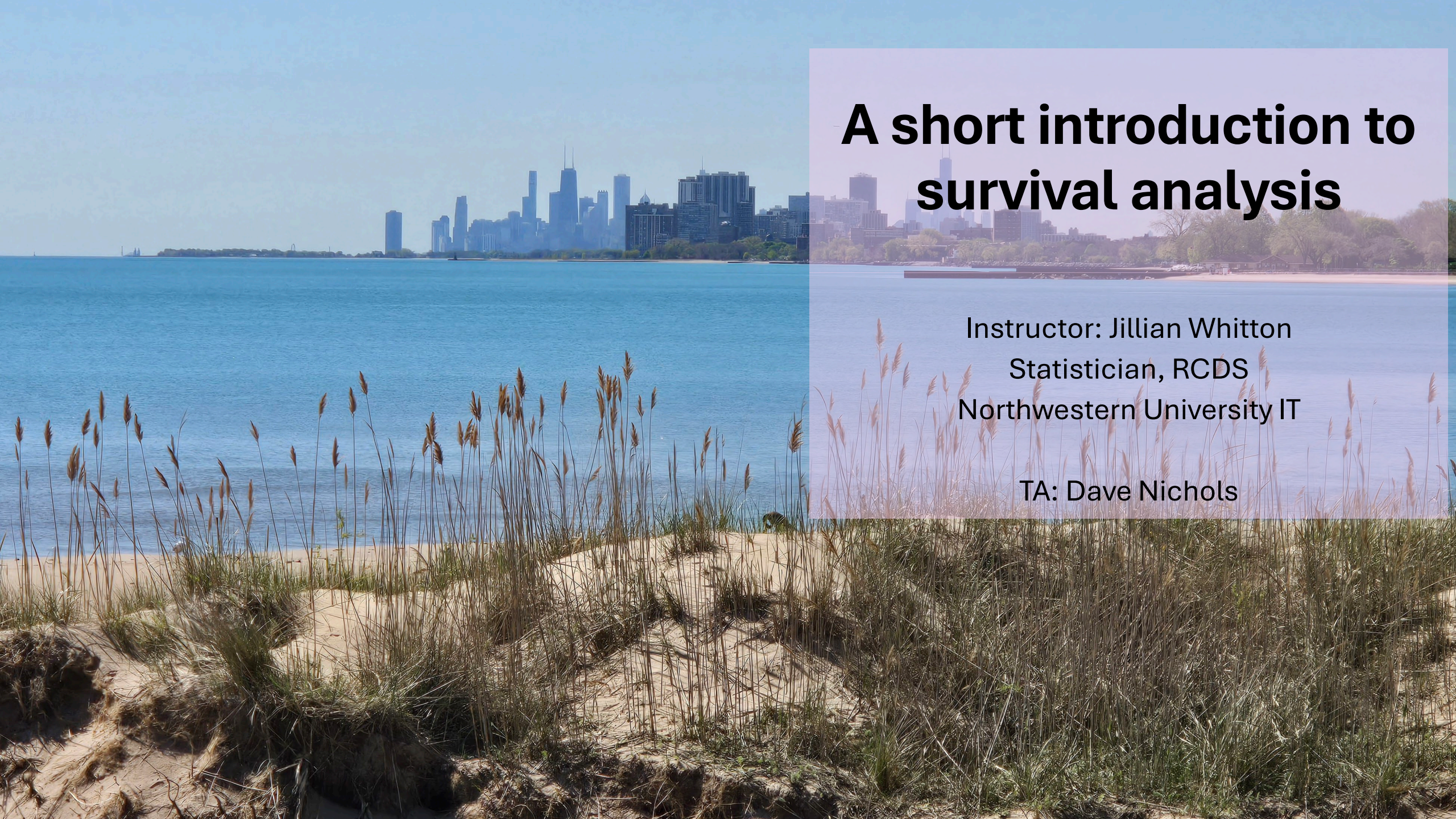


A short introduction to survival analysis

Instructor: Jillian Whitton
Statistician, RCDS
Northwestern University IT

TA: Dave Nichols

This workshop is brought to you by:

Northwestern IT
Research Computing and Data Services

Need help?

- AI, Machine Learning, Data Science
- Statistics
- Visualization
- Collecting web data (scraping, APIs), text analysis, extracting information from text
- Cleaning, transforming, reformatting, and wrangling data
- Automating repetitive research tasks
- Research reproducibility and replicability
- Programming, computing, data management, etc.
- R, Python, SQL, MATLAB, Stata, SPSS, SAS, etc.

Request a **FREE** consultation at **bit.ly/rcdsconsult**.

Logistics

- **Ask Questions** [in the zoom chat].
  - If you know the answer, feel free to respond (we may politely clarify if needed).

- **If my internet goes out.**
  - Take a 5 minute break, and we will meet back in the same zoom room.

Introduction and Goals

- This class will introduce the most common approaches to the analysis of survival data.

- At the end, you should be able to:
  - Identify survival data;
  - Be familiar with the most common ways of analyzing them;
  - Recognize situations when you might need to do something more involved.

- This isn’t a coding class – the techniques we cover can be used in any of the standard statistical software packages: R, SAS, SPSS, etc.
- I used R for the examples. My code is on Github at https://github.com/nuitrcs/stats. Or you can email me at jillian.whitton@northwestern.edu.
- And a reminder: whatever software you use, our consultancy service can help: **bit.ly/rcdsconsult**
- I’ll repeat this information at the end.

Evaluating COVID-19 Vaccine Efficacy Using Kaplan–Meier Survival Analysis

by Waleed Hilal [1], Michael G. Chislett [2,*], Yuandi Wu [1], Brett Snider [3], Edward A. McBean [2], John Yawney [4] and Stephen Andrew Gadsden [1]

Coronary-Artery Bypass Surgery in Patients with Ischemic Cardiomyopathy

Eric J Velazquez [1], Kerry L Lee [1], Robert H Jones [1], Hussein R Al-Khalidi [1], James A Hill [1], Julio A Panza [1], Robert E Michler [1], Robert O Bonow [1], Torsten Doenst [1], Mark C Petrie [1], Jae K Oh [1], Lilin She [1], Vanessa L Moore [1], Patrice Desvigne-Nickens [1], George Sopko [1], Jean L Rouleau [1]; for the STICHES Investigators[1,*]

