How long will I live? The statistics behind prognosis in cancer research

Emily C. Zabor

New York R Conference, April 21, 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

Survival time, and conversely time to death, is a time-to-event endpoint

Time-to-event endpoints are very common in many contexts, not just cancer

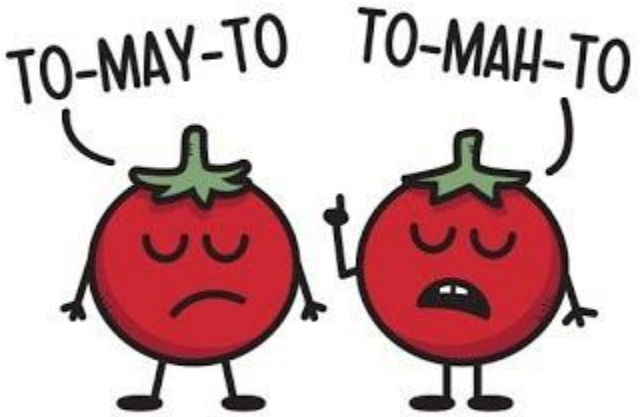
- Time from HIV infection to development of AIDS
- Time to heart attack
- Time to onset of substance abuse
- Time to initiation of sexual activity
- Time to machine malfunction

It is common for time-to-event endpoints to be analyzed incorrectly

You say tomayto, I say tomahto

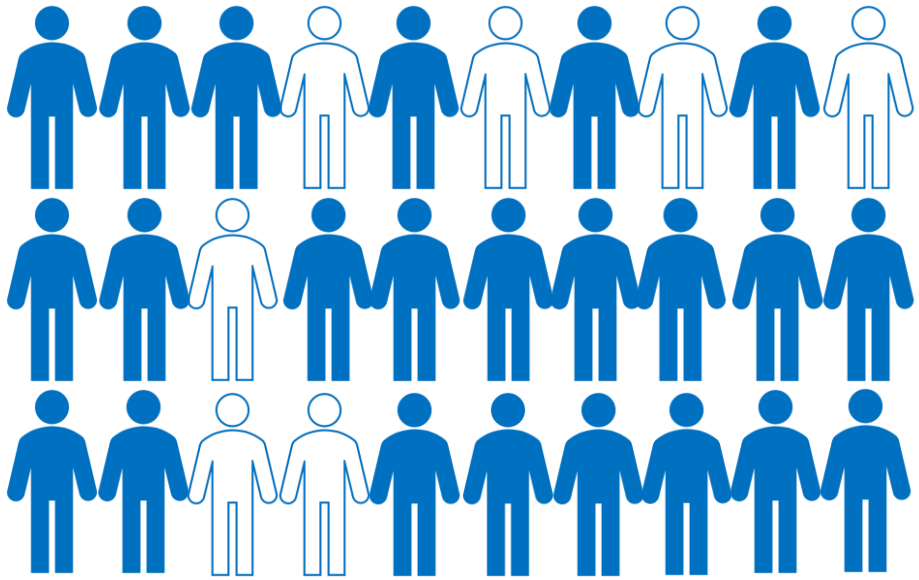
What's called survival analysis in the healthcare field goes by many other names in other fields:

- ✓Reliability analysis
- ✓Duration analysis
- ✓Event history analysis
- ✓Time-to-event analysis



