Teaching Survival Analysis to Clinical Collaborators

Emily C. Zabor

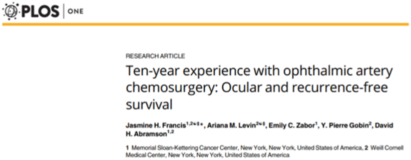
R/Medicine Conference

September 7, 2018

1884 Memorial Sloan Kettering Cancer Center

COLUMBIA UNIVERSITY | MAILMAN SCHOOL of PUBLIC HEALTH
BIOSTATISTICS

The most common questions in cancer research relate to disease survival



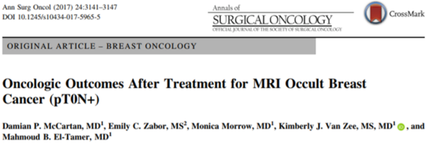
PLOS | ONE

RESEARCH ARTICLE

Ten-year experience with ophthalmic artery chemosurgery: Ocular and recurrence-free survival

Jasmine H. Francis[1,2☯‡*], Ariana M. Levin[2☯‡], Emily C. Zabor[1], Y. Pierre Gobin[2], David H. Abramson[1,2]

1 Memorial Sloan-Kettering Cancer Center, New York, New York, United States of America, 2 Weill Cornell Medical Center, New York, New York, United States of America



Ann Surg Oncol (2017) 24:3141–3147
DOI 10.1245/s10434-017-5965-5

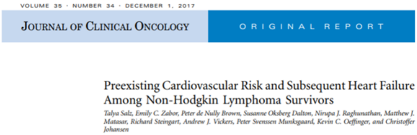
Annals of SURGICAL ONCOLOGY
OFFICIAL JOURNAL OF THE SOCIETY OF SURGICAL ONCOLOGY

CrossMark

ORIGINAL ARTICLE – BREAST ONCOLOGY

Oncologic Outcomes After Treatment for MRI Occult Breast Cancer (pT0N+)

Damian P. McCartan, MD[1], Emily C. Zabor, MS[2], Monica Morrow, MD[1], Kimberly J. Van Zee, MS, MD[1], and Mahmoud B. El-Tamer, MD[1]

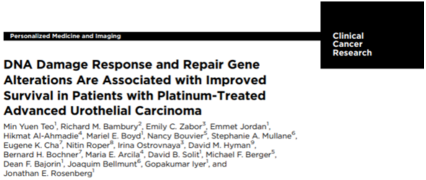


VOLUME 35 · NUMBER 34 · DECEMBER 1, 2017

JOURNAL OF CLINICAL ONCOLOGY | ORIGINAL REPORT

Preexisting Cardiovascular Risk and Subsequent Heart Failure Among Non-Hodgkin Lymphoma Survivors

Talya Salz, Emily C. Zabor, Peter de Nully Brown, Susanne Oksberg Dalton, Nirupa J. Raghunathan, Matthew J. Matasar, Richard Steingart, Andrew J. Vickers, Peter Svenssen Munksgaard, Kevin C. Oeffinger, and Christoffer Johansen



Personalized Medicine and Imaging

Clinical Cancer Research

DNA Damage Response and Repair Gene Alterations Are Associated with Improved Survival in Patients with Platinum-Treated Advanced Urothelial Carcinoma

Min Yuen Teo[1], Richard M. Bambury[2], Emily C. Zabor[3], Emmet Jordan[1], Hikmat Al-Ahmadie[4], Mariel E. Boyd[1], Nancy Bouvier[5], Stephanie A. Mullane[6], Eugene K. Cha[7], Nitin Roper[8], Irina Ostrovnaya[3], David M. Hyman[9], Bernard H. Bochner[7], Maria E. Arcila[4], David B. Solit[1], Michael F. Berger[5], Dean F. Bajorin[1], Joaquim Bellmunt[6], Gopakumar Iyer[1], and Jonathan E. Rosenberg[1]