Time-to-Event Prediction with Neural Networks and Cox Regression

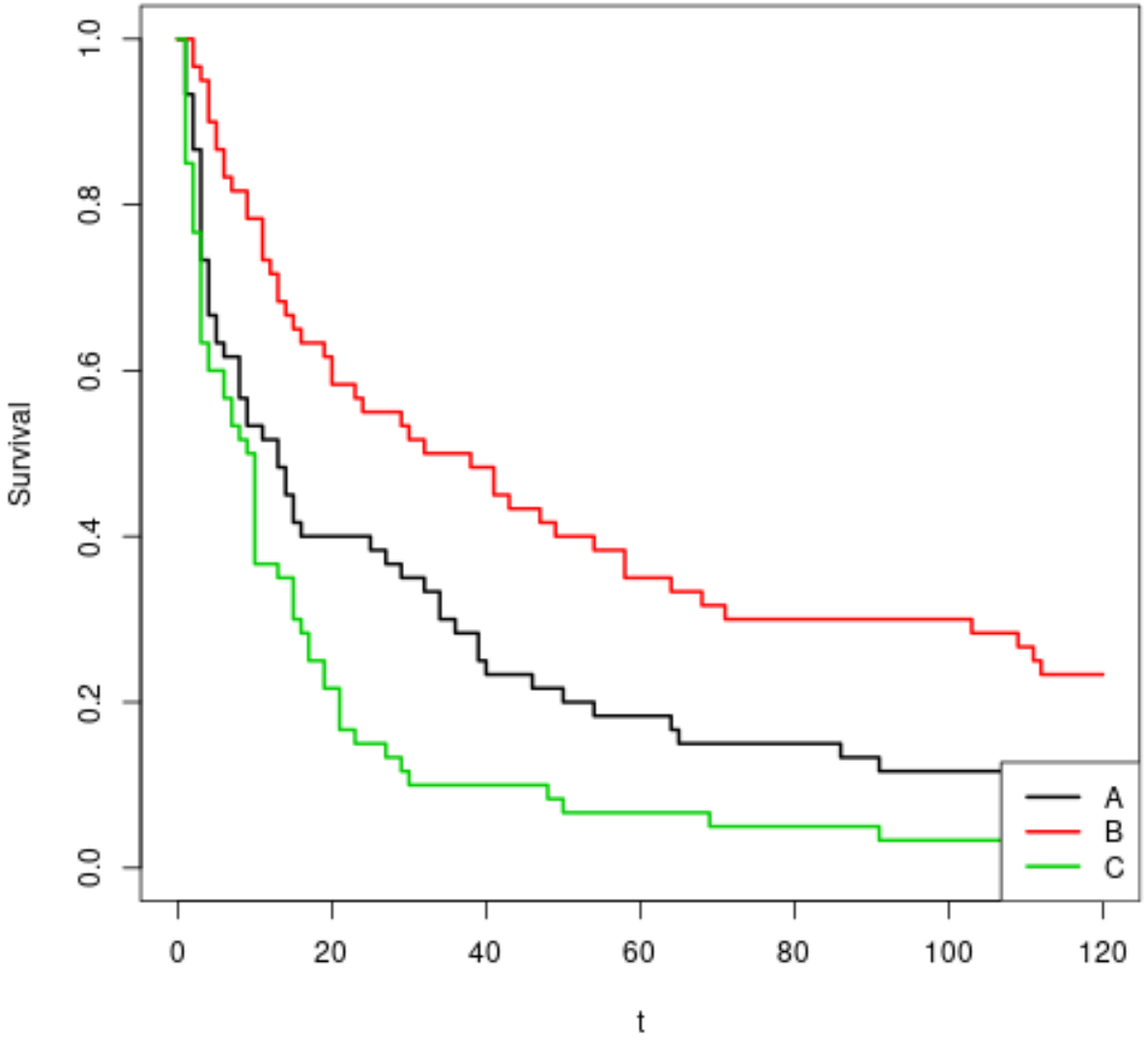
Håvard Kvamme
Ørnulf Borgan
Ida Scheel
Department of Mathematics
University of Oslo
P.O. Box 1053 Blindern
0316 Oslo, Norway

HAAVAKVA@MATH.UIO.NO
BORGAN@MATH.UIO.NO
IDASCH@MATH.UIO.NO

The Survival Analysis for a Predictive Maintenance in Manufacturing

Bahrudin Hrnjica, Selver Softic

Survival data



- Survival data are common.
- We typically have a population at time 0, which reduces in size as time passes.
- Our population may decrease because an outcome has occurred, or simply because of drop-out.
- Common in medicine (survival after diagnosis, disease development after treatment).
- Also common in engineering and industry (equipment failure time).
- And social science: job duration, marriage, housing… (event history analysis)

What do we want to know?

- What distinguishes these data is the *time* component.
- We're interested in time-related outcomes:
  - Describe survival times;
  - Compare survival rates of several groups;
  - Describe the effects of covariates (predictors, either categorical or continuous) on survival rates.

- In many applications, we start with a population at time 0.
- The population remaining in our study decreases as time passes.
- We're interested in the survival function $S(t)$:

$$S(t) = Pr(T > t)$$

the probability of survival beyond time t.

- The population can decrease for two reasons:
  - An event of interest happens.
  - Our observation period comes to an end. If that happens to an individual before the end of a study, it's called censorship.
    - Examples for human subjects include death, not completing follow-up surveys, or no longer meeting study criteria.
- We need to accommodate both events and censored observations.

Are you working with survival data?

Post a brief description in the chat.

A very simple example

- We have a group of 10 patients who we're monitoring for a year after hospitalization. Note that the hospitalizations don't all have to be on the same date.
- We want to know whether they are admitted to hospital again.
- The next slide describes their outcomes:

Patient	# days	Outcome	Event or censorship?		
1	34	Hospitalized	event		
2	57	Hospitalized	event		
3	64	Died	censored		
4	78	Hospitalized	event		
5	138	Moved out of state	censored		
6	186	Died	censored		
7	246	Hospitalized	event		
8	293	Hospitalized	event		
9	365	End of study	censored		
10	365	End of study	censored		

We'll be coming back to this table…

Estimating the survival function

- The survival function for data in our table can be estimated as

$$\hat{S}(t) = \prod_{t \leq t_i} \left(1 - \frac{d_i}{n_i}\right)$$

where d_i is the number of events at time t_i, and n_i is the risk set: the number of individuals at risk (no event or censorship) at that time.

- We don't need specialist software for this. Let's go back to our example.