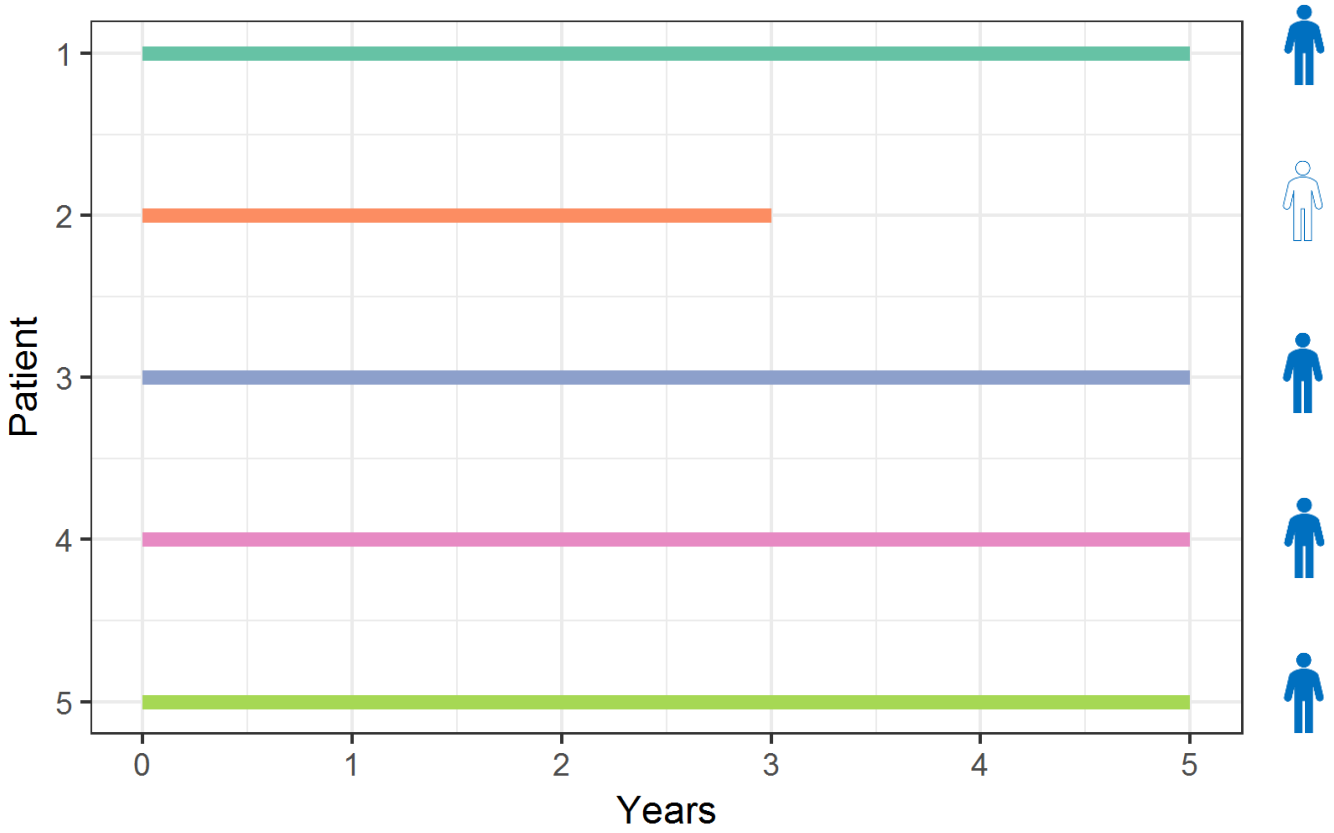
https://www.quora.com/What-do-people-mean-when-they-say-Tomato-Tomato

In cancer research we often want to know the probability of survival and survival time



- What is the probability of survival to a certain number of years?
- What is the average survival time?

