Wrangling Survival Data

From Time-dependent Covariates to Multistate Endpoints

Beth Atkinson September 12, 2019

MAYO CLINIC

Outline

- Time to event - one observation per subject
- Start/Stop data
  - Why needed?
  - New tools: `tmerge`, `survSplit`
  - Check data: `survcheck`
  - Common mistakes
- Multistate data
  - Competing risk

MAYO CLINIC

Basics

Logistics

- All code shown based on the latest/greatest version of the `survival` package (3.0)
- Slides/Example code available at `https://github.com/bethatkinson/rmed2019_surv`
  - Examples loaded into RStudio Cloud - `https://rstudio.cloud/project/475200`
- Email: `atkinson@mayo.edu`

MAYO CLINIC

Background

- I am a statistician working in medical research
- Many of the questions I work with are "time until . . . "
  - Fracture
  - Diagnosis of a chronic comorbidity
  - Liver transplant
  - Death
  - . . .
- I study osteoporosis in population-based cohorts, so many of my examples deal with fractures
- I started off using Splus in 1990 so my code is a mix of base R and tidyverse

MAYO CLINIC

Premise

Most statistics discussions focus on the analysis and assume the data is already in shape. The reality is that:

- Data wrangling takes much of the time
- Doing it correctly is critical
- . . . so that's what I'll talk about

MAYO CLINIC

Some principles of data creation

- Correct is more important than fast: Don't worry if the code takes a bit to run. We often do dozens of fits using one dataset
- Correct is more important than clever
- Readable is more important than short
- Use every data check opportunity available
- Comments are your friend, or better yet make the data creation an Rmd file with text explaining the code

Key Principle

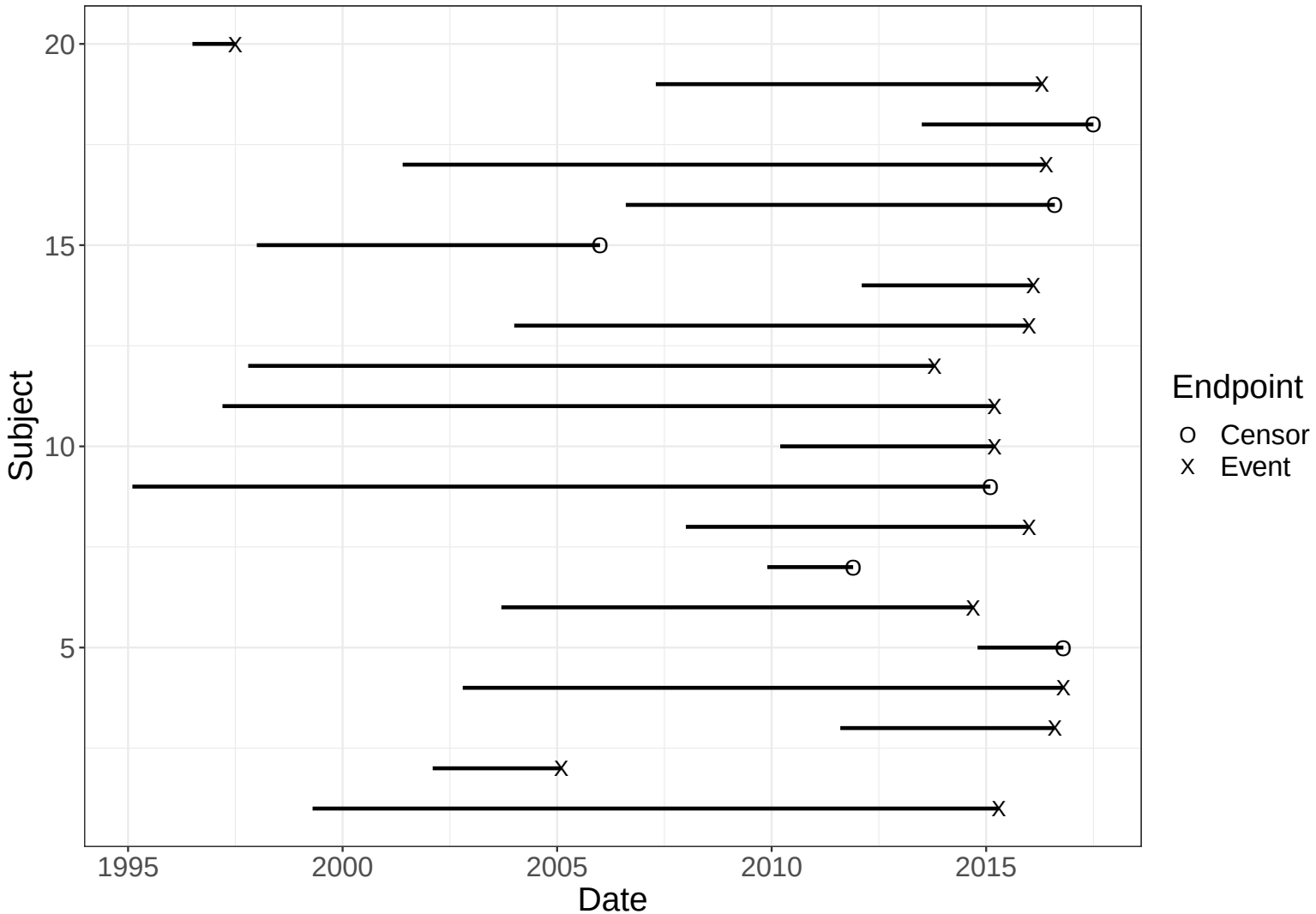
"It takes time to observe time"

Challenges:

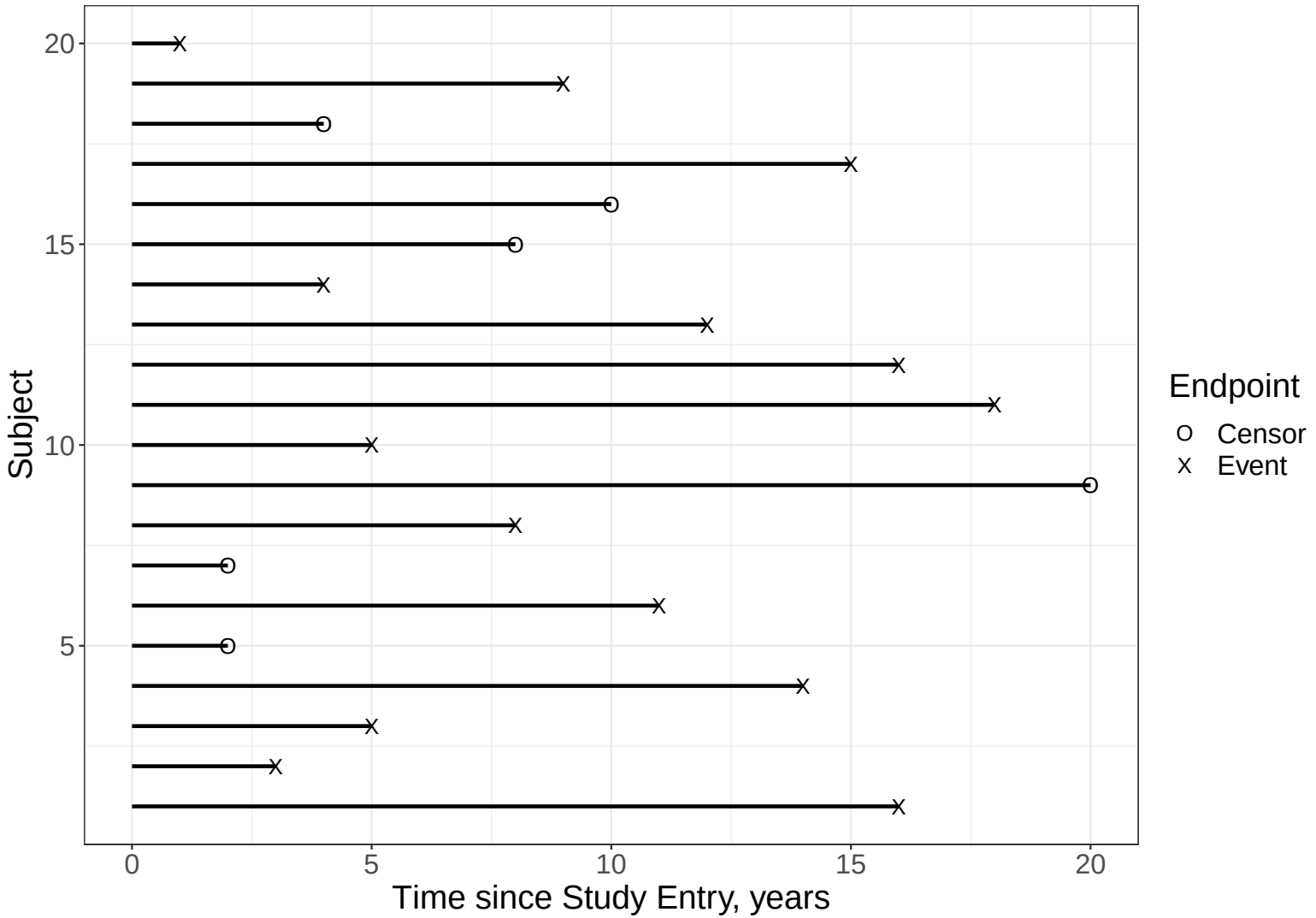
- Incomplete information (*censoring*). At the time of an analysis, not everyone will have yet had the event.
- Dated results.
  - In order to report 5 year survival, from a treatment, patients need to be enrolled and then followed for 5+ years.
  - By the time recruitment and follow-up is finished, the final report on the treatment might be 8 years old and considered out of date.

MAYO CLINIC

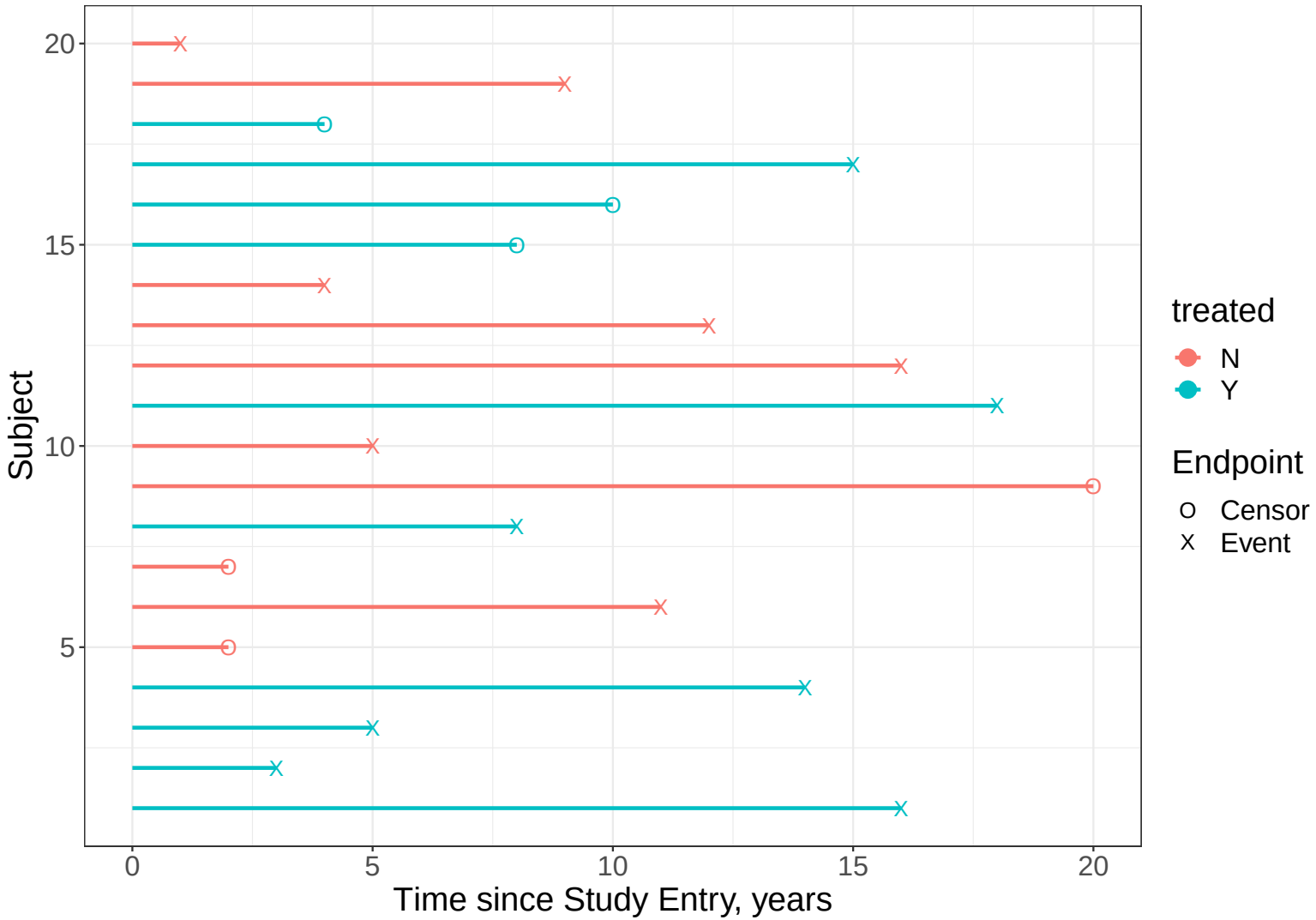
Calendar Year Scale



Time from Study Entry Scale



Compare Baseline Treatment Groups



Example: AML data

Does maintainance of the standard course of chemotherapy improve survival for patients with Acute Myelogeous Leukemia?

```
> dim(aml)
[1] 23  3
> head(aml)
  time status           x
1    9      1 Maintained
2   13      1 Maintained
3   13      0 Maintained
4   18      1 Maintained
5   23      1 Maintained
6   28      0 Maintained
```