Patient	# days	Outcome	Event or censorship?	Population at risk	Survival function Π(1-d/n)
1	34	Hospitalized	event	10	9/10 = **0.9**
2	57	Hospitalized	event	9	0.9 * 8/9 = **0.8**
3	64	Died	censored	8	
4	78	Hospitalized	event	7	0.8 * 6/7 = **0.69**
5	138	Moved out of state	censored	6	
6	186	Died	censored	5	
7	246	Hospitalized	event	4	0.69 * 3/4 = **0.51**
8	293	Hospitalized	event	3	0.51 * 2/3 = **0.34**
9	365	End of study	censored	2	
10	365	End of study	censored	2	

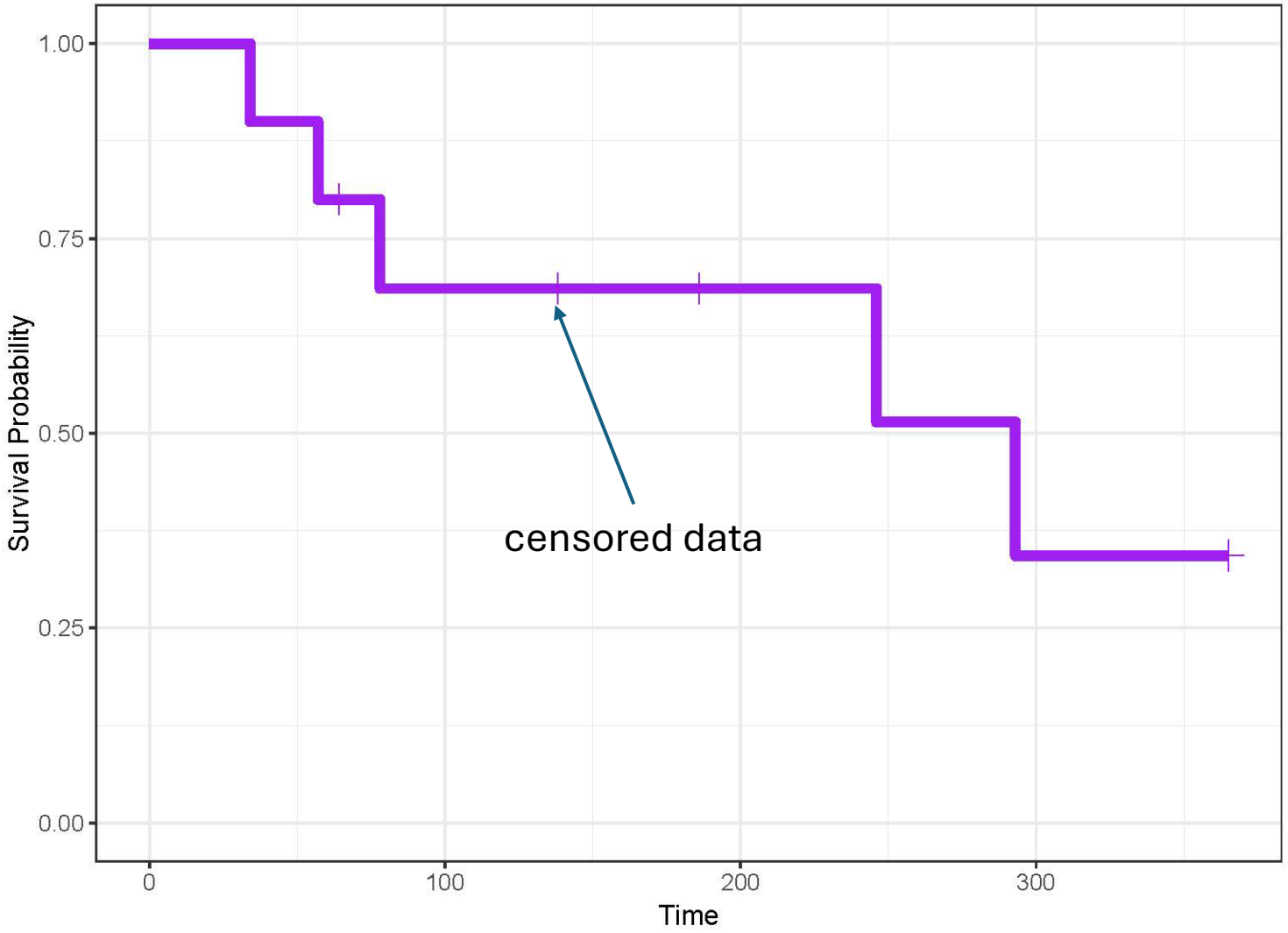
Something else we can do here

Patient	# days	Outcome	Event or censorship?	Population at risk	Survival function Π(1-d/n)
1	34	Hospitalized	event	10	0.9
2	57	Hospitalized	event	9	0.8
3	64	Died	censored	8	
4	78	Hospitalized	event	7	0.69
5	138	Moved out of state	censored	6	
6	186	Died	censored	5	0.51
7	246	Hospitalized	event	4	
8	293	Hospitalized	event	3	0.34
9	365	End of study	censored	2	
10	365	End of study	censored	2	

