Teaching Survival Analysis to Clinical Collaborators

Emily C. Zabor

R/Medicine Conference

September 7, 2018

Memorial Sloan Kettering Cancer Center

COLUMBIA UNIVERSITY | MAILMAN SCHOOL of PUBLIC HEALTH
BIOSTATISTICS

The most common questions in cancer research relate to disease survival

- What is the probability of survival to a certain point in time?
- What is the average survival time?

Survival analysis is a complex statistical procedure, so communication with collaborators is key

Strategies for conveying important information about survival analysis:

☐ Be ready with **examples** to explain complex ideas

☐ Use detailed **graphics** to accomplish this

☐ Accompany numbers such as p-values and hazard ratios with detailed **explanations**

Example of a dataset with censored data for a clinical application

The `lung` dataset is available from the `survival` package in R. The data contain subjects with advanced lung cancer from the North Central Cancer Treatment Group.

Variable descriptions, from the documentation:

- inst: Institution code
- time: Survival time in days
- status: censoring status 1=censored, 2=dead
- age: Age in years
- sex: Male=1 Female=2
- ph.ecog: ECOG performance score (0=good 5=dead)
- ph.karno: Karnofsky performance score (bad=0-good=100) rated by physician
- pat.karno: Karnofsky performance score as rated by patient
- meal.cal: Calories consumed at meals
- wt.loss: Weight loss in last six months

Example of a dataset with censored data for a clinical application

The `lung` dataset is available from the `survival` package in R. The data contain subjects with advanced lung cancer from the North Central Cancer Treatment Group.

Variable descriptions, from the documentation:

- inst: Institution code
- time: Survival time in days
- status: censoring status 1=censored, 2=dead
- age: Age in years
- sex: Male=1 Female=2
- ph.ecog: ECOG performance score (0=good 5=dead)
- ph.karno: Karnofsky performance score (bad=0-good=100) rated by physician
- pat.karno: Karnofsky performance score as rated by patient
- meal.cal: Calories consumed at meals
- wt.loss: Weight loss in last six months

Example of a dataset with censored data for a clinical application

The `lung` dataset is available from the `survival` package in R. The data contain subjects with advanced lung cancer from the North Central Cancer Treatment Group.

Variable descriptions, from the documentation:

- inst: Institution code
- time: Survival time in days
- status: censoring status 1=censored, 2=dead
- age: Age in years
- sex: Male=1 Female=2
- ph.ecog: ECOG performance score (0=good 5=dead)
- ph.karno: Karnofsky performance score (bad=0-good=100) rated by physician
- pat.karno: Karnofsky performance score as rated by patient
- meal.cal: Calories consumed at meals
- wt.loss: Weight loss in last six months

Censoring can be straightforward to understand in the context of a clinical trial
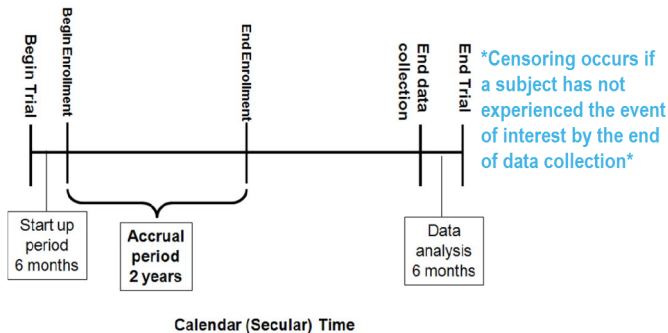
RICH JT, NEELY JG, PANIELLO RC, VOELKER CCJ, NUSSENBAUM B, WANG EW. A PRACTICAL GUIDE TO UNDERSTANDING KAPLAN-MEIER CURVES. Otolaryngology head and neck surgery: official journal of American Academy of Otolaryngology Head and Neck Surgery. 2010;143(3):331-336. doi:10.1016/j.otohns.2010.05.007.

In retrospective data follow-up time is not fixed, but censoring still occurs
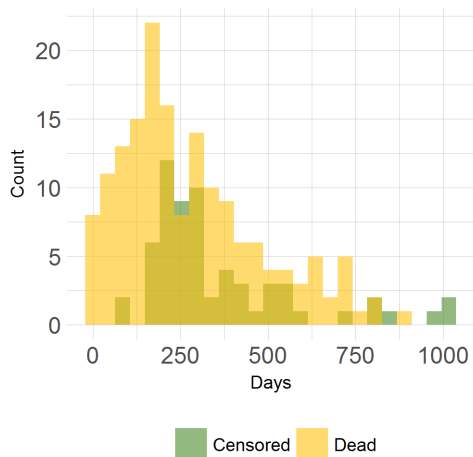
- Censored subjects had not yet died at date of data extraction