MAYO CLINIC

Create endpoint in `survival` package

A time-to-event outcome consists of 2 pieces of information:

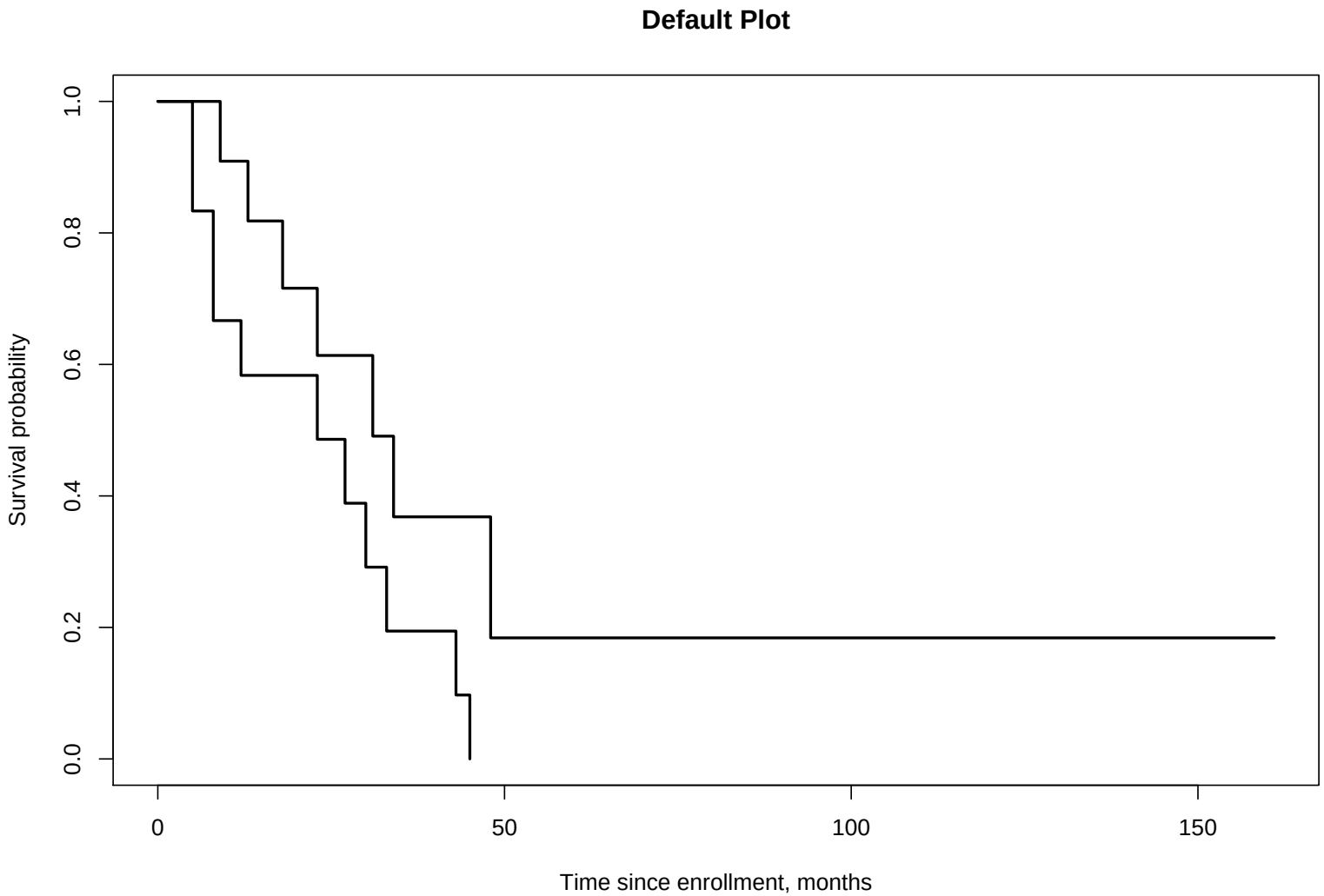
- Length of time over which the patient was observed
- Presence/absence of the event at the end of the time period
  - 0=censor/1=event
  - FALSE=censor/TRUE=event
  - 1=censor/2=event

```
> with(aml, Surv(time=time, event=status))[1:6]
[1]  9   13   13+ 18   23   28+
> aml$status[1:6]
[1] 1 1 0 1 1 0
```

Kaplan-Meier Curves: default

```
> fit <- survfit(Surv(time, status) ~ x, data=aml)
> print(fit)
Call: survfit(formula = Surv(time, status) ~ x, data = aml)

                 n events median 0.95LCL 0.95UCL
x=Maintained    11      7     31      18      NA
x=Nonmaintained 12     11     23       8      NA
>
> plot(fit,
      xlab='Time since enrollment, months',
      ylab='Survival probability')
```

MAYO CLINIC

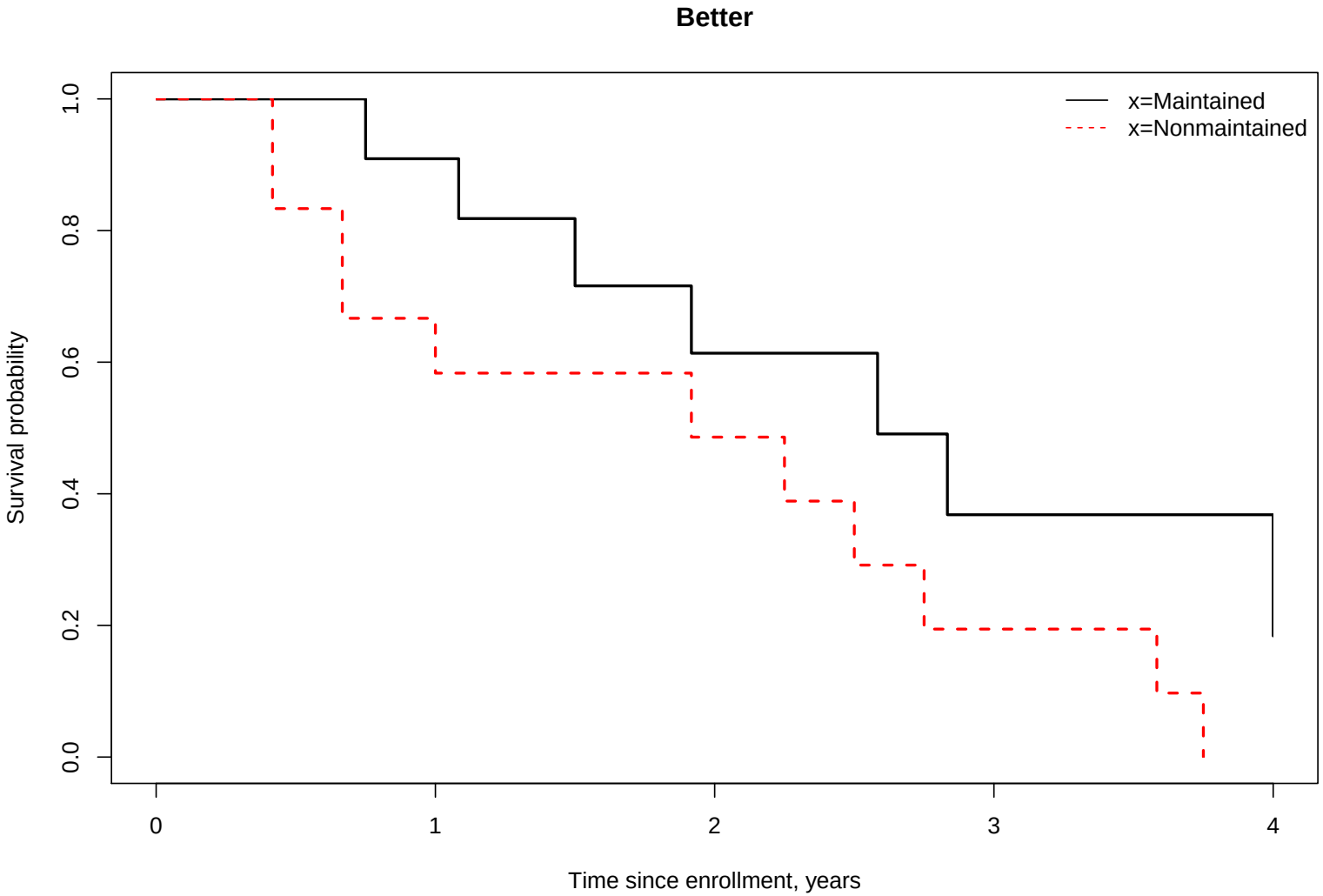

Default Plot
Survival probability
0.0
0.2
0.4
0.6
0.8
1.0
0
50
100
150
Time since enrollment, months

MAYO CLINIC

Kaplan-Meier Curves: better

```
> print(fit, scale=12)
Call: survfit(formula = Surv(time, status) ~ x, data = aml)

                 n events median 0.95LCL 0.95UCL
x=Maintained    11      7   2.58   1.500      NA
x=Nonmaintained 12     11   1.92   0.667      NA
>
> plot(fit, xscale=12, xlim=c(0, 4*12),
      col=1:2, lty=1:2,
      xlab='Time since enrollment, years',
      ylab='Survival probability')
> legend('topright', legend=names(fit$strata),
       col=1:2, lty=1:2, bty='n')
```