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 convey information

□ Accompany important numbers with **interpretations**

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

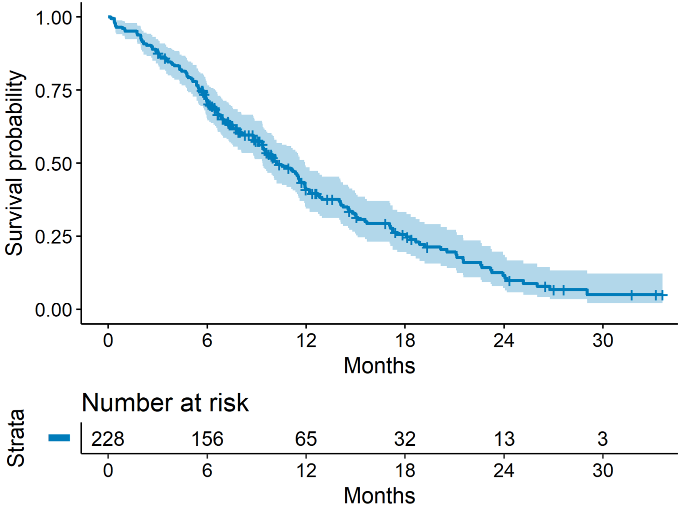
The `lung` dataset is available from the `survival` package in `R`. The data contain subjects with advanced lung cancer from the North Central Cancer Treatment Group.

Variable descriptions, from the documentation:

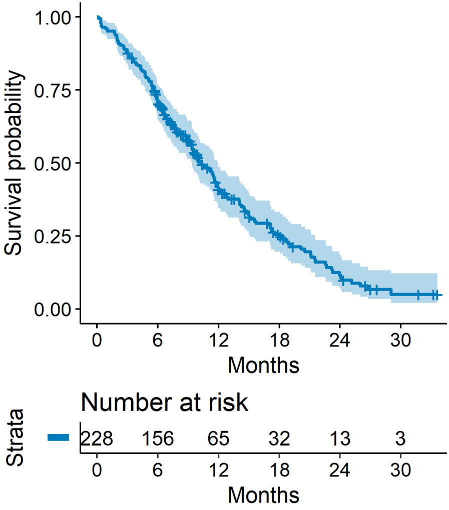
- inst: Institution code
- time: Survival time in days
- status: censoring status 1=censored, 2=dead
- age: Age in years
- sex: Male=1 Female=2
- ph.ecog: ECOG performance score (0=good 5=dead)
- ph.karno: Karnofsky performance score (bad=0-good=100) rated by physician
- pat.karno: Karnofsky performance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

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

Using the `ggsurvplot` function from the `survminer` package:



First just describe what is on the plot

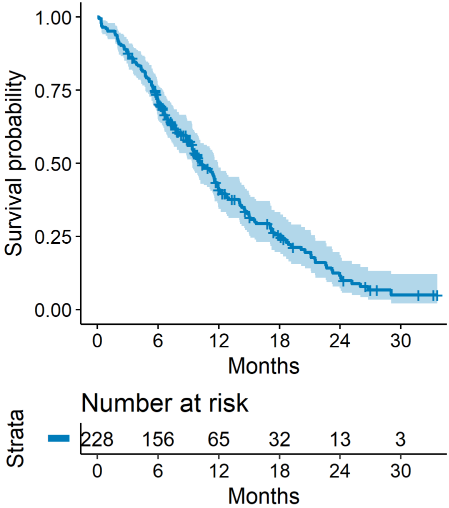


The x-axis is time and the y-axis is the survival function

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

Censored subjects are usually denoted by tick marks

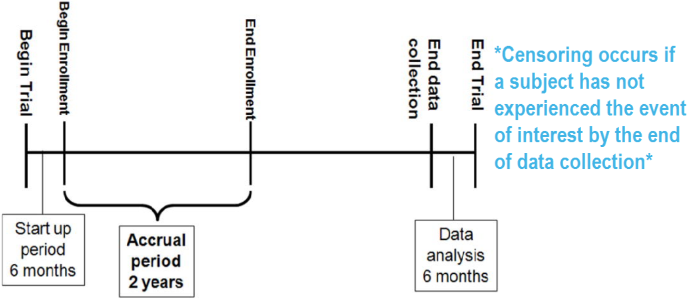
If appropriate give a basic description of the underlying calculations



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

Question from collaborator: What is censoring?

