf.int=TRUE, pval=TRUE, risk.table=FALSE,
           legend.labs=c("Female", "Male"), legend.title="Sex",
           palette=c("orchid2", "dodgerblue2"),
           title="Kaplan-Meier Curve for AML Cancer Survival")
```



Here's what we can see: females generally had higher survival probabilities than males – up until about 110 months, when their probabilities dropped off earlier.

---

## Part 5: Cox Proportional-Hazard Model

To dive deeper, we may want to see how multiple variables influence a patient's survival. To do this, we'll use a Cox regression model.

As previously mentioned, the variables we are looking at are `` `AgeAtDiagnosis` ``, `` `Sex` ``, `` `MutationCount` ``, `` `NumCumulativeTreatments` ``, and `` `SpecimenType` ``.

```
fit <- coxph(Surv(OverallSurvivalMonths, OverallSurvivalStatus)~AgeAtDiagnosi
s+
               Sex+MutationCount+NumCumulativeTreatments+
               SpecimenType, data = data) %>%
  tbl_regression(exp = TRUE)
fit
```

| Characteristic | HR[1] | 95% CI[1] | p-value |
|---|---|---|---|
| AgeAtDiagnosis | 1.03 | 1.02, 1.04 | <0.001 |
| Sex | | | |
| Female | — | — | |
| Male | 1.25 | 1.05, 1.49 | 0.013 |
| MutationCount | 1.00 | 0.99, 1.01 | 0.5 |
| NumCumulativeTreatments | 0.75 | 0.68, 0.84 | <0.001 |
| SpecimenType | | | |
| Bone Marrow Aspirate | — | — | |
| Leukapheresis | 2.91 | 1.74, 4.87 | <0.001 |
| Peripheral Blood | 1.38 | 1.16, 1.64 | <0.001 |

[1] HR = Hazard Ratio, CI = Confidence Interval

What is of interest to us is the hazard ratio (HR). If HR < 1, it means that there is a reduced hazard/chance of death whereas a HR > 1 means that there is an increased hazard/chance of death.

Highlights from our results:

- Age at Diagnosis: Older patients had a higher hazard (chance of death).
- Sex: Males had a 25% higher hazard (chance of death) compared to females.
- Specimen Type: Patients with leukapheresis samples face significantly higher hazards, which makes sense because it indicates a more aggressive disease.

---

## Conclusion

Our analysis sheds light on key factors influencing survival in AML patients: age, sex, and sample type. But there's still more to explore. For instance, why doesn't the number of mutations seem to impact survival? Could other variables—like lifestyle or additional genomic factors—hold the missing clues?

By predicting survival outcomes, we're not just crunching numbers; we're paving the way for more personalized treatments and better outcomes for patients. That's what makes this work exciting.