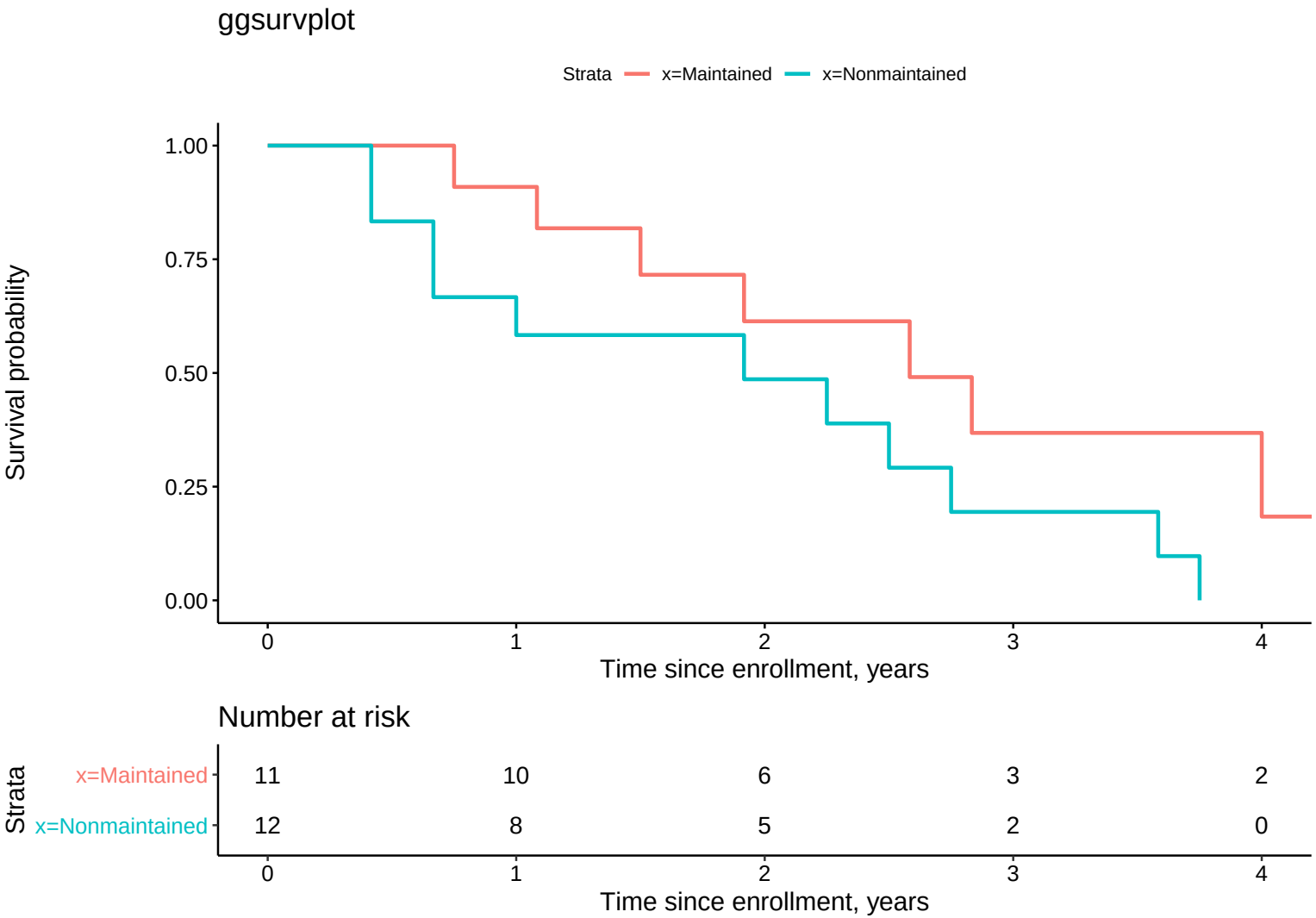
MAYO CLINIC

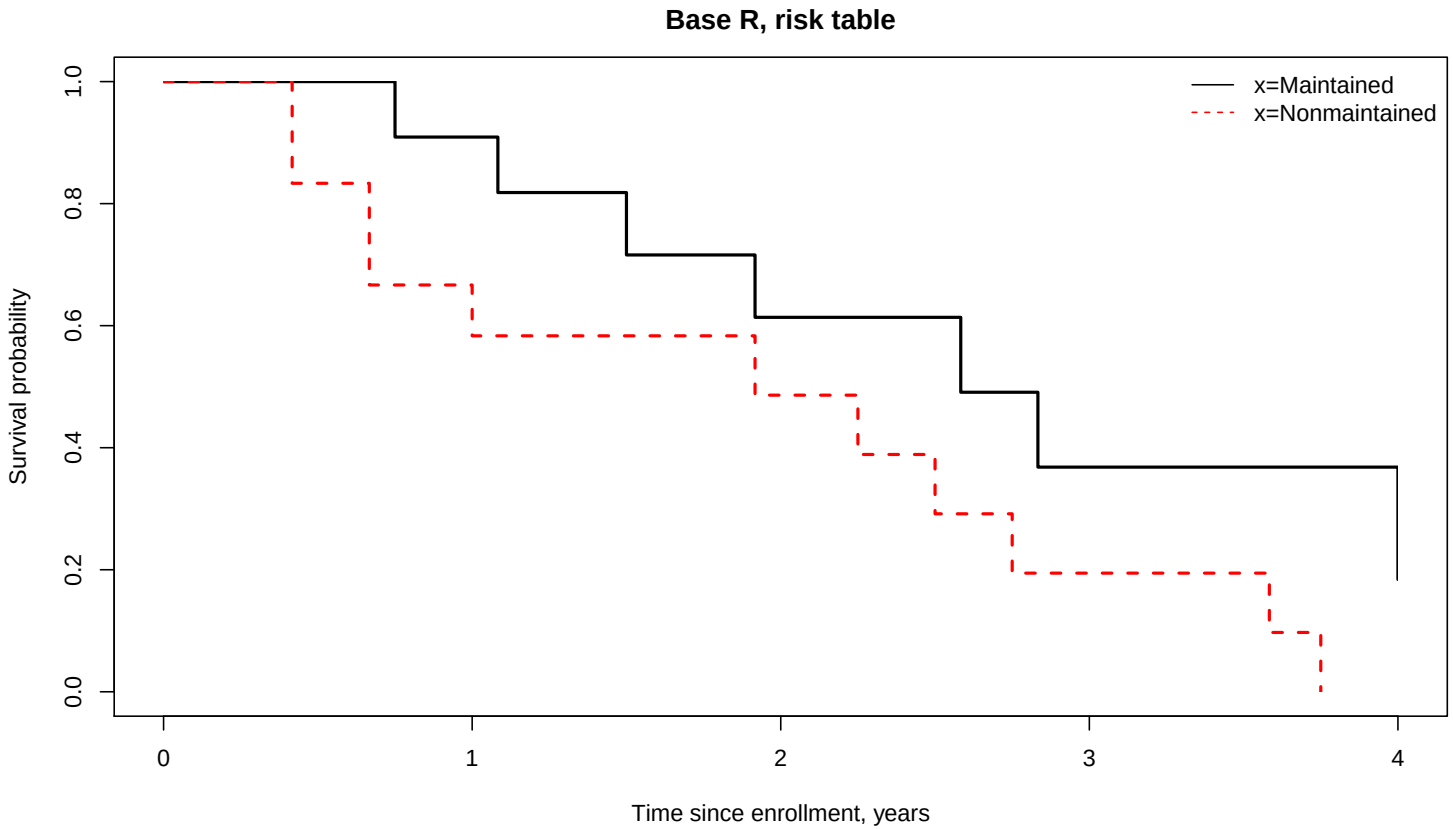

Better
x=Maintained
x=Nonmaintained
Survival probability
0.0
0.2
0.4
0.6
0.8
1.0
0
1
2
3
4
Time since enrollment, years

Kaplan-Meier Curves: ggsurvplot

```
> library(survminer)
>
> ggsurvplot(fit, xscale=12, xlim=c(0, 4*12),
             censor=FALSE, break.x.by=12,
             risk.table=TRUE,
             xlab='Time since enrollment, years',
             ylab='Survival probability')
```

MAYO CLINIC


ggsurvplot
Strata x=Maintained x=Nonmaintained
Survival probability
1.00
0.75
0.50
0.25
0.00
0
1
2
3
4
Time since enrollment, years
Number at risk
Strata
x=Maintained
x=Nonmaintained
11
10
6
3
2
12
8
5
2
0
0
1
2
3
4
Time since enrollment, years



Maintained	11	10	6	3	2
Nonmaintained	12	8	5	2	0

Cox Models

```
> fit <- coxph(Surv(time, status) ~ x, data=aml)
> fit
Call:
coxph(formula = Surv(time, status) ~ x, data = aml)

                coef exp(coef) se(coef)     z      p
xNonmaintained 0.9155    2.4981   0.5119 1.788 0.0737

Likelihood ratio test=3.38  on 1 df, p=0.06581
n= 23, number of events= 18
```

Your Turn - Run basic analysis

- See exercises/1.basic_survival.Rmd

Start/Stop Data

Use Cases

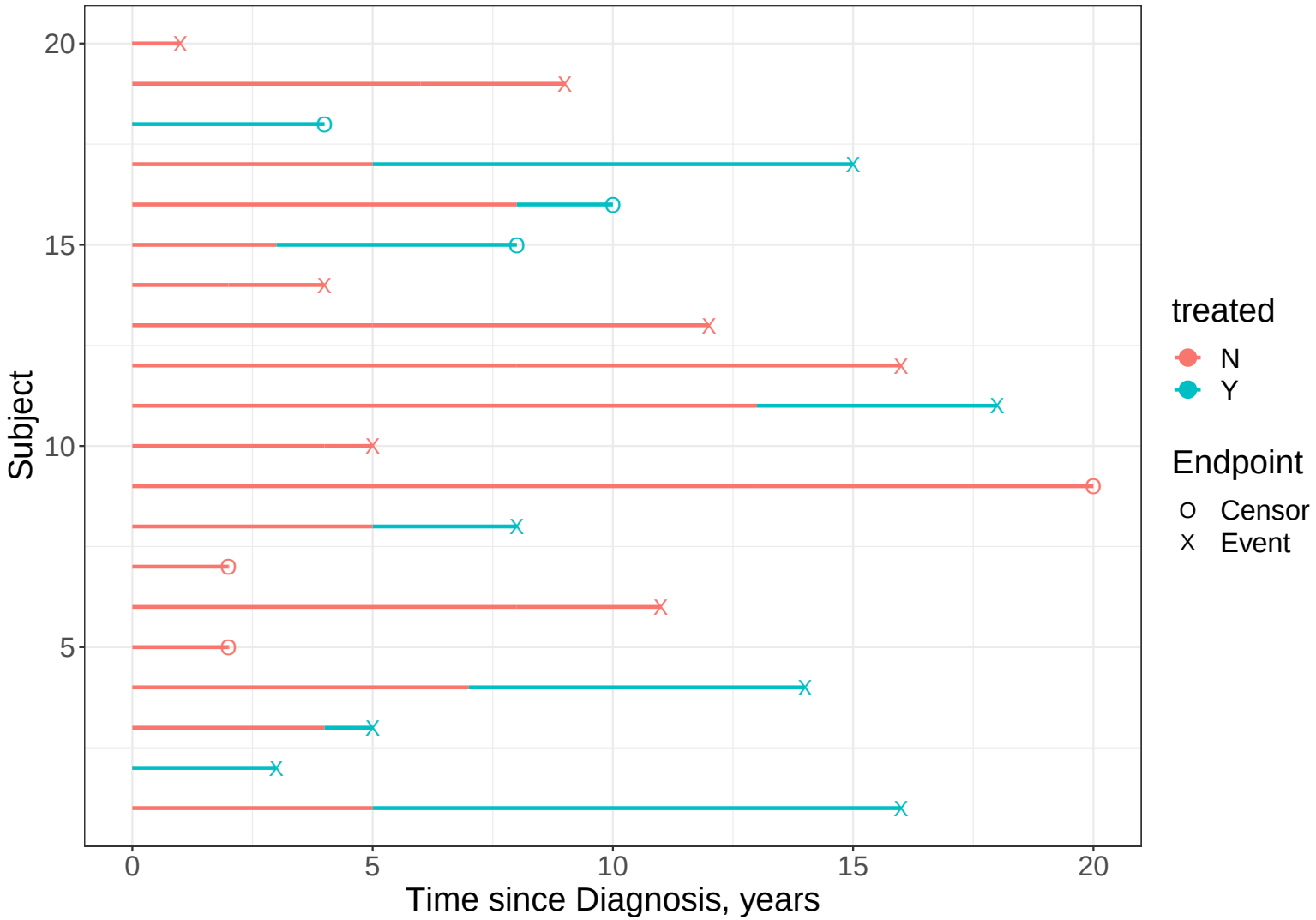
When is start/stop data needed?