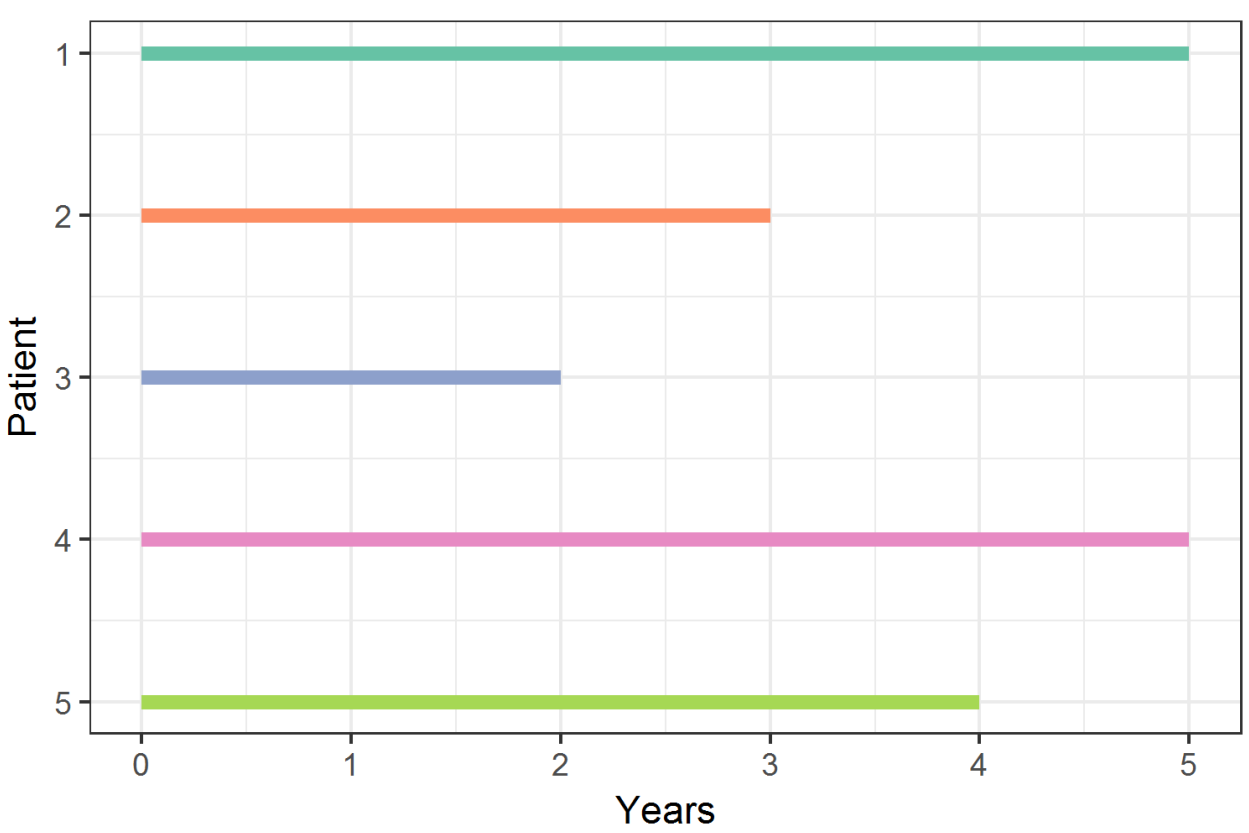
If everyone is followed for a set amount of time, no problems



- All patients followed for 5 years
- 1 patient dead at 3 years
- 4 still alive
- 5-year probability of survival:

$$\left(1 - \frac{1}{5}\right) \times 100 = 80\%$$

If follow-up time is variable, problems

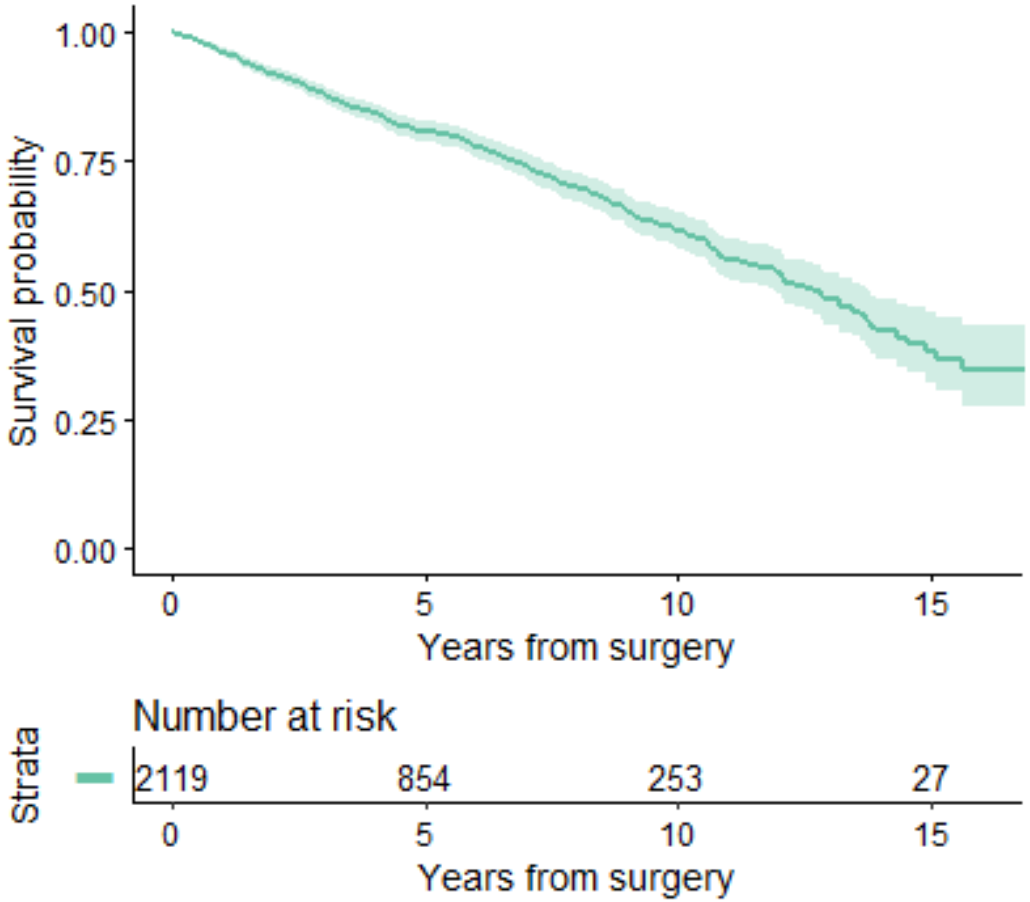


- 2 patients alive at 5 years
- 1 patient dead at 3 years
- 2 still alive, but censored
- 5-year probability of survival:

?

The Kaplan-Meier estimate of survival is appropriate for censored time-to-event data

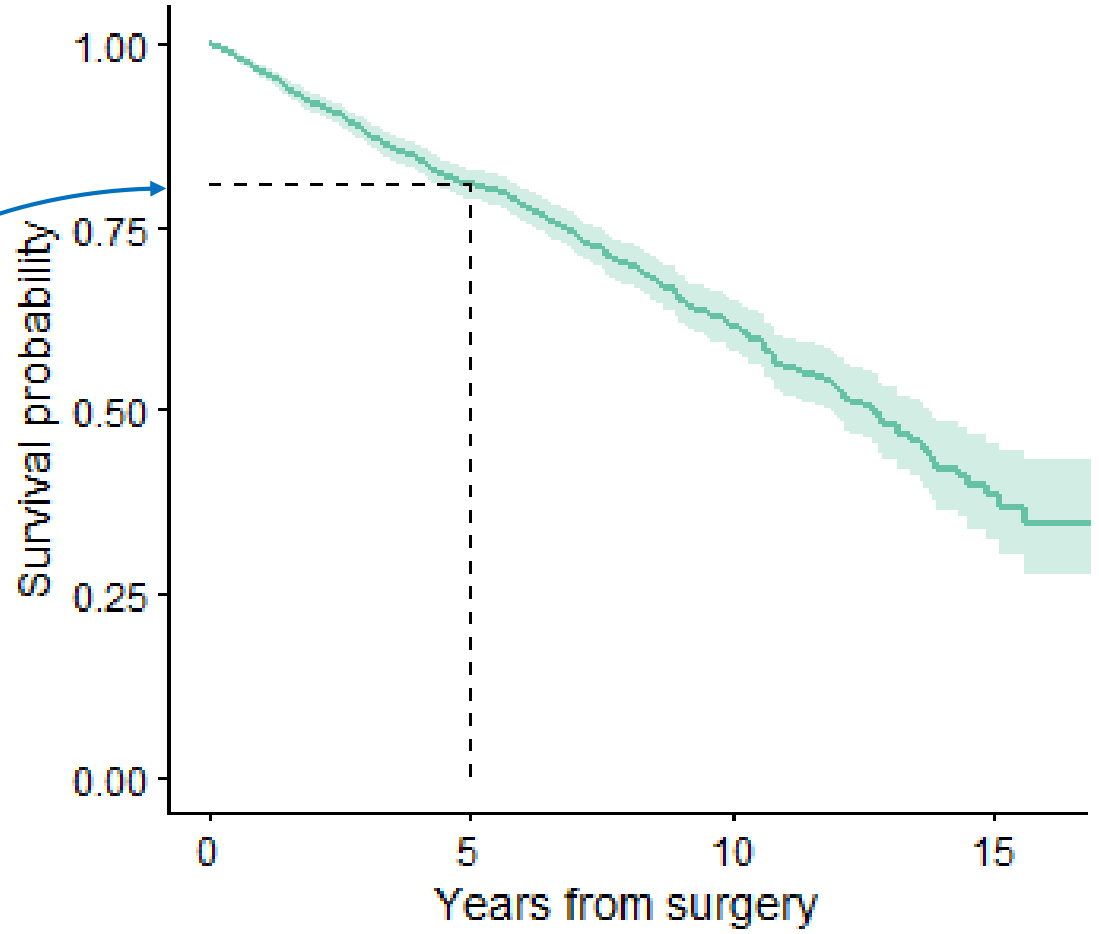
```r
# Load the survival and survminer packages
library(survival)
library(survminer)

# Create the survival curve using survfit
# Plot with ggsurvplot
survfit(Surv(os_yrs, os) ~ 1, data = df) %>%
    ggsurvplot(palette = "Set2",
               censor.shape = "",
               xlab = "Years from surgery",
               xlim = c(0, 16),
               legend = "none",
               risk.table = TRUE,
               risk.table.y.text = FALSE)
```