100 days Survival = 0.69

365 days Survival = 0.34

A plot of our simple example

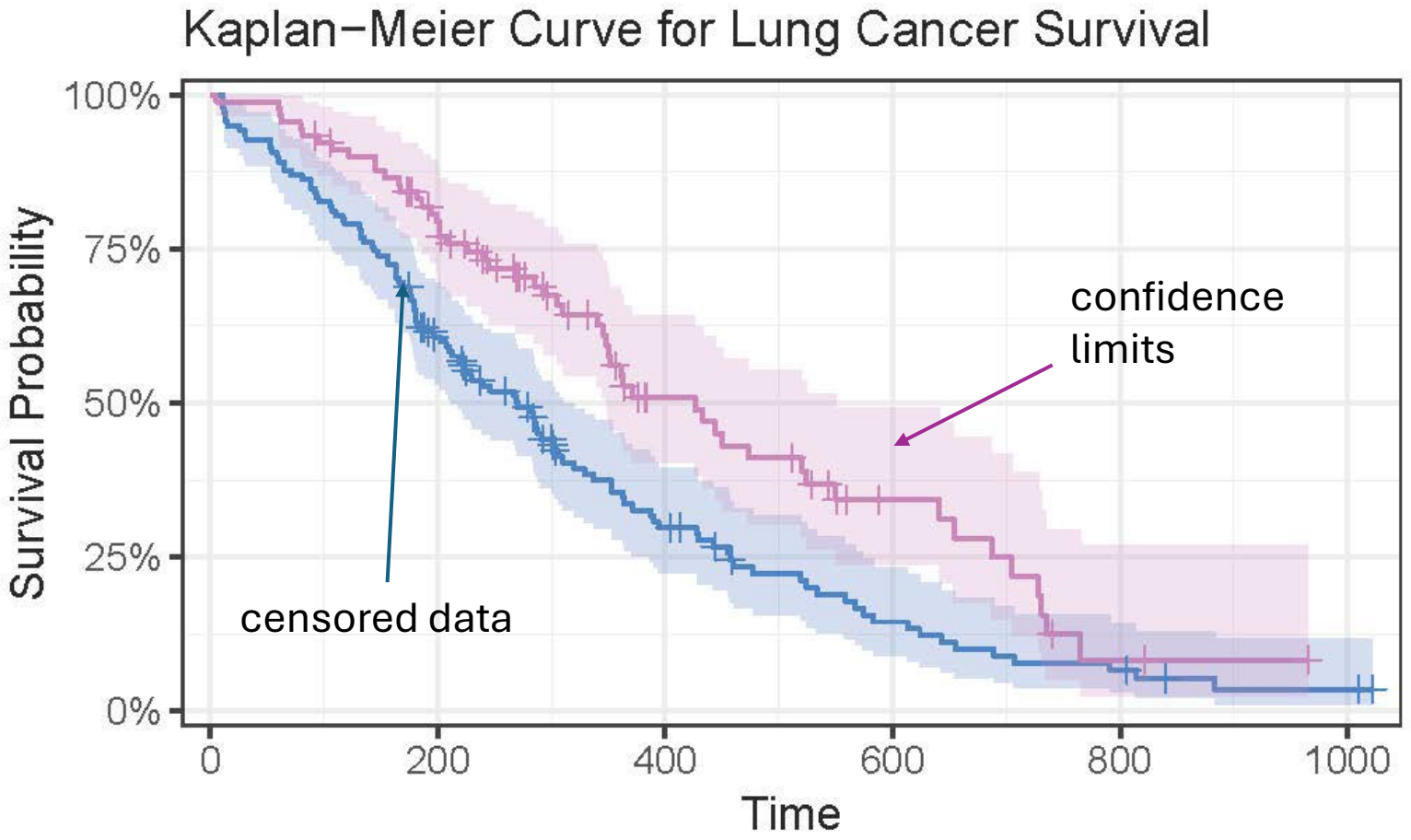


Patient	# days	Outcome	Event or censorship?	Population at risk	Survival function Π(1-d/n)
1	34	Hospitalized	event	10	0.9
2	57	Hospitalized	event	9	0.8
3	64	Died	censored	8	
4	78	Hospitalized	event	7	0.69
5	138	Moved out of state	censored	6	
6	186	Died	censored	5	0.51
7	246	Hospitalized	event	4	
8	293	Hospitalized	event	3	0.34
9	365	End of study	censored	2	
10	365	End of study	censored	2	

Non-parametric (Kaplan-Meier) survival function

- Normally we'd have a lot more observations…
- As we've seen, the estimates are relatively simple to calculate. But we wouldn't usually do them by hand or in a spreadsheet.
- The R package **survival** will generate survival function estimates. We can also add confidence intervals, and log-rank significance tests if we have more than one group. (For example, does survival differ by age group?)
- The R package **ggsurvfit** can be used to plot our data.
- Other statistical software (SPSS, SAS, Stata, etc.) has similar applications.

A more typical plot (real-world data)



- An aside: in this workshop, we're discussing *right-censored* data, where we know the start point (date of manufacture, date of study entry…) but there's no observed end point. This is the most common situation.
- But be aware that *left-censored* data (where we don't know the start point) also sometimes exist, especially in population studies.

What about regression models?

- We can use non-parametric methods to compare survival for different groups, and to get a measure of statistical significance.
- What if we need to estimate the scale of the difference between groups? Can we construct a regression model?
- We can – often, but not always.
- The most common approach is the *Cox proportional-hazards model*.
- Like all regression models, we need to understand when we can and cannot use this. We'll get to that…

The survival function and the hazard function

- Recall the survival function $S(t)$:

$$S(t) = Pr(T > t)$$

  the probability of survival beyond time t.

- Related to this is the hazard function $\lambda(t)$: the instantaneous risk of the event happening at time t, given survival to that point. It's the observed rate, not a probability.
- We can model the effects of our covariates relative to a baseline hazard function $\lambda_0(t)$, at baseline levels of our covariates (such as: lowest age group, no disease, no treatment). This is similar to the intercept in other types of regression models.

Cox regression

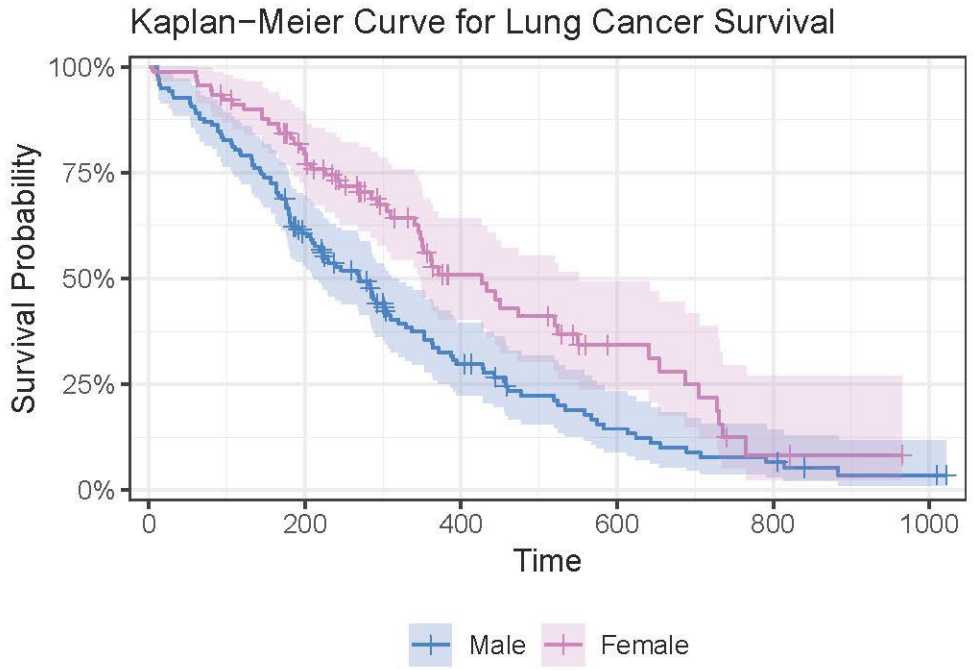
- We assume that our covariates affect the hazard function *proportionally*.
- If this assumption holds, we do not have to worry about the form of the hazard function; we can just work with the proportions.
- If the baseline function is $\lambda_0(t)$, we can model the effects of the covariates x_i with regression parameters β_i as

$$\lambda(t|x_i) = \lambda_0(t)\exp(\beta_1 x_1 + \beta_2 x_2 \ldots + \beta_i x_i)$$

- We can model this, giving us estimates of the regression parameters (which are multiplicative effects on the baseline) with their significance levels.