- Time-dependent covariates
- Multiple events of the same type per subject
- Left truncation or gaps in observation
- Analysis by time periods
- Multistate

=> Deceptively simple task, easy to do incorrectly

MAYO CLINIC

Time-Dependent Covariates

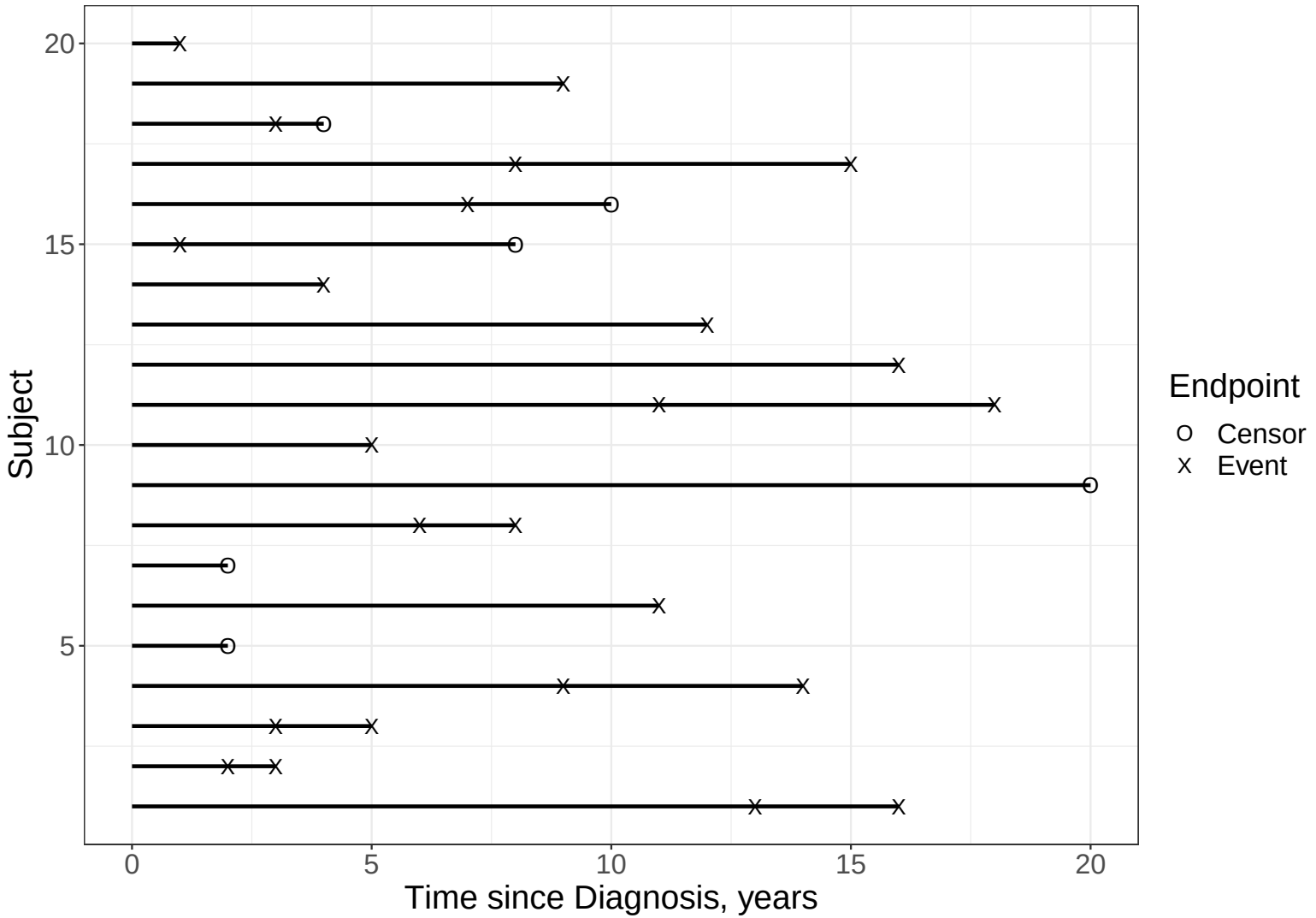


Time-Dependent Covariates

- Lab values that change over time (`pbcseq` data)
- Medication
  - Ever exposed
  - Cumulative dose
  - On and off
- Diagnosis of new comorbidity (e.g., diabetes)
- Surgery

MAYO CLINIC

Multiple Events/Same Type

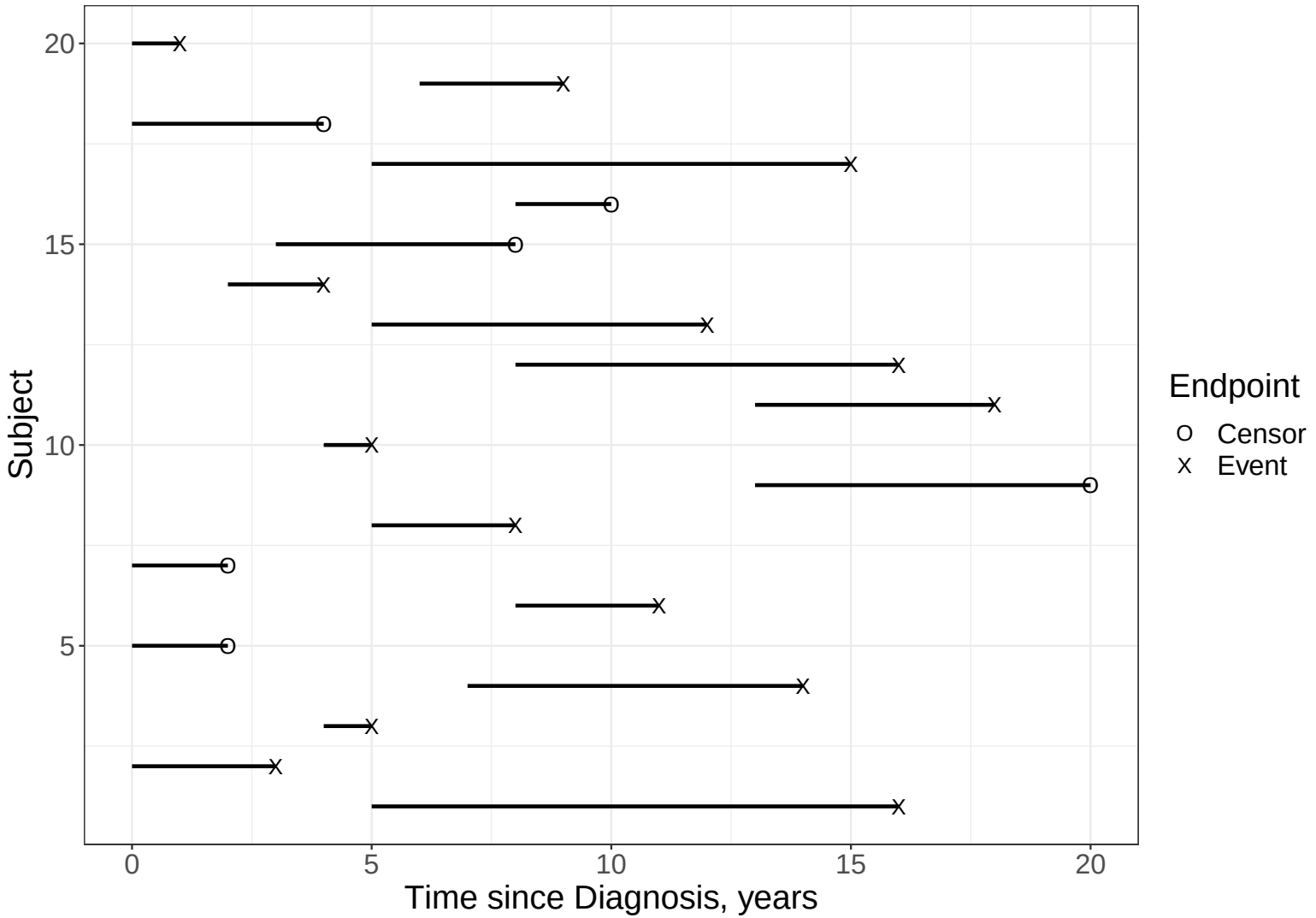


Multiple Events/Same Type

- Fractures
- Repeat infections (`rhDNase`, `cgd` data)
- Number of recurrences of cancer (`bladder` data)

MAYO CLINIC

Left Truncation

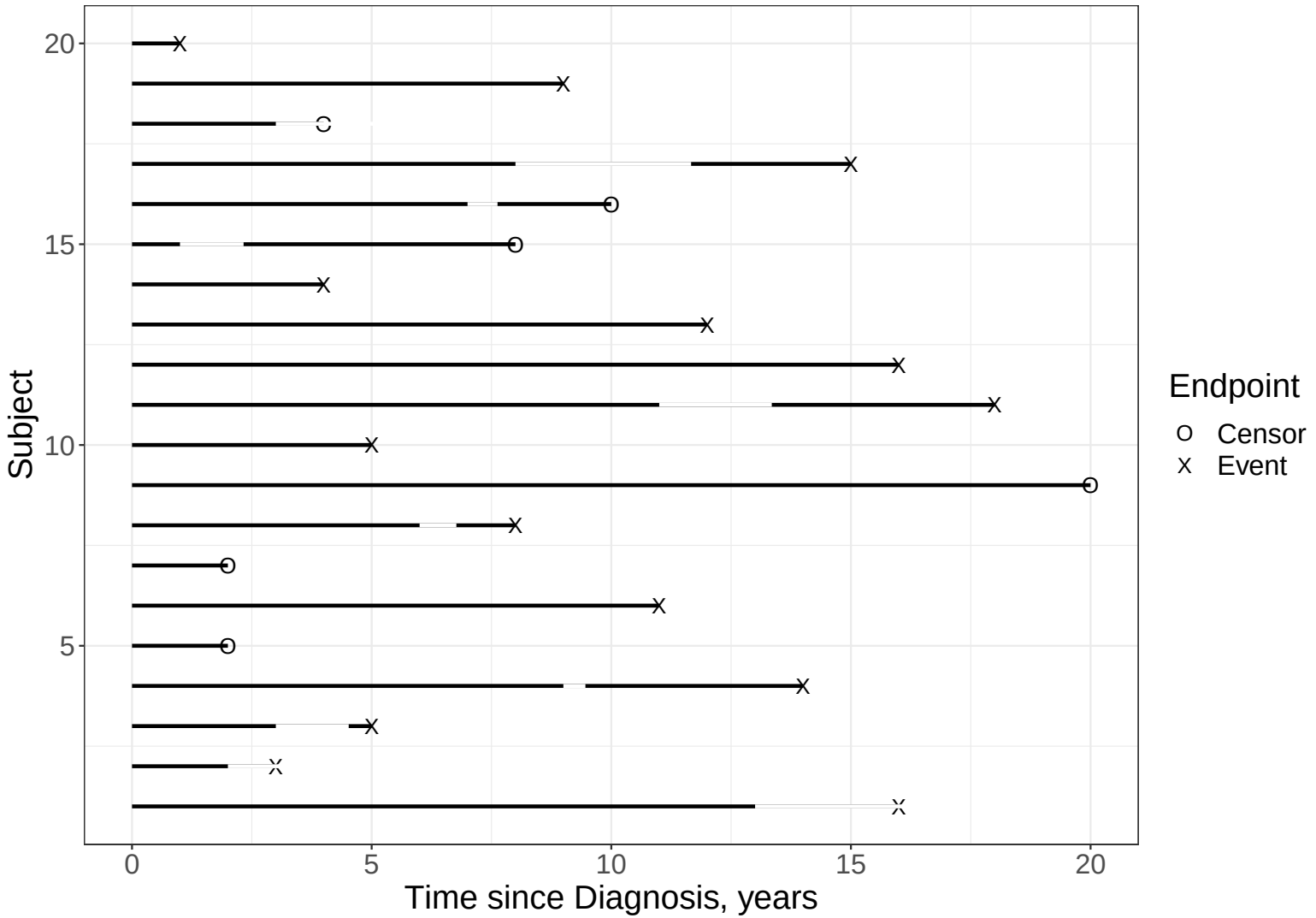


Left Truncation

- Disease started prior to diagnosis, want time-scale to be time-since-onset
- Population-based cohort, interested in "age" as a time-scale