5-year survival is the survival probability corresponding to 5 years

```
# Use summary to get 5-year probability
summary(survfit(Surv(os_yrs, os) ~ 1,
                data = df), times = 5)
```

5-year probability of survival: 81%



Ignoring censoring leads to an incorrect estimate of the probability of survival

- 2119 patients total
- 297 patients dead at 5 years

INCORRECT

5-year probability of survival:

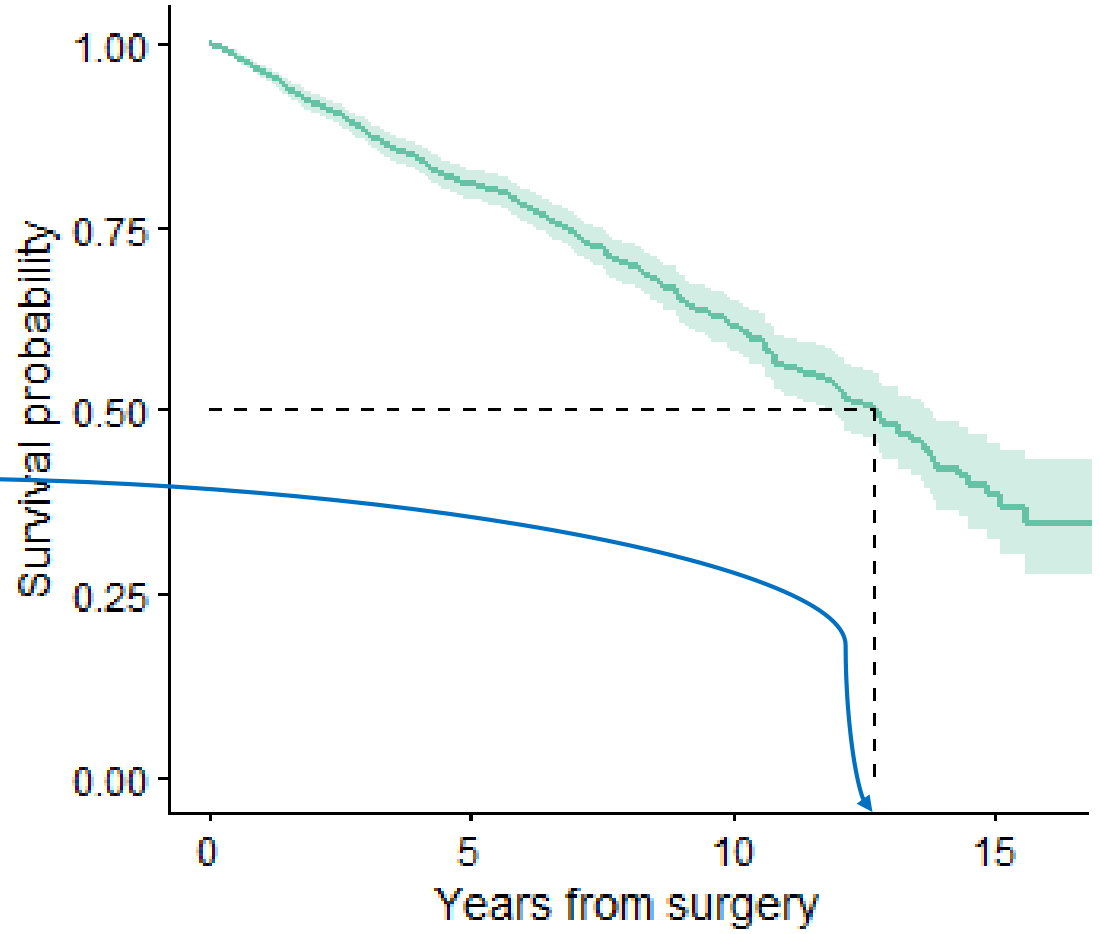
$$\left(1 - \frac{297}{2119}\right) \times 100 = 86\%$$ ✗

*Ignores the fact that **968** patients were censored before 5 years*

Median survival is the time corresponding to a survival probability of 50%

```
# Print survfit object to get median
survfit(Surv(os_yrs, os) ~ 1, data = df)
```