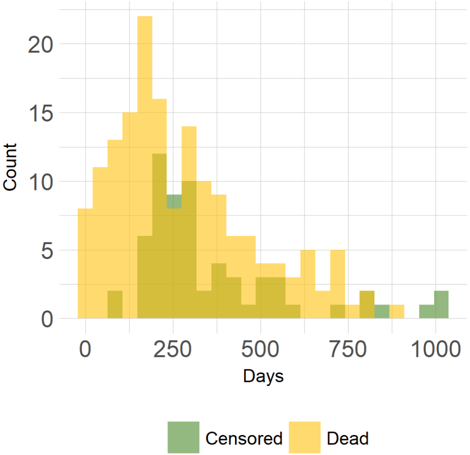
In the context of a clinical trial:



Censoring occurs if a subject has not experienced the event of interest by the end of data collection

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.

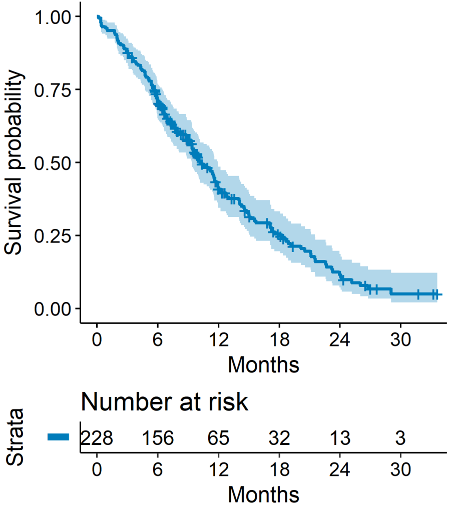
Question from collaborator: Why do I need specialized methods to analyze time-to-event data?



1. Censored subjects still provide information so must be appropriately included in the analysis

2. The distribution of follow-up times is skewed

Question from collaborator: Can I report the percentage of events out of the total study population?

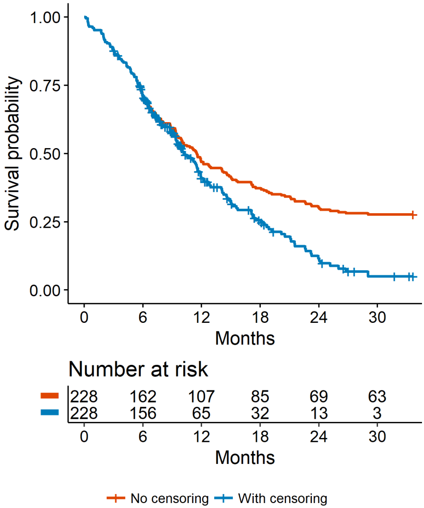


In the `lung` data this would lead to an estimate of survival probability at the end of the study of $1 - \frac{165}{228} = 0.28$.

But this is **incorrect**.

Why?

The survival curve is lower when there is censoring during follow-up

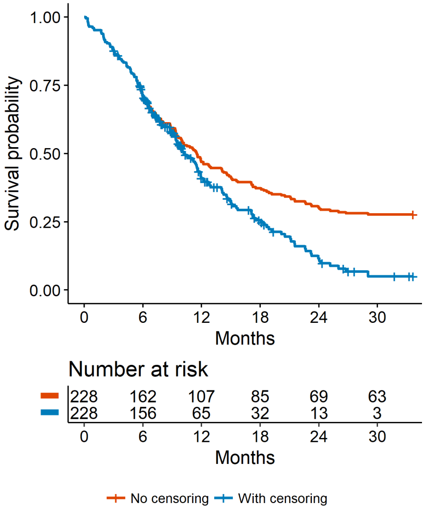


Imagine two studies, each with 228 subjects.