MAYO CLINIC

Gaps in Follow-up

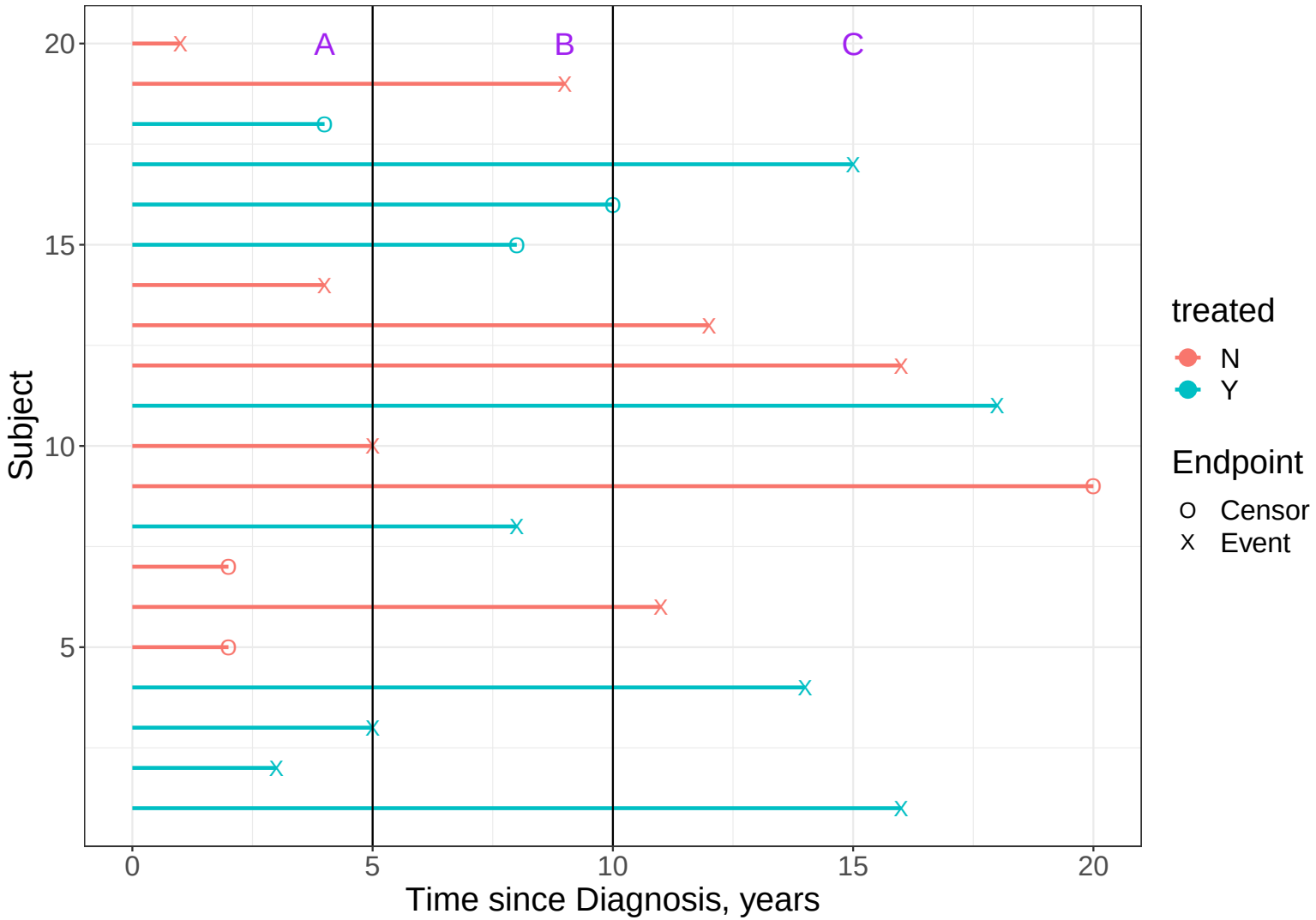


Gaps in Follow-up

- After an event, subjects are not at risk during the course of antibiotics or for 6 days after treatment ends (`rhDNase` data)
- Subjects move out of the region temporarily and are not "at risk" during that time

MAYO CLINIC

Analysis by Time Period



Analysis by Time Period

- Risk of event during first 5 years after cancer is different than afterwards
- Effect decreases over time
  - baseline lab variable
  - treatment

Simple Example: Data we have

- Initial dataset has 1 observation per subject
- Surgery is a time-dependent covariate

id	age	tm_fu	event	tm_surg
1	40	10	0	5
2	20	20	1	8
3	50	30	1	NA

Simple Example: What we want

id	tstart	tstop	death	age	surgery
1	0	5	0	40	0
1	5	10	0	40	1
2	0	8	0	20	0
2	8	20	1	20	1
3	0	30	1	50	0

MAYO CLINIC

Counting Process data

- (time1, time2] time interval
- status at the end of time2
- covariates as of time1

Creating Start/Stop Data

The tmerge function

- tmerge function in survival package: tool for creating start/stop data
- Sequential insertion
  - Build the dataset one covariate or endpoint at a time
  - Each addition will be "slipped in" to the original data in the same way that one would slide a new card into an existing deck of cards

MAYO CLINIC

The tmerge function

- The basic form of the function call is

```
> newdata <- tmerge(data1, data2, id,
                    newvar=tdc(time, value), ...)
```

- primary arguments:
  - data1: baseline data to be retained in the analysis dataset
  - data2: source for new data including events and time-dependent covariates
  - id: subject identifier used to merge the data together
  - ...: additional arguments that add variables to the dataset
  - tstart, tstop: used to set the time range for each subject
  - options

MAYO CLINIC

The `tmerge` function

- The key part of the call are the "..." arguments, which can be one of 4 types:
  - tdc() and cumtdc() add a time-dependent covariate
  - event() and cumevent() add a new endpoint
- Resulting dataset has 3 new variables (at least):
  - `id`: identifier indicating which rows belong to the same subject
  - `tstart`: start of the interval
  - `tstop`: end of the interval

Example

- Baseline data: d1

```
  id age
1  1  40
2  2  20
3  3  50
```

- Time varying data: d2

```
  id tm_fu event tm_surg
1  1    10     0       5
2  2    20     1       8
3  3    30     1      NA
```

Example: step 1 - create start/stop time

```
> step1 <- tmerge(data1=d1, data2=d2, id=id,
                  death=event(tm_fu, event))
> step1
  id age tstart tstop death
1  1  40      0    10     0
2  2  20      0    20     1
3  3  50      0    30     1
```

MAYO CLINIC

Example: step 2 - create time-dependent covariate

```
> step2 <- tmerge(data1=step1, data2=d2, id=id,
                  surgery=tdc(tm_surg))
> step2
  id age tstart tstop death surgery
1  1  40      0     5     0       0
2  1  40      5    10     0       1
3  2  20      0     8     0       0
4  2  20      8    20     1       1
5  3  50      0    30     1       0
```

MAYO CLINIC

This can also be done in just one step:

```
> tmerge(data1=d1, data2=d2, id=id,
         death=event(tm_fu, event),
         surgery=tdc(tm_surg))
```

tcount attribute

```
> attr(step2, "tcount")
```

```
         early late gap within boundary lead trail tied
death        0    0   0      0        0    0     3    0
surgery      0    0   0      2        0    0     0    0
```

MAYO CLINIC

tcount attribute

`tcount` - a tool to check data

- Time outside the specified time frame.
  - "early" occur before the first interval for a subject
  - "late" occur after the last interval for a subject
  - "gap" times fall into a gap
  - These events will be discarded.
  - A TD covariate value will be applied to later intervals
- "within" fall inside an existing interval and cause it to be split into two

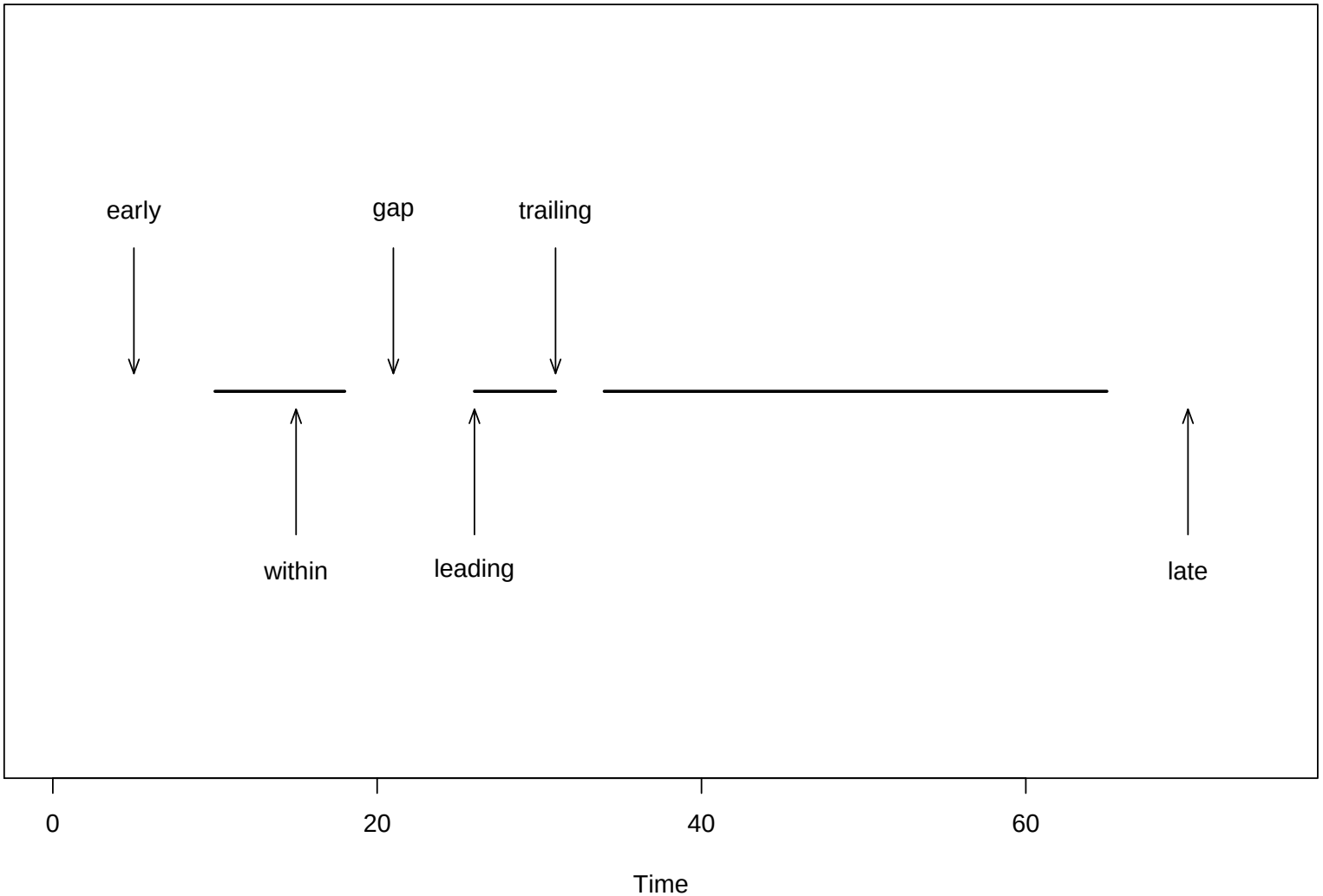
MAYO CLINIC

`tcount` attribute

- Observations that fall exactly on the edge of an interval but within the (min, max] time for a subject are counted as being on a "leading" edge, "trailing" edge or "boundary".
- "tied" shows # of additions where the id and time point were identical.

MAYO CLINIC

tcount attribute



tcount attribute

- 3 *trailing* deaths
- 2 *within* splits with surgery

	id	age	tstart	tstop	death	surgery
1	1	40	0	5	0	0
2	1	40	5	10	0	1
3	2	20	0	8	0	0
4	2	20	8	20	1	1
5	3	50	0	30	1	0

MAYO CLINIC

Example: Original Analysis, Stanford heart transplant data

- Original analysis used: futime, fustat, transplant status, and age
  - Transplant happened after baseline
  - `jasa` dataset

```
Call:
coxph(formula = Surv(futime, fustat) ~ age + transplant, data =

               coef exp(coef) se(coef)      z        p
age         0.06020   1.06205  0.01531  3.933 8.40e-05
transplant -1.80047   0.16522  0.27225 -6.613 3.76e-11

Likelihood ratio test=44.46  on 2 df, p=2.214e-10
n= 103, number of events= 75
```

==> **Immortal time bias** <==

Your Turn - Create the Correct Data

- Stanford heart transplant data (`jasa`)
  - wait.time: time before transplant (tx)
  - futime: follow-up time
  - fustat: dead or alive
  - age
- Create

id	tstart	tstop	death	age	tx
1	.	.	.	.	.

See the file `exercises/2.jasa.Rmd`.