Cox regression

- All the major statistics/data analysis packages can perform Cox regressions: here I'll be using **survival** and **survminer** in R.
- An example: let's look at the lung cancer patients we used for our Kaplan-Meier plot.



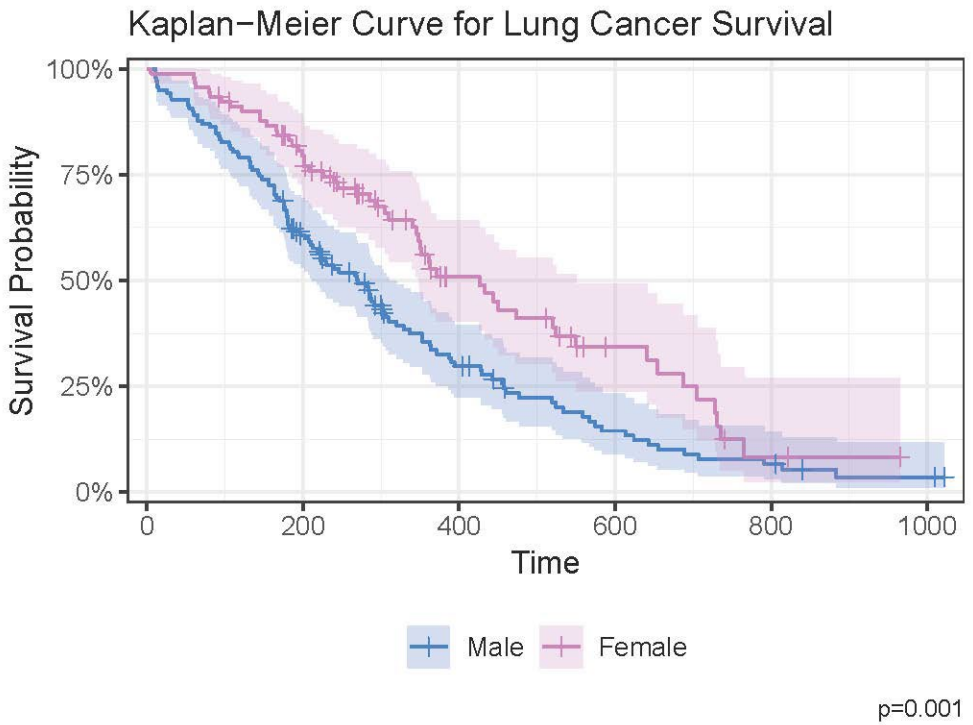
A univariate example

- We can estimate the effect of sex on lung cancer survival with a univariate Cox model.
- Using R, we get the following output:

```
Call:
coxph(formula = Surv(time, status) ~ sex, data = lung)

  n= 228, number of events= 165

       coef exp(coef) se(coef)      z Pr(>|z|)
sex -0.5310    0.5880   0.1672 -3.176  0.00149 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

    exp(coef) exp(-coef) lower .95 upper .95
sex     0.588      1.701    0.4237     0.816
```

- So we estimate that female patients have a hazard rate 0.588 x that of male patients, with $p = 0.001$ (which matches what we saw with Kaplan-Meier).

- But we have to be careful!
- Remember that we are assuming the hazards are proportional.
- Look again at our graph:

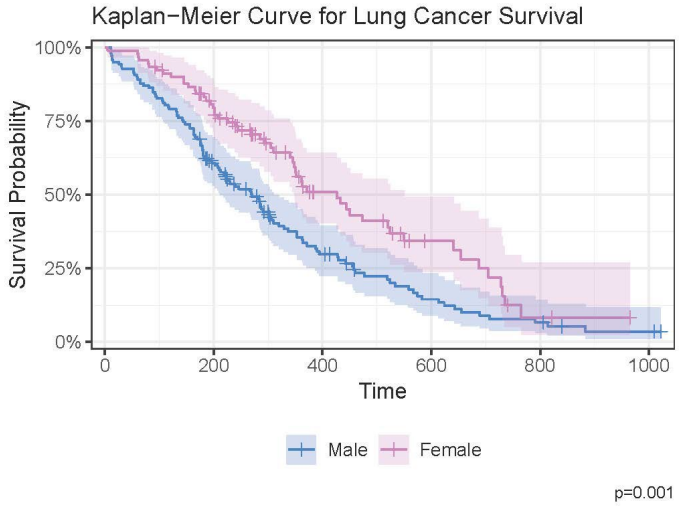


- We should test for proportionality.
- Easy to do in R (the cox.zph function) and also in other packages.

```
        chisq df     p
sex      2.86  1 0.091
GLOBAL   2.86  1 0.091
```