There are 165 deaths in each study.

The orange study has no censoring during follow-up, the blue study has subjects censored throughout (the true lung data).

When there is no censoring during follow-up, the survival estimate is straightforward



Number at risk

	0	6	12	18	24	30
No censoring	228	162	107	85	69	63
With censoring	228	156	65	32	13	3

Months

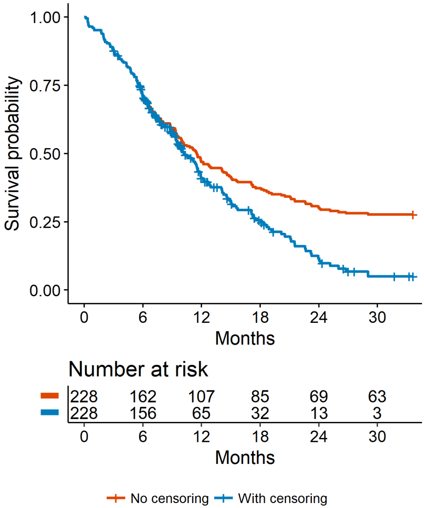
In the orange study everyone is followed until the last follow-up time (i.e. no censoring during follow-up).

The survival probability at last follow-up is 0.28

This is equivalent to an estimating the raw percent of patients still alive:

$$1 - \frac{165}{228} = 0.28$$

When there is censoring during follow-up, a naive estimate will be incorrect



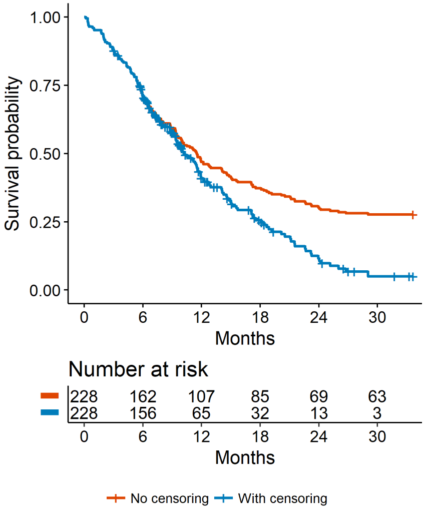
	0	6	12	18	24	30
🟧	228	162	107	85	69	63
🟦	228	156	65	32	13	3

The blue study has 63 subjects censored during follow-up. Still 165 subjects died.

The blue curve is lower than the orange curve, especially at the end.

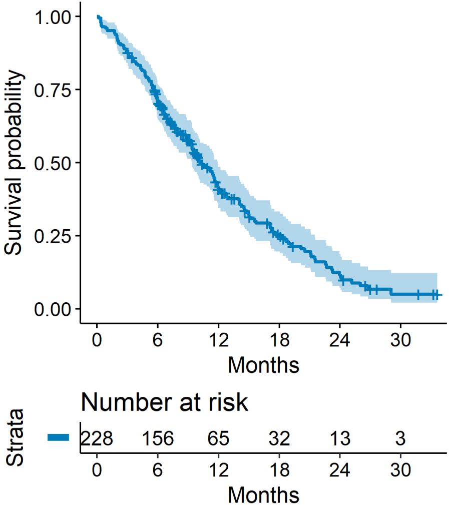
The survival probability at end of study = something < 0.28 (actually 0.005)

Ignoring censoring leads to an overestimate of the overall survival probability



This occurs because the censored subjects only contribute information for **part** of the follow-up time, and then fall out of the risk set, thus pulling down the cumulative probability of survival.

Question from collaborator: What is the 1-year survival probability?