MAYO CLINIC

Stanford Heart Transplant

```
> jasa$id <- 1:nrow(jasa)
> sdata <- tmerge(jasa, jasa, id=id,
                  death = event(futime, fustat),
                  tx = tdc(wait.time))
```

```
Error in tmerge(jasa, jasa, id = id,
          death = event(futime, fustat),
          transplant = tdc(wait.time)) :
found an ending time of 0,
      the default starting time of 0 is invalid
```

MAYO CLINIC

What went wrong?

```
> jasa %>% filter(futime==0) %>%
   select(id, futime, fustat, wait.time)
  id futime fustat wait.time
1 15      0      1        NA
```

- **1 subject died on the day of entry.** (0,0) is an illegal time interval for coxph.
  It suffices to have them die on day 0.5.

```
> jasa$futime <- pmax(0.5, jasa$futime)
```

MAYO CLINIC

Rerun

```
> sdata <- tmerge(jasa, jasa, id=id,
                  death = event(futime, fustat),
                  tx = tdc(wait.time))
```

```
> attr(sdata, "tcount")
```

```
      early late gap within boundary lead trail tied
death     0    0   0      0        0    0   103    0
tx        0    0   0     66        0    2     1    0
```

What does "trailing" mean?

```
> jasa %>% filter(wait.time==futime) %>%
          select(id, futime, fustat, wait.time)
  id futime fustat wait.time
1 38      4      1         4
```

- **Subject died on the same day as their procedure.** The problem is resolved by moving the transplant back 0.5 day.

```
> jasa$wait.time <- if_else(jasa$wait.time==jasa$futime,
                            jasa$wait.time - .5,
                            jasa$wait.time)
```

MAYO CLINIC

Rerun again

```
> sdata <- tmerge(jasa, jasa, id=id,
                  death = event(futime, fustat),
                  tx = tdc(wait.time))
```

```
> attr(sdata, "tcount")
```

```
      early late gap within boundary lead trail tied
death     0    0   0      0        0    0   103    0
tx        0    0   0     67        0    2     0    0
```

Yay!

Cox Model

```
> fit <- coxph(Surv(tstart, tstop, death) ~ age + tx,
               data=sdata)
> fit
Call:
coxph(formula = Surv(tstart, tstop, death) ~ age + tx, data = sd

        coef exp(coef)  se(coef)      z      p
age  0.030758  1.031236  0.014496  2.122 0.0339
tx  -0.006058  0.993960  0.311750 -0.019 0.9845

Likelihood ratio test=5.17  on 2 df, p=0.07541
n= 170, number of events= 75
```

Example: Continuous values that change over time

- `pbcseq` is from the Mayo Clinic trial in primary biliary cirrhosis (PBC) of the liver conducted between 1974 and 1984. 312 subjects were randomized to placebo or D-penicillamine.
- The data has 1945 observations with repeated laboratory values + baseline variables

id	futime	status	trt	day	alk.phos	bili
1	400	2	1	0	1718	14.5
1	400	2	1	192	1612	21.3
5	1505	1	0	0	671	3.4
5	1505	1	0	199	689	1.9
5	1505	1	0	391	652	2.5
5	1505	1	0	769	554	5.7
5	1505	1	0	1098	588	5.2
5	1505	1	0	1455	377	19.0

MAYO CLINIC

Create baseline data

```
> # baseline
> pbc_b <- pbcseq %>% select(id:sex) %>% distinct()
> head(pbc_b)
  id futime status trt      age sex
1  1    400      2   1 58.76523   f
2  2   5169      0   1 56.44627   f
3  3   1012      2   1 70.07255   m
4  4   1925      2   1 54.74059   f
5  5   1505      1   0 38.10541   f
6  6   2503      2   0 66.25873   f
```

MAYO CLINIC

Look at status

```
> table(pbc_b$status)

  0   1   2
143  29 140
```

After discussion with investigator, decided that in this instance, transplant (1) and death (2) can both be treated as death.

```
> pbc_b$status2 <- as.numeric(pbc_b$status>0)
```

MAYO CLINIC

Set range

```
> # set range
> newpbc <- tmerge(pbc_b, pbc_b, id=id,
                   death = event(futime, status2))
>
> print(head(newpbc),digits=2)
  id futime status trt age sex status2 tstart tstop death
1  1    400      2   1  59   f       1      0   400     1
2  2   5169      0   1  56   f       0      0  5169     0
3  3   1012      2   1  70   m       1      0  1012     1
4  4   1925      2   1  55   f       1      0  1925     1
5  5   1505      1   0  38   f       1      0  1505     1
6  6   2503      2   0  66   f       1      0  2503     1
```

MAYO CLINIC

Create new TDC variables

```
> newpbc <- tmerge(newpbc, pbcseq, id = id,
                   ascites = tdc(day, ascites),
                   bili = tdc(day, bili),
                   albumin = tdc(day, albumin))
```

MAYO CLINIC

id	tstart	tstop	death	sex	ascites	bili	albumin
1	0	192	0	f	1	14.5	2.60
1	192	400	1	f	1	21.3	2.94
2	0	182	0	f	0	1.1	4.14
2	182	365	0	f	0	0.8	3.60
2	365	768	0	f	0	1.0	3.55
2	768	1790	0	f	0	1.9	3.92

Example: Continuous values that change over time

```
> attr(newpbc, "tcount")
```

	early	late	gap	within	boundary	lead	trail	tied
death	0	0	0	0	0	0	312	0
ascites	0	0	0	1573	0	312	0	0
bili	0	0	0	60	1573	312	0	0
albumin	0	0	0	0	1633	312	0	0

MAYO CLINIC

Example: Continuous values that change over time

- Missing values in time or value from `data2` are ignored
  - Consequence: "last value carried forward"
- Default can be changed by adding `options=list(na.rm=FALSE)` to the second call
  - Any `tdc` calls with a missing time are still ignored, independent of the na.rm value, since we would not know where to insert them.

MAYO CLINIC

Code available in `3.pbc.Rmd`

How covariates differ from events

- Time-dependent covariates
  - Apply from the *start* of a new interval
  - Persist for all remaining intervals unless subsequently changed
- Events
  - Occur at the *end* of an interval
  - Only occur once

MAYO CLINIC

Your Turn

- Chronic Granulotamous Disease (cgd0)
  - id, treat, sex, age
  - futime: follow-up time
  - etime1-etime7: up to 7 infection times/subject
- Create

id	tstart	tstop	infect	treat	enum
1	.	.	.	.	.

where `enum` is the interval number/id

See `exercises/4.cgd.Rmd`

MAYO CLINIC

CGD

```
> newcgd <- tmerge(data1=cgd0, data2=cgd0,
                   id=id, tstop=futime,
                   infect=event(etime1), infect=event(etime2),
                   infect=event(etime3), infect=event(etime4),
                   infect=event(etime5), infect=event(etime6),
                   infect=event(etime7))
> newcgd <- tmerge(newcgd, newcgd, id=id,
                   enum=cumtdc(tstart))
```

MAYO CLINIC

CGD

```
> attr(newcgd, "tcount")
```

	early	late	gap	within	boundary	lead	trail	tied
infect	0	0	0	44	0	0	0	0
infect	0	0	0	16	0	0	1	0
infect	0	0	0	8	0	0	0	0
infect	0	0	0	3	0	0	0	0
infect	0	0	0	2	0	0	0	0
infect	0	0	0	1	0	0	0	0
infect	0	0	0	1	0	0	0	0
enum	0	0	0	0	75	128	0	0

MAYO CLINIC

CGD

```
> newcgd %>% filter(id==2) %>%
   select(id, tstart, tstop, infect, enum)
  id tstart tstop infect enum
1  2      0     8      1    1
2  2      8    26      1    2
3  2     26   152      1    3
4  2    152   241      1    4
5  2    241   249      1    5
6  2    249   322      1    6
7  2    322   350      1    7
8  2    350   439      0    8
```

CGD

```
> fit <- coxph(Surv(tstart,tstop,infect) ~ treat +
             steroids + inherit, id=id, data=newcgd)
> fit
Call:
coxph(formula = Surv(tstart, tstop, infect) ~ treat + steroids +
    inherit, data = newcgd, id = id)

            coef exp(coef) se(coef) robust se      z
treat    -1.0722    0.3422   0.2619    0.3118 -3.438
steroids -0.7726    0.4618   0.5169    0.4687 -1.648
inherit   0.1777    1.1944   0.2356    0.3180  0.559
                p
treat    0.000585
steroids 0.099310
inherit  0.576395

Likelihood ratio test=22.49  on 3 df, p=5.149e-05
```

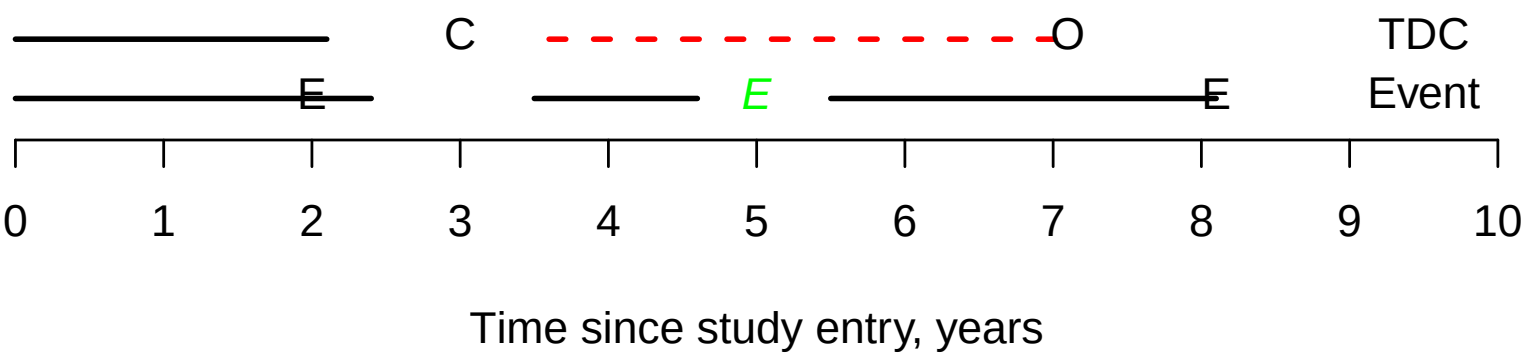
CGD

- Look at the first infection versus all infections

```
> fit0 <- coxph(Surv(tstart,tstop,infect) ~ treat + steroids +
                  inherit, id=id, data=newcgd, subset=enum==1)
> round(cbind(first=coef(fit0),all=coef(fit)), 3)
          first    all
treat    -1.093 -1.072
steroids -0.809 -0.773
inherit   0.025  0.178
```

Gaps

- *Time dependent covariates* that occur before the start of a subject's follow-up interval or during a gap in time do not generate a new interval split, but they do set the value of that covariate for future times.
  - During a subject's time under observation we would like the variable "Has diabetes" to be accurate
- *Events* that occur in a gap are not counted.
  - Don't know the appropriate comparison group, so we ignore those events.



Example: Intentional gaps, rhDNase

- Randomized clinical trial examining a treatment for cystic fibrosis
- Infection is the event of interest, indicated by `ivstart`
- For 6 days after `ivstop`, the subject is not at risk for a new infection

id	inst	trt	entry.dt	end.dt	fev	ivstart	ivstop
1	1	1	1992-03-20	1992-09-04	28.8	NA	NA
129	12	0	1992-02-23	1992-08-07	105.6	5	26
129	12	0	1992-02-23	1992-08-07	105.6	44	58
129	12	0	1992-02-23	1992-08-07	105.6	87	108
129	12	0	1992-02-23	1992-08-07	105.6	124	143
129	12	0	1992-02-23	1992-08-07	105.6	163	166

MAYO CLINIC

Your Turn

Use the `rhDNase` data found in the `survival` package:

1. Create range for when subjects are under observation (`tmerge`)
2. Create event for each infection (`tmerge`)
3. Create intervals where they are not at risk (`tmerge`)
4. Remove intervals where not at risk
5. Add a counter for each person (`tmerge`)
6. Check data and `tcount` attribute

See `exercises/5.dnase.Rmd`

MAYO CLINIC

rhDNase

Quick look at the data

```
> dim(rhDNase)
[1] 767   8
> head(rhDNase)
  id inst trt   entry.dt     end.dt  fev ivstart ivstop
1  1    1   1 1992-03-20 1992-09-04 28.8      NA     NA
2  2    1   1 1992-03-24 1992-09-09 64.0      NA     NA
3  3    1   0 1992-03-24 1992-09-08 67.2      65     75
4  4    1   1 1992-03-26 1992-09-10 57.6      NA     NA
5  5    1   0 1992-03-24 1992-09-11 57.6      NA     NA
6  6    1   1 1992-03-27 1992-09-09 25.6      NA     NA
```