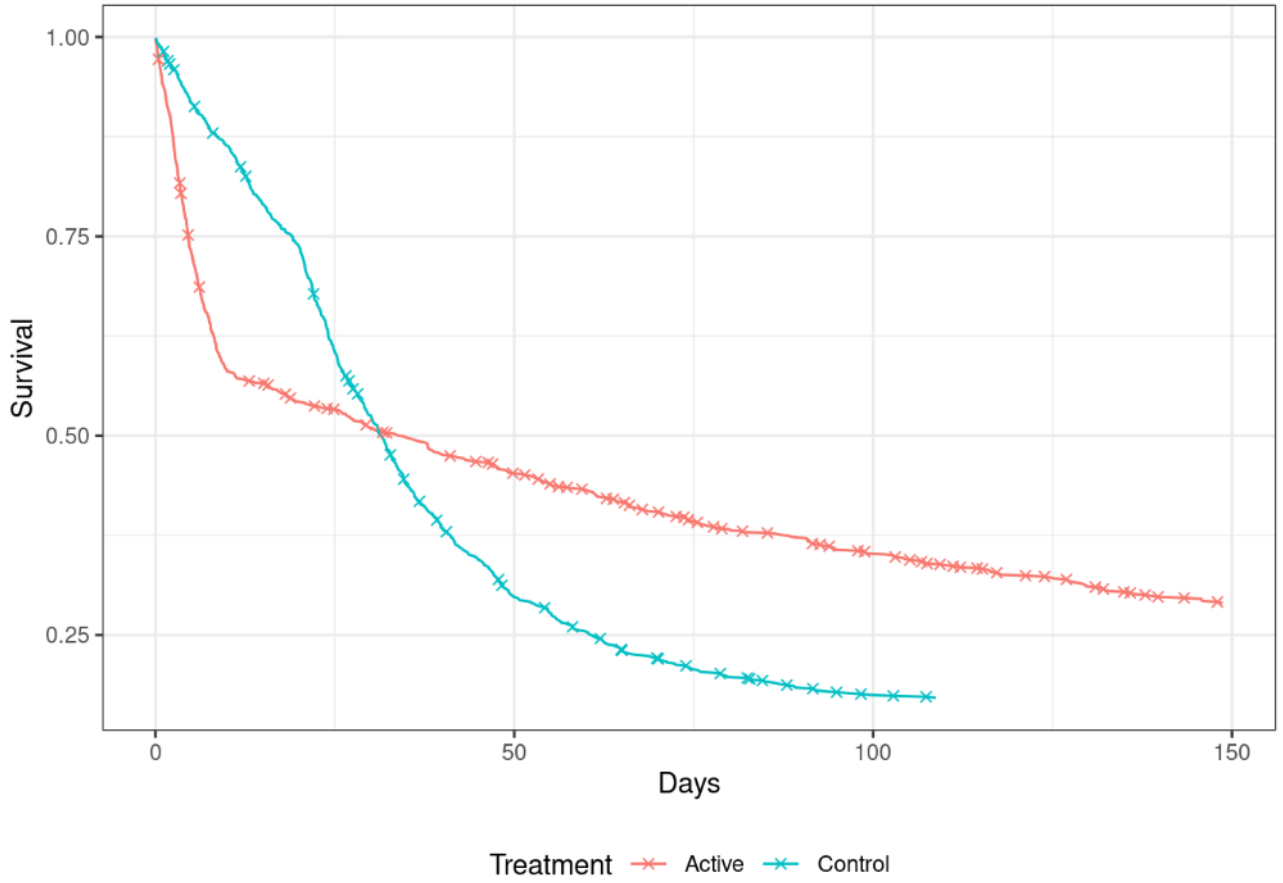
- So we're OK to accept proportionality with our lung-cancer data data ($p < 0.05$ would indicate a problem, if we're using 0.05 as our criterion).

Looking again at our graph: the lines converge toward the right, where data are sparse.



Always view and plot your data before you model them! This applies to every analysis of every kind.

- Sometimes we can look at a survival plot and see that hazards are probably not proportional.
- Take a look at this plot. Do you see any issues?



Multivariable Cox regression

- It's easy to extend this to more than one covariate
- Here's the output using R, for a model of the lung cancer data incorporating age, sex, and weight loss:

```
Call:
coxph(formula = Surv(time, status) ~ age + sex + wt.loss, data = lung)

  n= 214, number of events= 152
   (14 observations deleted due to missingness)

              coef  exp(coef)   se(coef)      z Pr(>|z|)
age       0.0200882  1.0202913  0.0096644  2.079   0.0377 *
sex      -0.5210319  0.5939074  0.1743541 -2.988   0.0028 **
wt.loss   0.0007596  1.0007599  0.0061934  0.123   0.9024
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

         exp(coef) exp(-coef) lower .95 upper .95
age         1.0203     0.9801    1.0011    1.0398
sex         0.5939     1.6838    0.4220    0.8359
wt.loss     1.0008     0.9992    0.9887    1.0130
```

- And again we test for proportionality.
- This is very important with multivariable analyses, where it can be hard to plot all the combinations.

	chisq	df	p
age	0.5077	1	0.48
sex	2.5489	1	0.11
wt.loss	0.0144	1	0.90
GLOBAL	3.0051	3	0.39

- As with any other model-building process, we can improve our model iteratively, while keeping our hypothesis in view.
- So in this case, we might want to drop weight loss from the model, unless weight loss is the thing we're interested in.

- The details are beyond the scope of this workshop, but we can also apply proportional-hazards models to data with time-dependent covariates (such as a risk factor which only applies to the first year after exposure).

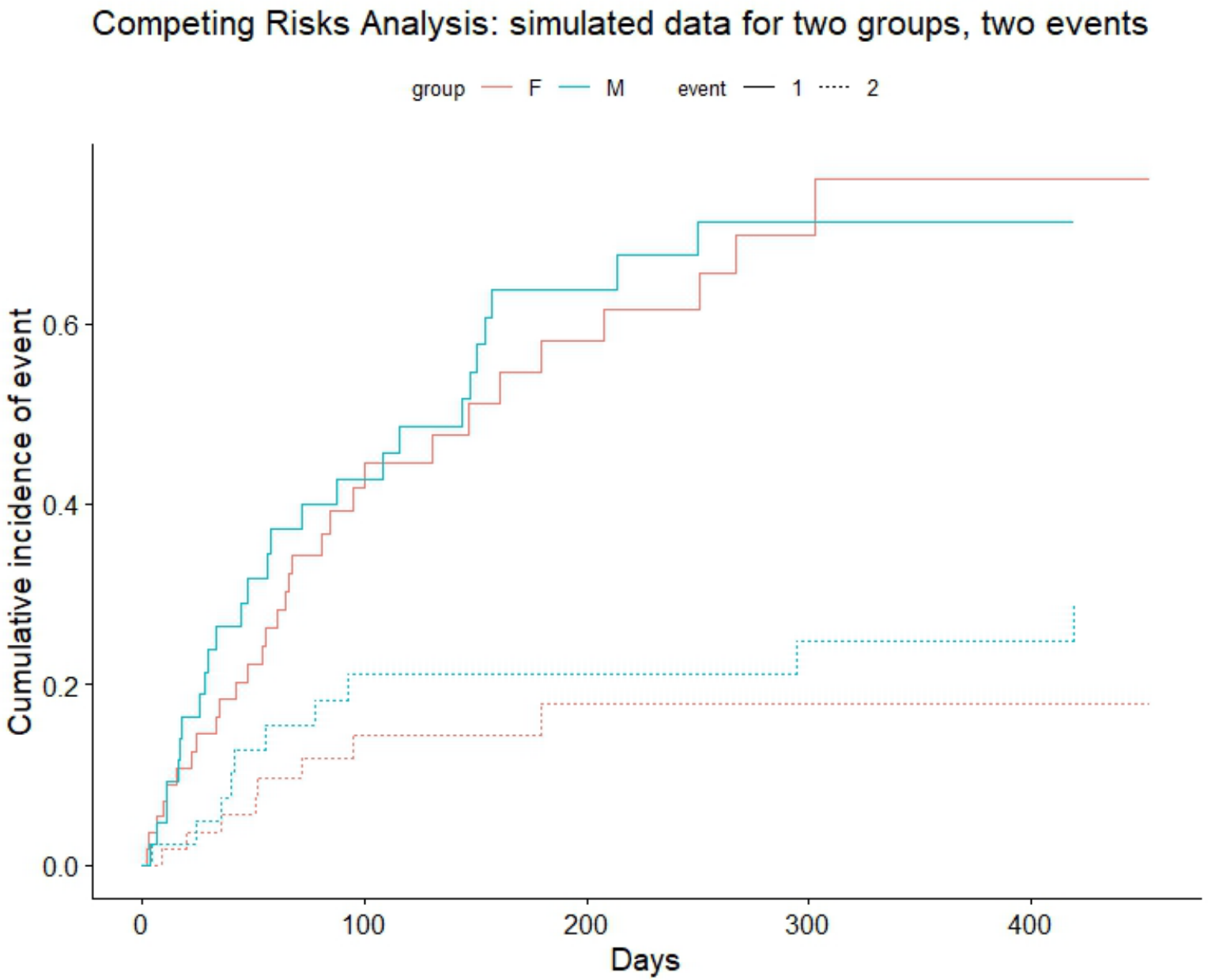
What else?