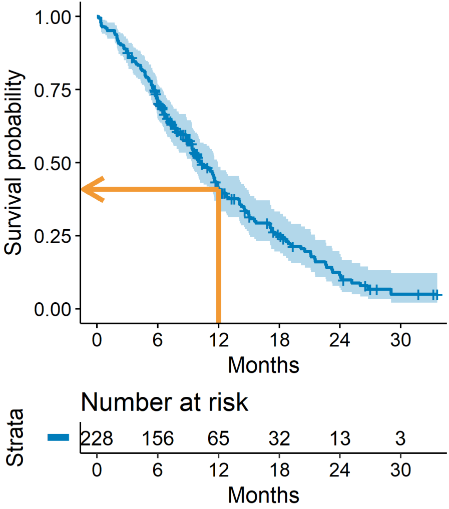


In the `lung` data the 1-year survival probability is 0.41

But where does this number come from?

And what does it mean?

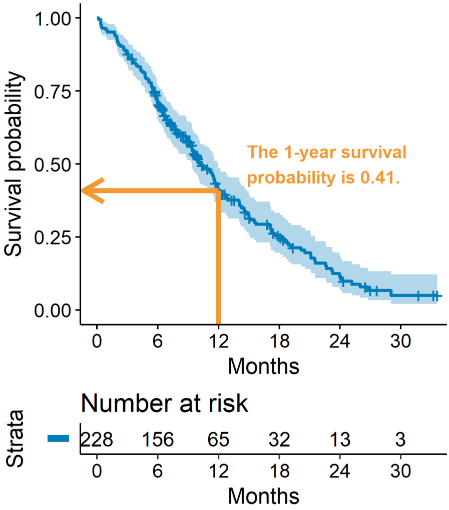
We start by showing what the 1-year survival probability is on the survival curve



The 1-year survival probability is the **probability on the y-axis** corresponding to **1-year on the x-axis**

The `geom_segment` function from `ggplot2` can help with this

Next we annotate the plot with text to state the probability alongside the curve



The `glue` function from the `glue` package provides an easy way to reproducibly annotate plots

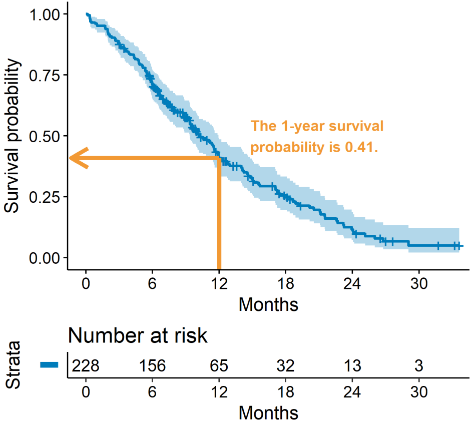
Additionally create reproducible text to make sure numbers are being correctly interpreted

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)}",
                "),\n which represents the estimated proportion of ",
                "patients who\n survived beyond 1 year.")
```

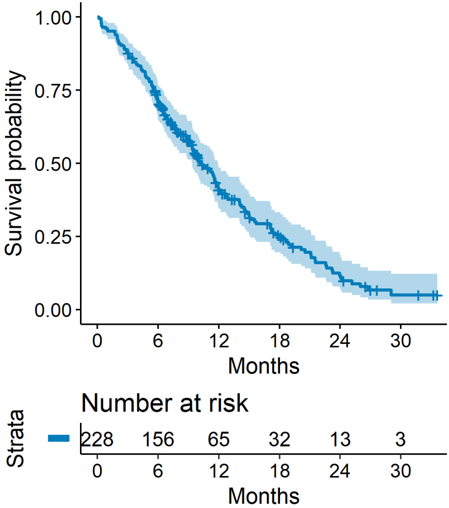
```
## The survival probability at 1-year is 0.41 (95% CI: 0.34 - 0.49),
## which represents the estimated proportion of patients who
## survived beyond 1 year.
```

Finally put everything together into a report to our collaborator about the 1-year survival probability