Median survival time: 12.7 years



Ignoring censoring leads to an incorrect estimate of median survival

- 476 patients died

INCORRECT

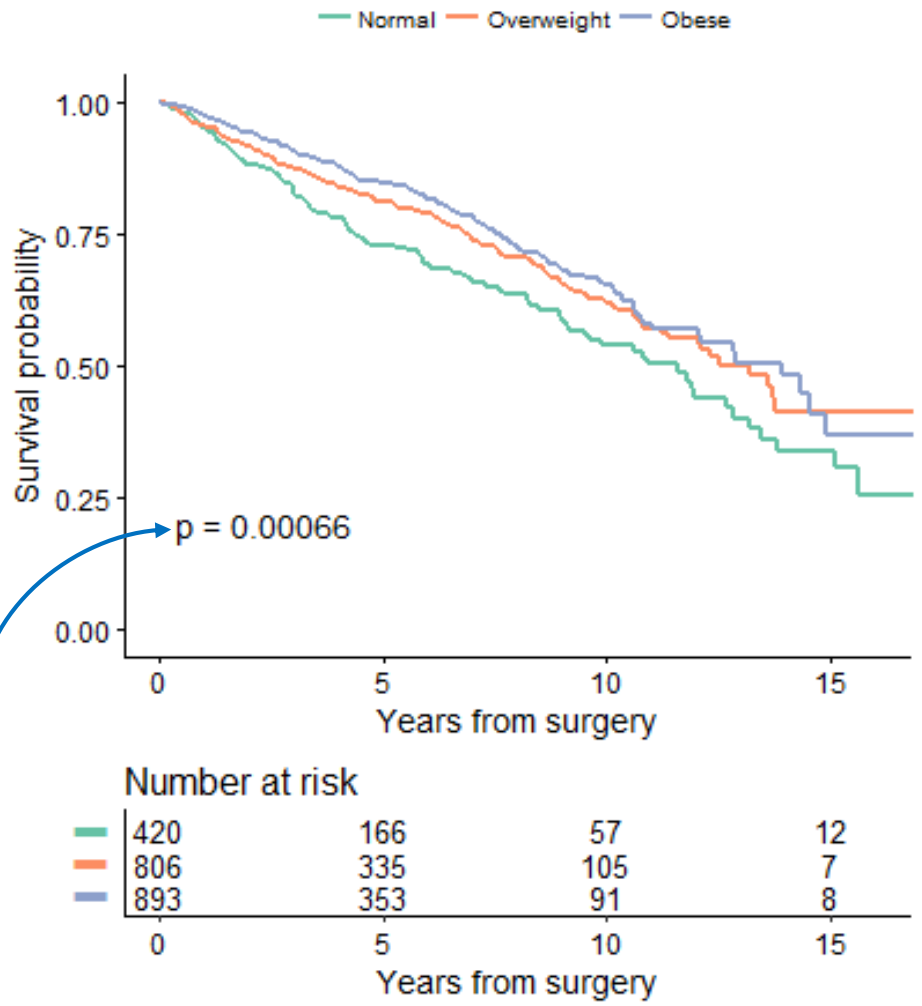
Median survival time among those who died: 3.5 years ✗

*Ignores the fact that **censored** patients also contribute follow-up time*

There are also tests to compare survival between groups

```r
# Add covariate to RHS to get curves by group
survfit(Surv(os_yrs, os) ~ bmi_cat, data = df) %>%
    ggsurvplot(palette = "Set2",
               risk.table = TRUE,
               xlab = "Years from surgery",
               legend.labs = c("Normal",
                               "Overweight",
                               "Obese"),
               legend.title = "",
               pval = TRUE,
               xlim = c(0, 16),
               risk.table.y.text = FALSE)

# Use survdiff for log-rank test
survdiff(Surv(os_yrs, os) ~ bmi_cat, data = df)
```

Log-rank test p-value: <.001



The Cox regression model can be used to fit a semi-parametric model

```
# coxph fits a Cox regression model
coxph(Surv(os_yrs, os) ~ factor(bmi_cat), data = df) %>%
    summary()
```

Factor	HR (95% CI)	p-value
BMI		<.001
Normal	1.00	
Overweight	0.75 (0.60 – 0.94)	
Obese	0.64 (0.51 – 0.81)	

A hazard ratio (HR) > 1 represents an increased hazard of death whereas a HR < 1 represents a reduced hazard of death