- Censored subjects still provide information

- The distribution of follow-up times is skewed

Question from collaborator: What is a survival curve?
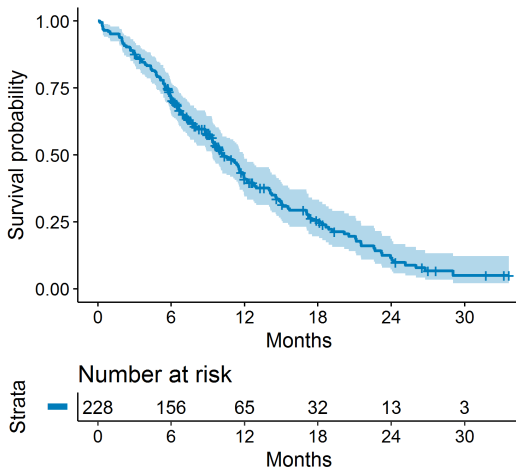
The **Kaplan-Meier survival curve** is fundamental to survival analysis.

The Kaplan-Meier survival function forms the basis of most survival analyses

Some basic facts:

- The x-axis is time and the y-axis is the survival function

- The survival function is calculated at each time as the ratio of subjects who did not experience the event by that time to the total number of subjects still at risk at that time

- Step function where each step down represents a time at which one or more events occurred

- Censored subjects are usually denoted by tick marks

Question from collaborator: Can I report the percentage of events out of the total study population?
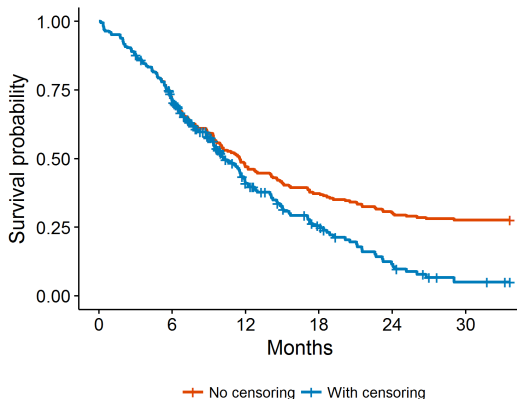
Imagine two studies, each with 228 subjects. There are 165 deaths in each study. The orange study has no censoring, the blue study includes censored subjects (the true `lung` data).

Ignoring censoring leads to an overestimate of the survival probability

- The orange study has no censoring. 165/228 subjects died $\rightarrow$ survival probability at end of study $= 1 - \frac{165}{228} \times 100 = 28\%$.

- The blue group has 63 subjects censored. Still 165/228 subjects died.

- But the blue curve is clearly lower than the orange curve, especially at the end $\rightarrow$ survival probability at end of study $=$ something $< 28\%$ (actually 0.05%).

- This occurs because the censored subjects only contribute information for **part** of the follow-up time, and then they **fall out of the denominator**.

Question from collaborator: What is the 1-year survival probability?
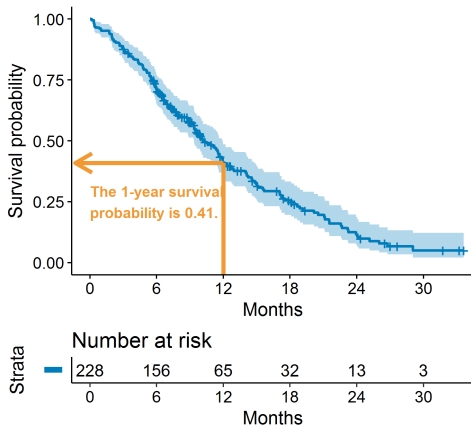
An annotated survival plot with reproducible text output using the `glue` package can help show that it is the probability on the y-axis corresponding to 1-year on the x-axis.

Additionally include reproducible text with the image to make sure numbers are being correctly interpreted

As part of a reproducible workflow that includes writing reports in R Markdown, the `glue_data` function from the `glue` package can easily print reproducible text as a corresponding description:

```r
fit1 <- survival::survfit(
  survival::Surv(time, status) ~ 1, data = lung)
glue::glue_data(summary(fit1, times = 365.25),
                "The survival probability at 1-year is ",
                "{round(surv, 2)} ",
                "(95% CI: ", "{round(lower, 2)} - ", "{round(upper, 2)}",
                "), which represents the proportion of ",
                "patients who survived beyond 1 year.")
```

```
## The survival probability at 1-year is 0.41 (95% CI: 0.34 - 0.49), which repr
```

Question from collaborator: What is the meaning of median survival time?
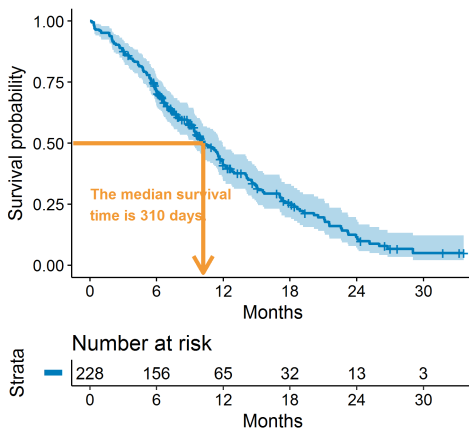
The **median survival time** is the point on the x-axis corresponding to a survival probability of 0.5 on the y-axis.