```
## The survival probability at 1-year is 0.41 (95% CI: 0.34 - 0.49),
## which represents the estimated proportion of patients who
## survived beyond 1 year.
```

Question from collaborator: What is the median survival time?

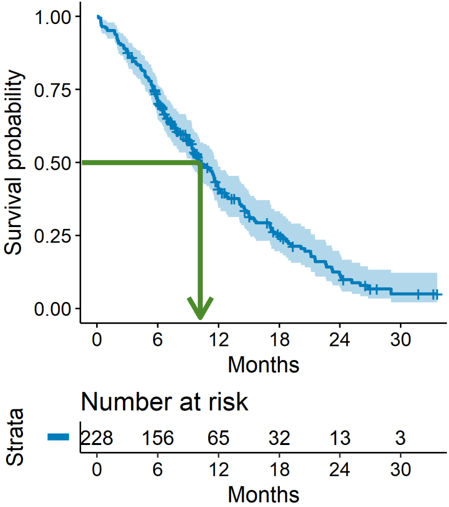


In the `lung` data the median survival time is 310 days

But where does this number come from?

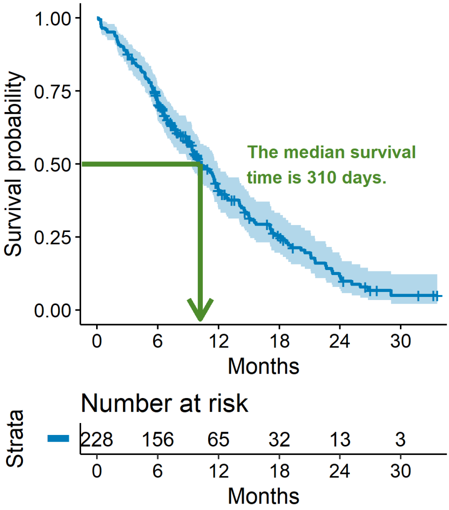
And what does it mean?

We start by showing what the median survival time is on the survival curve



The median survival time is the **time on the x-axis** corresponding to a **survival probability of 0.5 on the y-axis**.

Next we annotate the plot with text to state the time estimate alongside the curve



We again use the `glue` function from the `glue` package to annotate our plot with reproducible text

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

First define the `survfit` `summary` object, using the `survival` package:

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

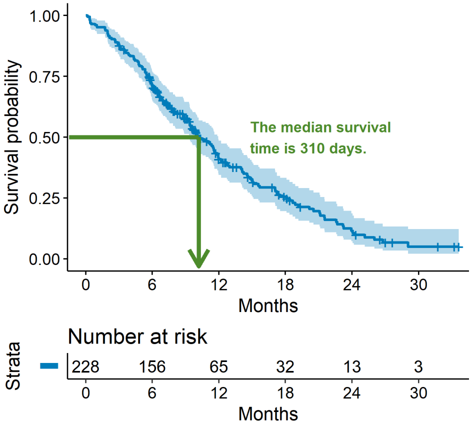
Then:

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

On knitting this will print our desired description:

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

Finally put everything together into a report to our collaborator about the median survival time



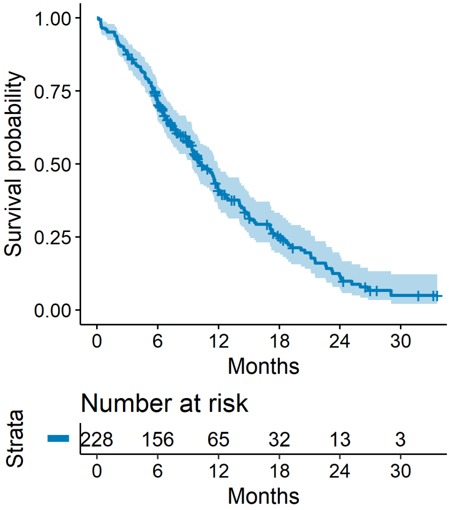
The median survival time is 310 days.