MAYO CLINIC

rhDNase

```
> table(table(rhDNase$id)) # number obs/id

  1   2   3   4   5
565  53  21   7   1
> table(!is.na(rhDNase$ivstart)) # number events

FALSE  TRUE
  400   367
```

rhDNase

Create range for when subjects are under observation

```
> # Make sure data is sorted by id, ivstart time
> dnase <- rhDNase %>% arrange(id,ivstart) %>%
   mutate(end.tm = as.numeric(end.dt - entry.dt))
>
> # 1st obs/id
> dnase.b <- dnase %>% distinct(id, .keep_all=TRUE)
>
> dn1 <- tmerge(dnase.b[,c('id','inst','trt','fev')],
                dnase.b, tstop=end.tm, id=id)
```

rhDNase

Create event for each infection

```
> dn2 <- tmerge(dn1, dnase,
                infect=event(ivstart), id=id)
> dn2[dn2$id==129,]
     id inst trt   fev tstart tstop infect
204 129   12   0 105.6      0     5      1
205 129   12   0 105.6      5    44      1
206 129   12   0 105.6     44    87      1
207 129   12   0 105.6     87   124      1
208 129   12   0 105.6    124   163      1
209 129   12   0 105.6    163   166      0
```

MAYO CLINIC

rhDNase

Create intervals where they are not at risk

```
> dn3 <- tmerge(dn2, dnase,
                no.risk=event(ivstop+6), id=id)
> dn3[dn3$id==129,]
     id inst trt   fev tstart tstop infect no.risk
271 129   12   0 105.6      0     5      1       0
272 129   12   0 105.6      5    32      0       1
273 129   12   0 105.6     32    44      1       0
274 129   12   0 105.6     44    64      0       1
275 129   12   0 105.6     64    87      1       0
276 129   12   0 105.6     87   114      0       1
277 129   12   0 105.6    114   124      1       0
278 129   12   0 105.6    124   149      0       1
279 129   12   0 105.6    149   163      1       0
280 129   12   0 105.6    163   166      0       0
```

MAYO CLINIC

rhDNase

Remove intervals where not at risk

```
> dn4 <- dn3[dn3$no.risk!=1,]
```

Add a counter for each person

```
> newdnase <- tmerge(dn4, dn4, enum=cumtdc(tstart), id=id)
```

MAYO CLINIC

rhDNase

Check to make sure code worked correct

```
> newdnase[newdnase$id==129,]
     id inst trt   fev tstart tstop infect no.risk enum
204 129   12   0 105.6      0     5      1       0    1
205 129   12   0 105.6     32    44      1       0    2
206 129   12   0 105.6     64    87      1       0    3
207 129   12   0 105.6    114   124      1       0    4
208 129   12   0 105.6    149   163      1       0    5
209 129   12   0 105.6    163   166      0       0    6
```

MAYO CLINIC

rhDNase: check tcount

```
> attr(newdnase, "tcount")
```

```
          early late gap within boundary lead trail tied
infect        6    0   0    358        0    0     3    0
no.risk       0   51   0    315        0    0     1    0
enum          0    0   0      0       46  958     0    0
```

tmerge summary

- tmerge is a simple to use, flexible tool to create multiple start/stop intervals per subject
  - time-dependent covariates - both binary and continuous
  - multiple outcomes per subject
  - allows for gaps in time
  - sometimes useful to create both tdc and event
- data checks can help avoid errors
  - tcount attribute

MAYO CLINIC

The `survSplit` function

- Another approach to create start/stop data
- Breaks follow-up at specified cut points
- Useful when you want separate coefficients within time periods

MAYO CLINIC

Go back to d2 data

	id	tm_fu	event	tm_surg
1	1	10	0	5
2	2	20	1	8
3	3	30	1	NA

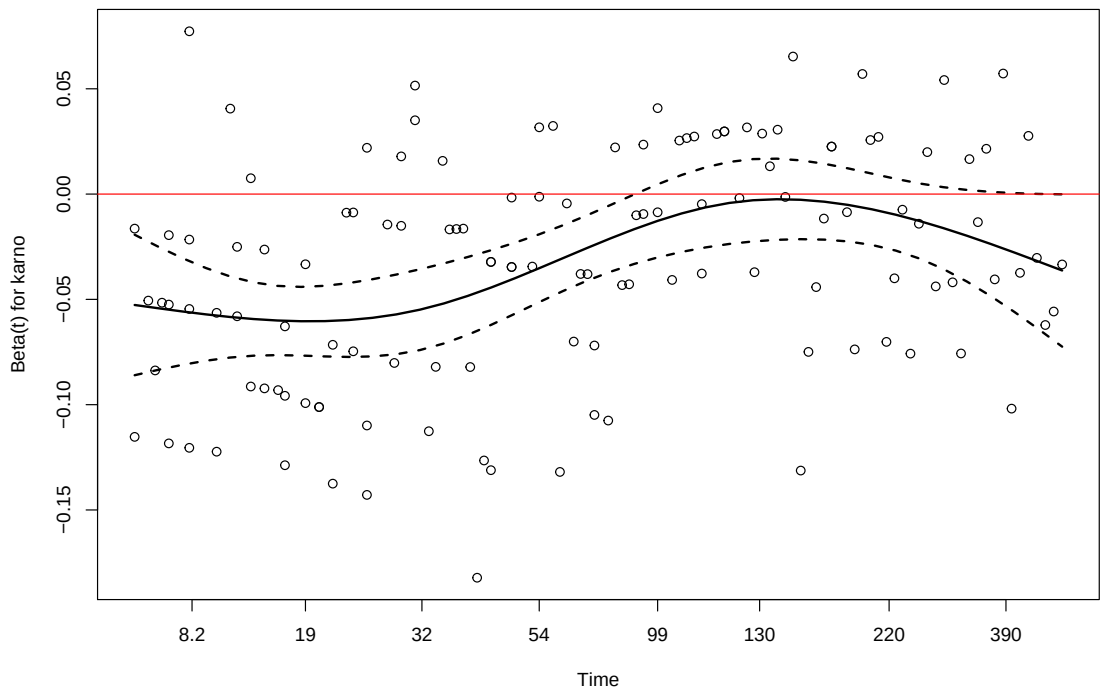
MAYO CLINIC

survSplit

```
> survSplit(Surv(tm_fu, event) ~ ., data=d2,
            cut=c(5,15), episode='timegp')
  id tm_surg tstart tm_fu event timegp
1  1       5      0     5     0      1
2  1       5      5    10     0      2
3  2       8      0     5     0      1
4  2       8      5    15     0      2
5  2       8     15    20     1      3
6  3      NA      0     5     0      1
7  3      NA      5    15     0      2
8  3      NA     15    30     1      3
```

Example: Veteran Data

```
> fit1 <- coxph(Surv(time, status) ~ karno, data=veteran)
> plot(cox.zph(fit1), lwd=2)
> abline(h=0, col=2)
```



MAYO CLINIC

Example: Veteran Data

```
> vet2 <- survSplit(Surv(time, status) ~., veteran,
                    cut=c(60, 120), episode ="timegp")
> fit2 <- coxph(Surv(tstart, time, status) ~
                    karno:strata(timegp),
                data = vet2)
```

MAYO CLINIC

```
> # original
> round(coef(fit1), 3)
 karno
-0.033
>
> # within time period
> tmp <- round(coef(fit2), 3)
> names(tmp) <- c('t0_60','t60_120','t120+')
>
> tmp
  t0_60 t60_120   t120+
 -0.048  -0.011  -0.007
```

MAYO CLINIC

Caution - don't create too many intervals

- Compute time for `coxph` is proportional to the number of observations in the dataset
- If there are no observations that span two event times, then further splitting won't change the results

MAYO CLINIC

Check data/Common Mistakes

The survcheck function

- The basic form of the function call is

```
> ck1 <- survcheck(formula, data, id, istate)
```

- checks include:
  - overlap: subject is in 2 places at the same time
  - gap: gap in timeline
  - teleport: 2 adjacent intervals, change in state
  - jump: hole in timeline, change in state

MAYO CLINIC

Example: Multiple Events/Same type (newcgd)

```
Call:
survcheck(formula = Surv(tstart, tstop, infect) ~ treat, data =
    id = id)

128 subjects available for analysis
Transitions table:
      to
from    1 (censored)
  (s0) 44          84
  1    32          43

Number of subjects with 0, 1, ... transitions to each state:
       count
state    0  1 2 3 4 5 7
  1     84 27 9 5 1 1 1
  (any) 84 27 9 5 1 1 1
```

Example: Gaps (newdnase)

```
Call:
survcheck(formula = Surv(tstart, tstop, infect) ~ trt, data = ne
    id = id)

647 subjects available for analysis
Transitions table:
      to
from     1 (censored)
  (s0) 243         404
  1    118         239

Number of subjects with 0, 1, ... transitions to each state:
       count
state     0   1  2  3 4 5
  1     404 162 53 20 7 1
  (any) 404 162 53 20 7 1
```

Example: Gaps (newdnase)

```
> length(tmp$gap$id)
[1] 224
> length(tmp$gap$row)
[1] 311
>
> # Look at first gap
> newdnase[newdnase$id==tmp$gap$id[1],]
   id inst trt  fev tstart tstop infect no.risk enum
3   3    1   0 67.2      0    65      1       0    1
4   3    1   0 67.2     81   168      0       0    2
```

MAYO CLINIC

Common Mistake: Responders vs non-responders

Group people, at baseline, according to whether they eventually had a response to therapy, and then draw the survival curves. Surprise – responders always do better! Why?

- Assume patients are evaluated every 4 weeks
- Response, if it occurs, will happen by week 12
- Anyone who dies before week 4 is a non-responder, and most of those who die, do so before week 8
- You have to live longer to be called a responder

MAYO CLINIC

Common Mistake: KM Curves using Time-Dep Covariates

Suppose you have created a time-dependent covariate and the researcher wants a Kaplan-Meier curve. Is that ok? How would you interpret it?

Instead, consider using landmark analysis. Consider a point in time (e.g., 1 year) and use the covariate status as it was known at 1 year. Start follow-up at that time point.

MAYO CLINIC

Common Mistake: Prophetic variables

Some time-dependent covariates are not predictors of an event as much as they are markers of a failure-in-progress:

- Multiple-organ failure
- Ventilation
- "Called the priest"
- Medication changes
  - Cessation of diuretics in heart failure
- PSA and prostate cancer, if measurement and declaration occur on the same visit

These will tend to be phenomenal predictors.

So what?

MAYO CLINIC

Evaluate Time delay

- For any dataset containing constructed time-dependent covariates, it is a good idea to re-run the analyses after adding a 7-14 day lag to key variables.
- When the results show a substantial change, understand why this occurred.

```
> newpbc <- tmerge(newpbc, pbcseq, id = id,
                   ascites2 = tdc(day, ascites),
                   bili2 = tdc(day, bili),
                   options= list(delay=14))
```

MAYO CLINIC

Common Mistake: Insidious look-ahead

Smoothed continuous variables:

- A particular lab test has values of
  - 120 on day 0
  - 150 on day 90
  - 180 on day 120
- What should we use for the value at day 100?
- It is tempting to use 160 (1/3 of the way between 150 and 180).
- Bad idea!

MAYO CLINIC

Common Mistake: Insidious look-ahead

Persistence:

- Patients with a solitary plasmacytoma are treated with local radiation
- The tumor produces an immunoglobulin spike
- If the spike is still present at the 1 year evaluation, this is a bad thing. (It mean that the 'solitary' lesion likely was not solitary.)
- Want to draw a curve for "survival, post 1 year".
- Does the patient evaluated at 13 months (with persistence) go in the 'persistent spike' or 'other' group?
- We know that the spike would have been present at 12 months, if the test had been done then.
- Still, it's a bad idea.

MAYO CLINIC

Common Mistake: Summaries by event status

Subjects with censored follow-up end up in the non-event category.

- Covariate summaries by event/non-event
- Standard ROC curves

Think about re-distribute to the right

MAYO CLINIC

Common Mistake: Using Future Data

"You can't use future information today"

- Mark an adverse event as midway between visits
- Delete subjects who do not complete treatment

MAYO CLINIC

Common Mistake: Immortal Time Bias

- "Last clinical FU" versus "last FU by any means"
  - Fractures can only be detected via clinical follow-up, but we have more knowledge about whether they are alive or dead.
- Subjects were recruited based on diagnosis at a tertiary care center, but we are interested in follow-up based on when the symptoms 1st appeared. Patients have to live long enough to be included in the study so use left-truncation.