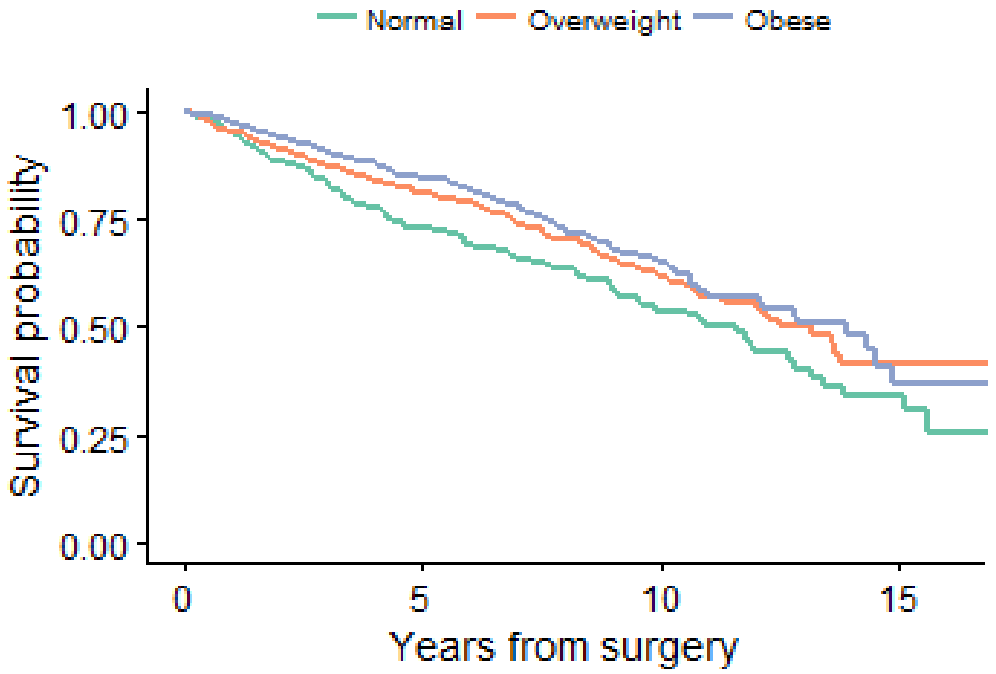
But hazard ratios are difficult to interpret and thus are commonly misunderstood

Commonly MISINTERPRETED as a 100x(1-HR)% reduction in the risk of death

Example:

HR = 0.64

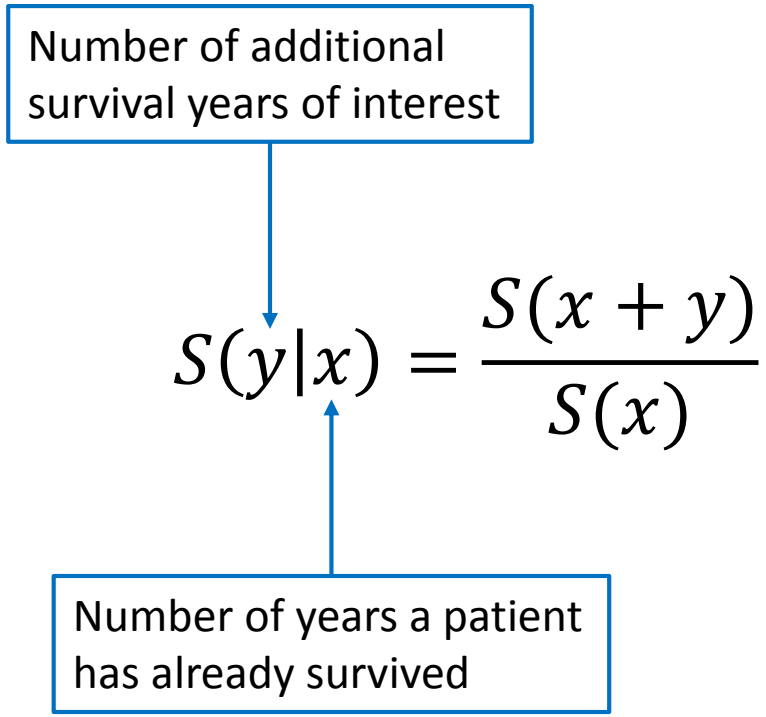
There is a 36% reduced risk of death for obese vs normal weight patients. ✗