Number at risk

Strata	0	6	12	18	24	30
▬	228	156	65	32	13	3

The median survival time is 310 days (95% CI: 285 - 363), which represents the estimated 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?

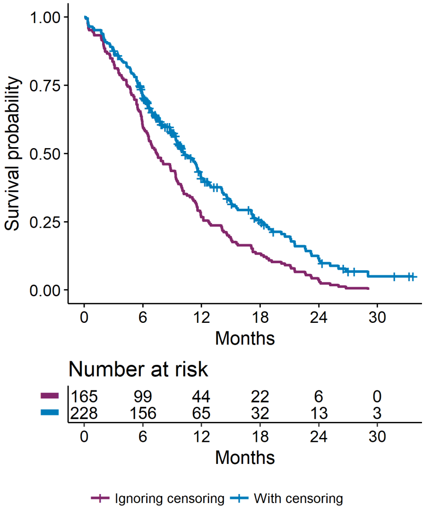


In the `lung` data this would lead to an estimate of median survival time of 226 days.

But this is **incorrect**.

Why?

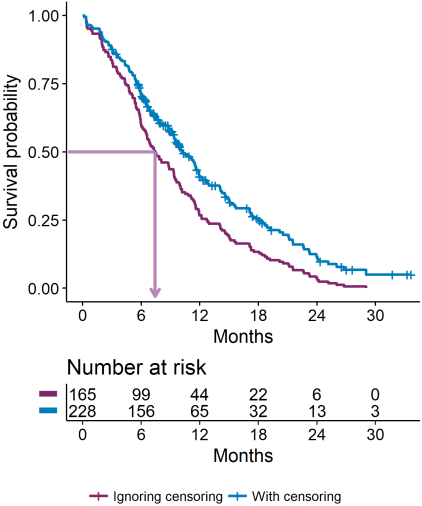
The true survival curve falls above a curve excluding the censored subjects



	0	6	12	18	24	30
🟪	165	99	44	22	6	0
🟦	228	156	65	32	13	3

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

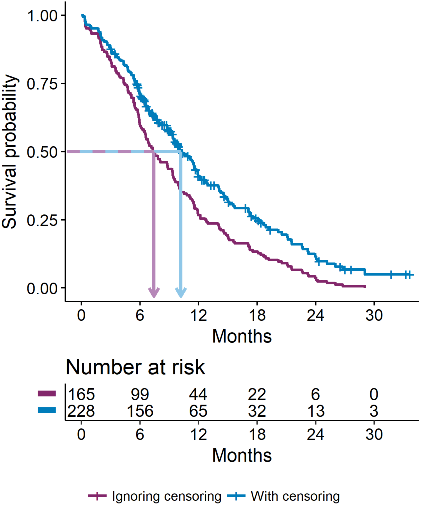
Ignoring censoring leads to an underestimate of the median survival time



The purple group ignores censored subjects.

This is equivalent to the naive estimate of median time to death **among subjects who died** of 226 days

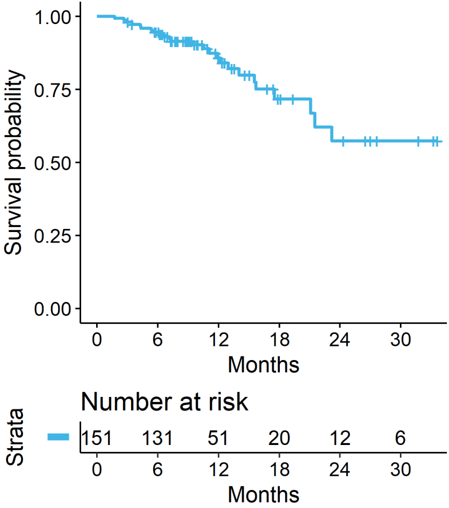
Censored subjects contribute information because we know their event occurred after the censoring time



The blue group includes the censored subjects.

The median survival time is longer, at 310 days.

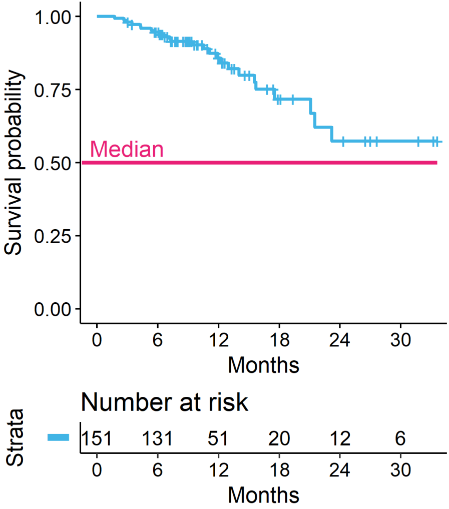
Question from collaborator: What does it mean when you say median survival is not reached?



Survival curve based on a sample of `lung` data subjects

In these data, we would report a median survival time of NA