- Kaplan-Meier estimates and/or Cox regression models will probably meet most of your needs most of the time.
- But they’re not universally applicable!
- There are a few other approaches you should be aware of. Here’s a (very) brief overview of a couple of them:

Competing risks

- We can be more flexible in our survival analysis by considering more than one event. For example, we might want to consider death from disease, and death from other causes.
- In a situation like this, “death from other causes” might be considered censoring, or it might be a second adverse outcome. It depends on what you’re trying to do.
- And there can sometimes be several competing risks: death from breast cancer vs. death from ovarian cancer vs. death from other cancer vs. other causes of death. (And censoring.)
- Again, there’s a lot of software available.

- The Kaplan-Meier approach to survival functions doesn't work for competing risks.
- Instead, we can work with the *cumulative incidence function*: the marginal probability of an event as a function of its cause-specific probability and the overall survival probability.
- Superficially, it looks like a "backward" survival function – starting at 0 and incrementing at each event.
- Here's an example, using simulated data and the R package **cmprsk**:


Competing Risks Analysis: simulated data for two groups, two events
group F M
event 1 2
Cumulative incidence of event
0.0
0.2
0.4
0.6
0
100
200
300
400
Days

Survival trees

- An alternative to regression models for multivariable data. It's useful if we want to identify high-risk groups.
- Use machine-learning techniques to partition a large dataset into subgroups with different survival rates, based on the levels of different variables.
- Each branch of the tree is a population with different characteristics, and we can, if we wish, draw a Kaplan-Meier plot for each one.
- We can generate a single tree or multiple versions from the same data (a survival forest).

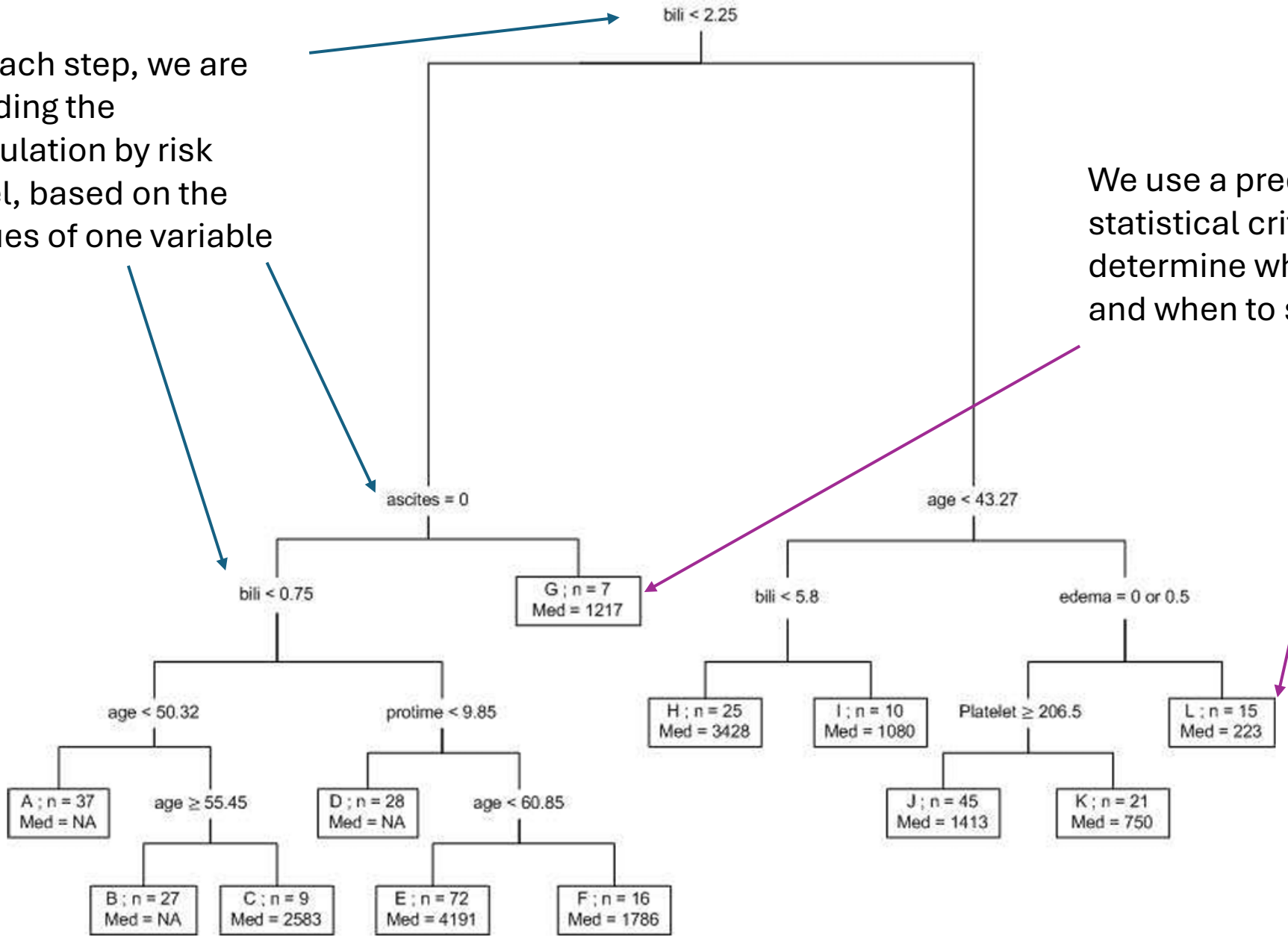
- The next slide shows a survival tree for a clinical trial of a drug for the treatment of cirrhosis of the liver.
- Variables:

1. Drug: 1=D–penicillamine, 0=placebo.
2. Age: age in years.
3. Sex: 0=male, 1-female.
4. Ascites: presence of ascites (0=no, 1=yes).
5. Hepatom: presence of hepatomegaly (0=no, 1=yes).
6. Spiders: presence of spiders (0=no, 1=yes).
7. Edema: presence of edema (0=no edema and no diuretic therapy for edema; 0.5=edema present for which no diuretic therapy was given, or edema resolved with diuretic therapy; 1=edema despite diuretic therapy).
8. Bili: Serum bilirubin, in mg/dl.
9. Albumin: in gm/dl.
10. Alkphos: alkaline phosphatase, in U/liter.
11. Platelet: platelet count, in number of platelets per–cubic–milliliter of blood divided by 1000.
12. Protime: prothrombin time, in seconds.

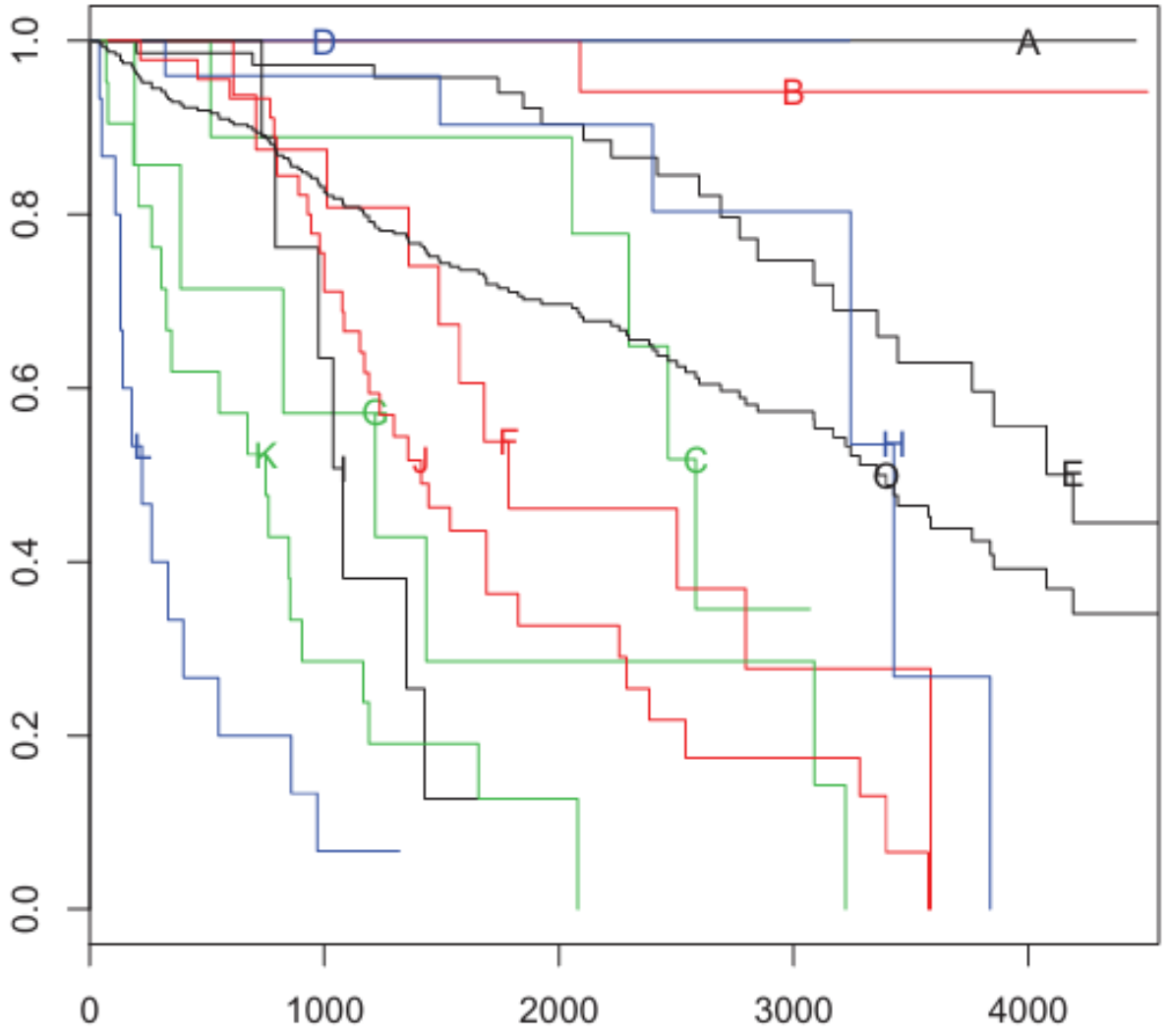
- Not all of these turn out to be important for modeling.

At each step, we are dividing the population by risk level, based on the values of one variable

We use a predetermined statistical criterion to determine when to split and when to stop



... and here are the survival plots for all those end points.



What's the point?

- We haven't built a traditional statistical regression model here, but this approach can be very useful.
- We've identified sectors of the population whose characteristics give them a higher risk.
- This can be useful for clinicians treating patients. In another context, it can be used by engineers to identify components at higher risk of damage.

Wrapping up

- Kaplan-Meier analyses and Cox regression models are both powerful and analytical approaches. They may be all you need much of the time – but be aware of their limitations.
- There are several other techniques which may be applicable to your data.
- If in doubt, talk to a statistician!
- RCDS free consultation service: **bit.ly/rcdsconsult**
- Questions, feedback, or for a copy of my R code: jillian.whitton@northwestern.edu. My code is also at https://github.com/nuitrcs/stats.

- Thanks for attending!
- I’m happy to answer questions – or you can email me jillian.whitton@northwestern.edu