Inline R code can also be used to incorporate reproducible text alongside graphical examples

First define the survfit summary object:

```
fit1 <- survfit(Surv(time, status) ~ 1, data = lung)
res <- summary(fit1)$table
```

Then:

The median survival time was `` `r round(res["median"])` `` days (95% CI: `` `r round(res["0.95LCL"])` `` - `` `r round(res["0.95UCL"])` ``), which represents the point in time that half of subjects will live beyond.

On knitting this will print our desired description:

The median survival time was 310 days (95% CI: 285 - 363), which represents the point in time that half of subjects will live beyond.

Question from collaborator: why can't I just estimate the median among those who had the event?
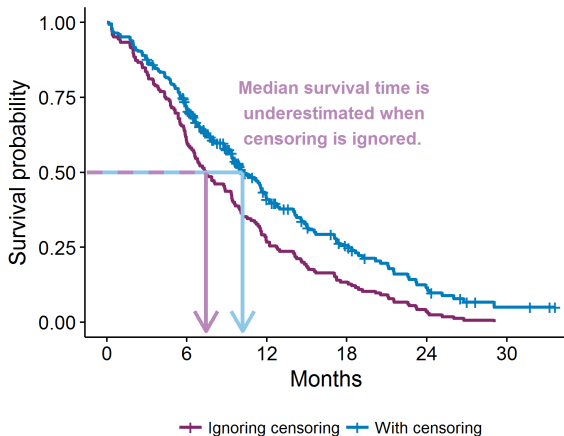

This graphical example compares the true survival curve in blue, and the curve excluding censored subjects in purple.

Ignoring censoring leads to an underestimate of median survival time

- The purple group ignores censored subjects. This is equivalent to a naive estimate of median time to death among subjects who died, which is 226 days.

- The blue group includes the censored subjects. The median survival time is clearly longer, at 310 days.

- This occurs because the censored subjects do contribute information toward calculation of median survival time.

- Even though we don't know **exactly** when these patients died, we know they died **after** the time of censoring.

Question from collaborator: what does it mean when you say median survival is not reached?
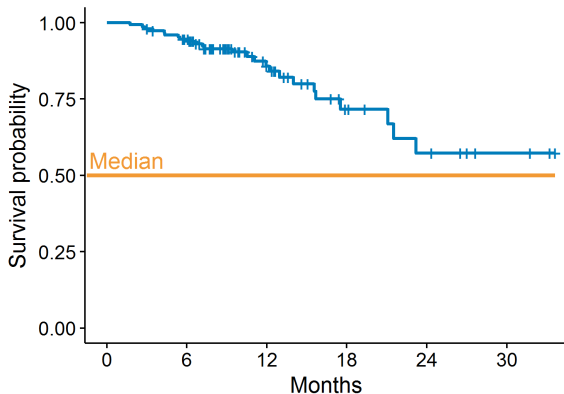
The survival curve must **cross** the survival probability of 0.5 before median survival time has been reached.

(Survival curve based on a random sample of `lung` data subjects, sampling a proportion of those who died and oversampling censored subjects.)

Median survival time is only observed when subjects are followed long enough

When the median is not reached:

- By the end of the study, we have not observed the point in time that half the subjects will survive beyond
- Subjects would need to be followed longer, so that more deaths could occur, before we could observe median survival time
- Does **not** mean that the median survival time is greater than the maximum observed time in the study
- Usually reported as "not reached" or "NR"

Question from collaborator: what is a hazard ratio?

A **hazard ratio (HR)** represents the instantaneous chance of the event occurring in one group, divided by the instantaneous chance of the event occurring in the other group.

A simple example can help. We conduct a study comparing recurrence rates between treatment A and treatment B:

- **HR = 0.5**: at any particular time, **half** as many patients on treatment A are experiencing the event as compared to treatment B
- **HR = 1**: at any particular time, the **same** number of patients on treatment A are experiencing the event as compared to treatment B
- **HR = 2**: at any particular time, **twice** as many patients on treatment A are experiencing the event as compared to treatment B

Useful R packages

- survival
- survminer
- ggplot2
- glue
- rmarkdown
- knitr

Thank you

Slides available at: https://github.com/zabore/r-medicine

Contact me:
@zabormetrics
@zabore
http://www.emilyzabor.com/