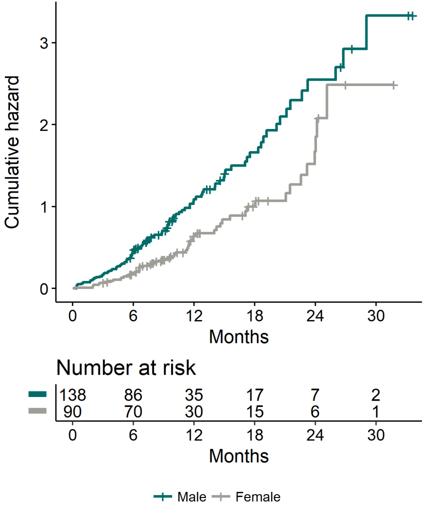
The survival curve must cross the survival probability of 0.5 before median survival time has been reached



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

Question from collaborator: What is a hazard ratio?



Number at risk

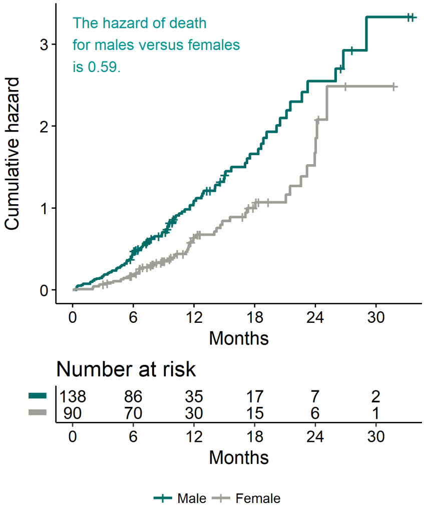
	0	6	12	18	24	30
Male	138	86	35	17	7	2
Female	90	70	30	15	6	1

Months

The hazard ratio (HR) represents the ratio of hazards between two groups at any particular point in time

The hazard of death for males versus females is 0.59 in the `lung` data

Hazard ratios apply at any given point in time, which is what differentiates them from relative risks



The hazard of death for males versus females is 0.59.

Number at risk

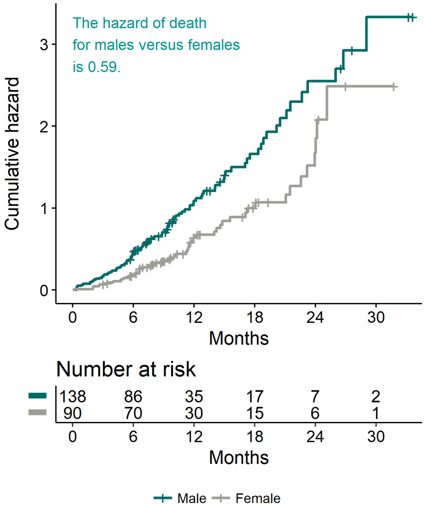
	0	6	12	18	24	30
Male	138	86	35	17	7	2
Female	90	70	30	15	6	1

Months

HR = 0.5 would mean that at any particular time, half as many females as males are dying

So our HR = 0.59 implies that around 0.6 times as many females are dying as males, at any given time

Put it all together into an interpretable, reproducible report for your collaborator



	0	6	12	18	24	30
▬	138	86	35	17	7	2
▬	90	70	30	15	6	1

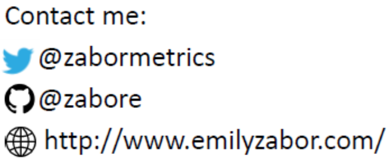
The hazard ratio of 0.59 implies that around 0.6 times as many females are dying as males, at any given time.

Essential 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/