Group	5-year OS	10-year OS
Normal	73%	54%
Obese	85%	65%

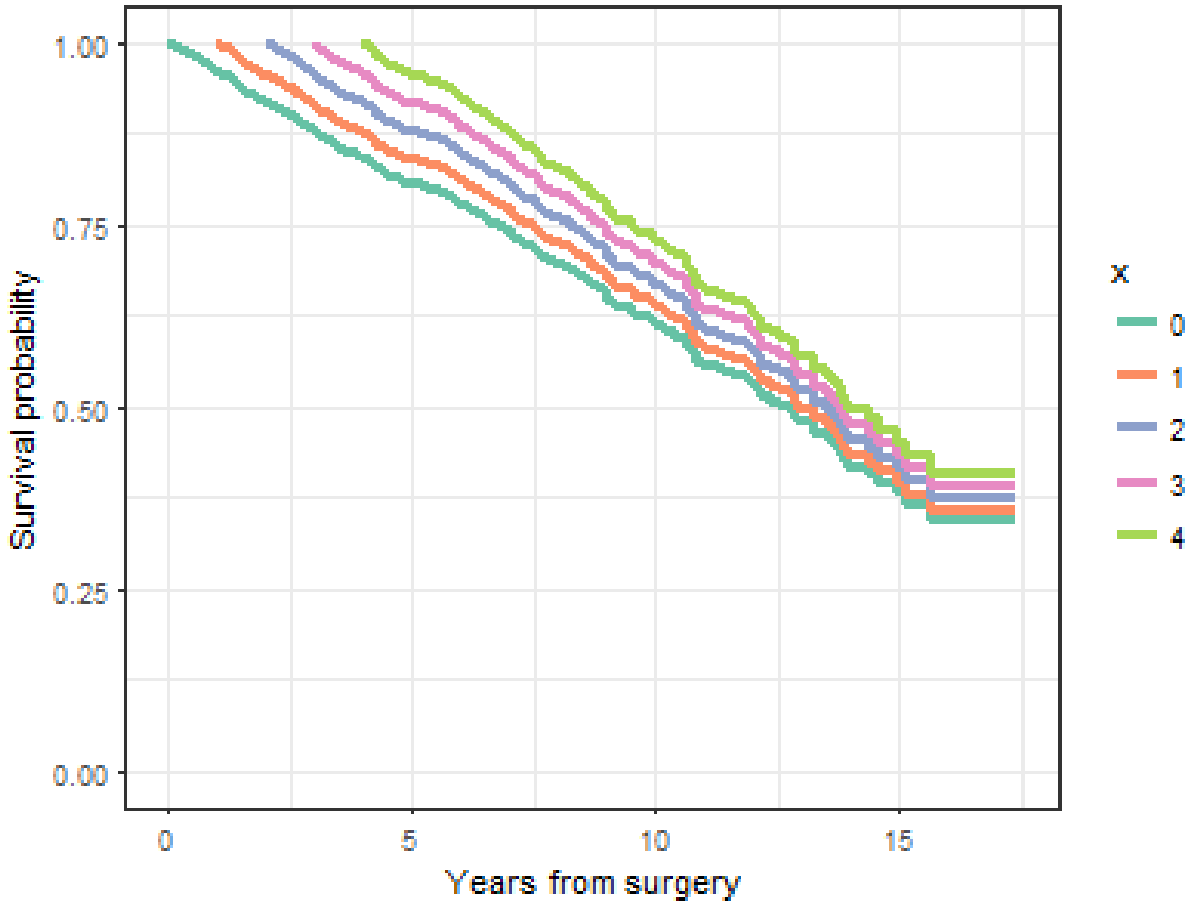
Patients commonly want updated estimates of survival after already living for some years

Does the 5-year probability of survival of 81% still apply to a patient who has already lived for 1 year? 2 years? 5 years?



Conditional survival provides an updated estimate of survival probability

Number of years already survived	Probability of surviving to 5 years
0	81%
1	84%
2	88%
3	92%
4	96%



https://github.com/zabore/condsurv

Code, data, and slides available on GitHub:

https://github.com/zabore/nyr2018

Contact me:

 @zabormetrics

 @zabore

 www.emilyzabor.com

Reference to original publication of the kidney and BMI data:

Hakimi, A. A., Furberg, H., Zabor, E. C., Jacobsen, A., Schultz, N., Ciriello, G., . . . Russo, P. (2013). An epidemiologic and genomic investigation into the obesity paradox in renal cell carcinoma. *J Natl Cancer Inst, 105*(24), 1862-1870.