MAYO CLINIC

Common Mistake: Electronic Studies

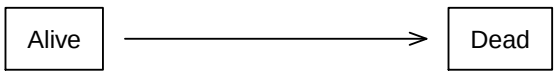
With electronic studies, it is easy to mess up.

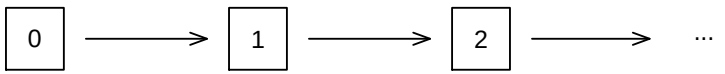
- One rule used for counting a diagnosis of a chronic condition is that there are "at least 2 instances, 30 days apart". Then an error is made, using the date of the first diagnostic code.

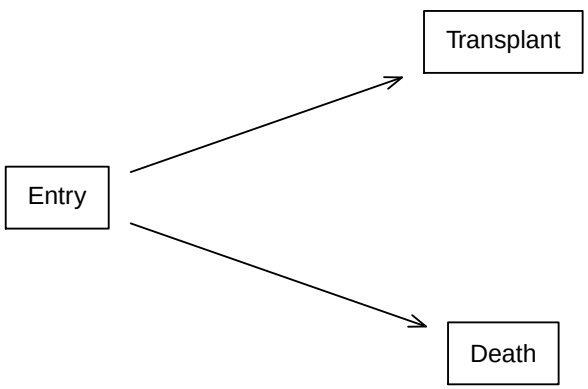
MAYO CLINIC

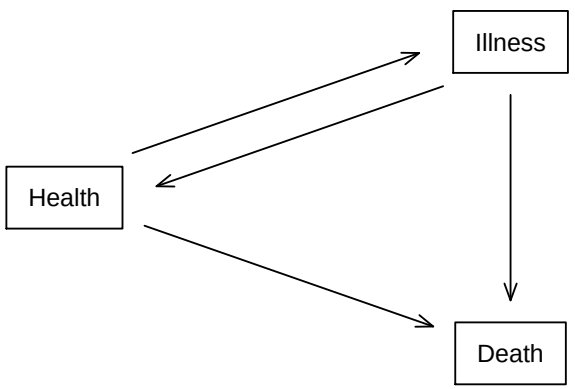
Multistate Data

Multistate Scenarios









MAYO CLINIC

Monoclonal Gammopathy of Undetermined Significance (MGUS)

- Subjects with a dominant clone in their plasma cell population, but without malignancy ($\geq$ 2% of plasma cells in the clone).
- Normally found incidentally to other tests.
- Should the patient be worried?
- About 1% per year convert to overt malignancy.
- Essentially independent of age and sex.

MAYO CLINIC

Example: Progression of MGUS

- 1384 subjects with monoclonal gammopathy of undetermined significance (MGUS)
- R. Kyle, New Engl J Med 346:564-569 (2002)
- Questions
  - Pattern of death and progression
  - Relationship to age, sex, hemoglobin, creatinine, and amount of protein in the "spike"

MAYO CLINIC

Example: MGUS data

The mgus2 dataset has two sets of variables that we are interested in. Time and event variables for:

- Progression (i.e., PCM): ptime and pstat.
- Death: futime and death.

id	age	sex	dxyr	hgb	ptime	pstat	futime	death
1379	73	M	1994	15.6	48	0	48	0
1380	69	M	1994	15.0	22	0	22	1
1381	78	M	1994	14.1	35	0	35	0
1382	66	M	1994	12.1	31	0	31	1
1383	82	F	1994	11.5	38	1	61	0
1384	79	M	1994	9.6	6	0	6	1

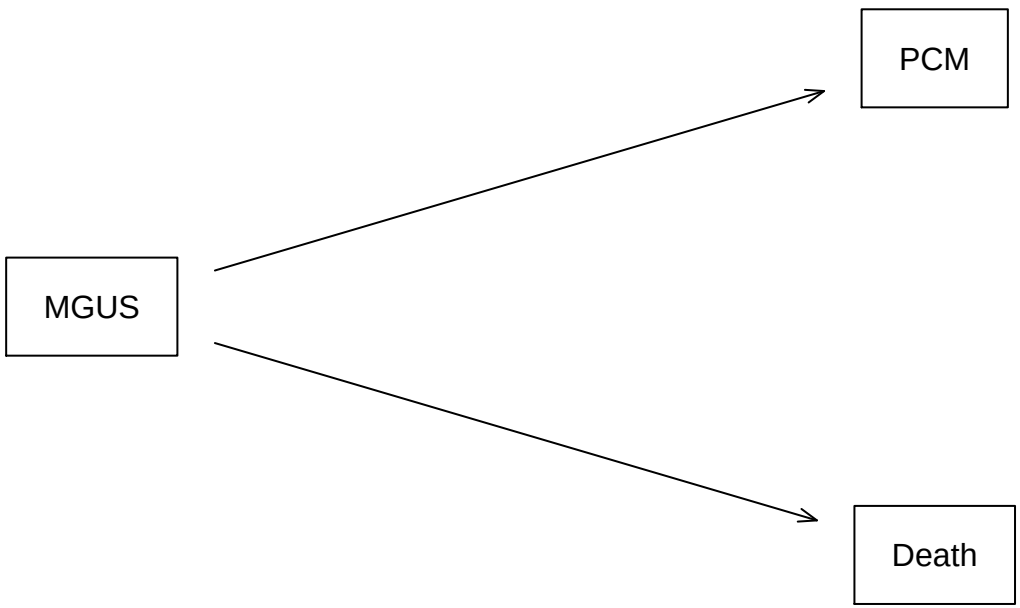
MAYO CLINIC

Create a Diagram - Competing Risk

```
> # Create names for the possible states
> states <- c("MGUS", "PCM", "Death")
> # Create matrix describing relationship between states
> connect <- matrix(0, nrow=3, ncol=3,
                    dimnames=list(states, states))
> # A non-zero element indicates that an arrow should be
> # drawn between state i (row) and state j (column)
> connect[1, c(2,3)] <- 1
> connect
      MGUS PCM Death
MGUS     0   1     1
PCM      0   0     0
Death    0   0     0
```

MAYO CLINIC

```
> # Plot
> statefig(layout=c(1,2), connect)
```



Example: MGUS - Competing Risk

Only need 1st event for each subject, so we only need 1 obs/person.

```
> # time variable will be follow-up time if there is no PCM,
> # and PCM time otherwise
> etime <- with(mgus2, ifelse(pstat==0, futime, ptime))
>
> # event variable will be 0 for censor or 2 for death
> # if there is no PCM, and 1 for PCM
> event <- with(mgus2, ifelse(pstat==0, 2*death, 1))
>
> # event variable must be a factor for multistate
> event <- factor(event, 0:2,
                  labels=c("censor", "pcm", "death"))
```

```
> # confirm coding makes sense
> library(arsenal)
> summary(freqlist(~ event+pstat+death, data=mgus2))
```

event	pstat	death	Freq	Cumulative Freq	Percent	Cumulative I
censor	0	0	409	409	29.55	
pcm	1	0	12	421	0.87	
		1	103	524	7.44	
death	0	1	860	1384	62.14	

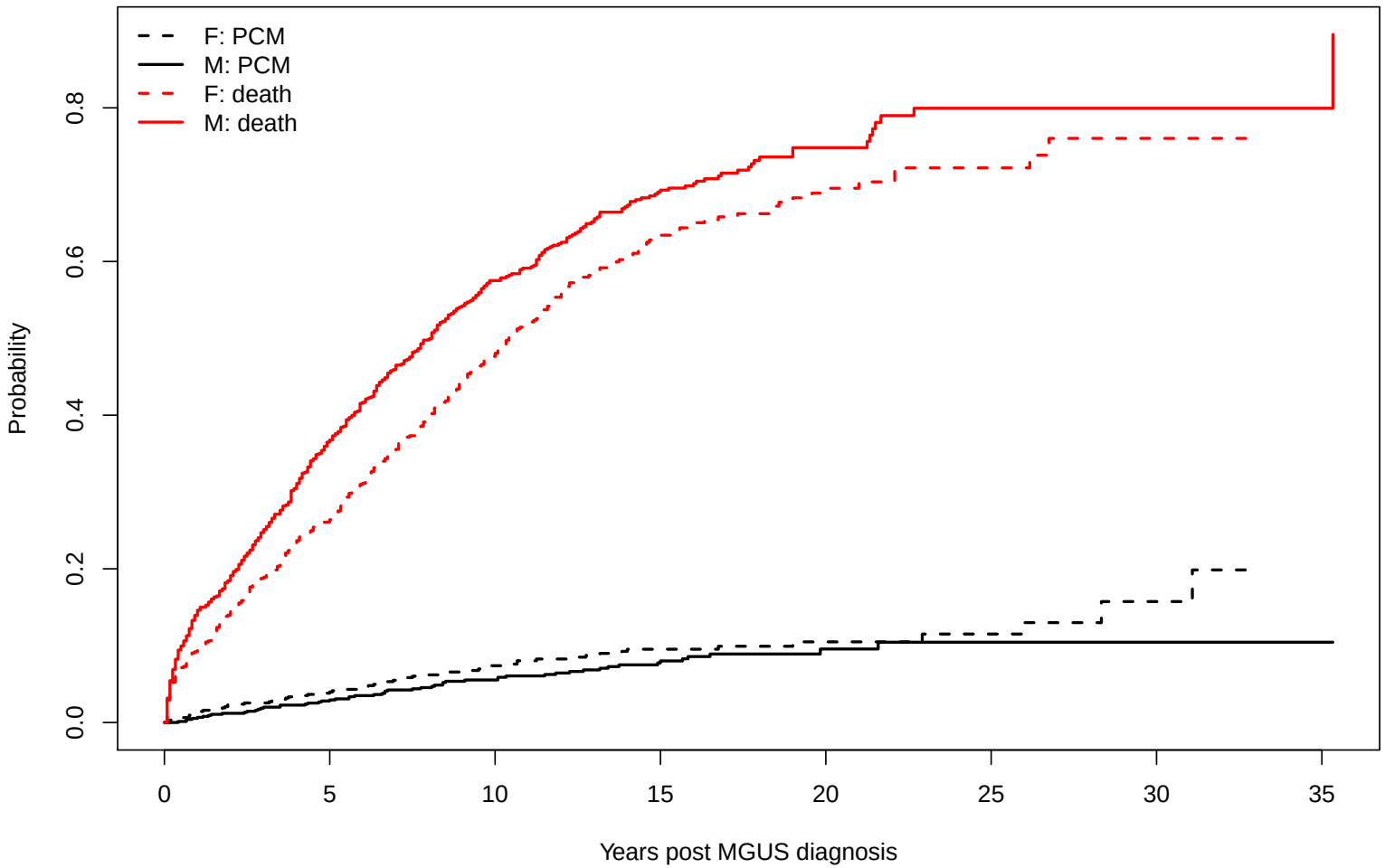
MAYO CLINIC

Example: MGUS - Aalen-Johansen estimate

```
> fit <- survfit(Surv(etime, event) ~ sex, data=mgus2)
> plot(fit, col=c(1,1,2,2), lty=c(2,1,2,1), xscale=12,
     xlab="Years post MGUS diagnosis", ylab="Probability")
>
> # short-cut for creating the group labels
> temp.label <- c(outer(c("F:","M:"),
                        c("PCM","death"), paste))
> legend("topleft", c(temp.label),
       col=c(1,1,2,2), lty=c(2,1,2,1), lwd=2, bty='n')
```

MAYO CLINIC

Example: MGUS - Aalen-Johansen estimate



MAYO CLINIC

Double check the legend. . .

- Pick a time on the x-axis and confirm results

```
> summary(fit, time=20, scale=12)
Call: survfit(formula = Surv(etime, event) ~ sex, data = mgus2)

                sex=F
   time   n.risk  n.event  P((s0))   P(pcm) P(death)
 20.0000 540.0000  93.0000   0.8523   0.0175   0.1302

                sex=M
   time   n.risk  n.event  P((s0))   P(pcm) P(death)
 20.0000 621.0000 137.0000   0.8181   0.0106   0.1713
```

MAYO CLINIC

Example: MGUS - Aalen-Johansen estimate

- Subset to PCM event

```
> fit$states # columns
[1] "(s0)"  "pcm"   "death"
> fit$strata # rows
sex=F sex=M
  227   227
```

- Do not plot this curve

```
> plot(fit, noplot="(s0)")
```

MAYO CLINIC

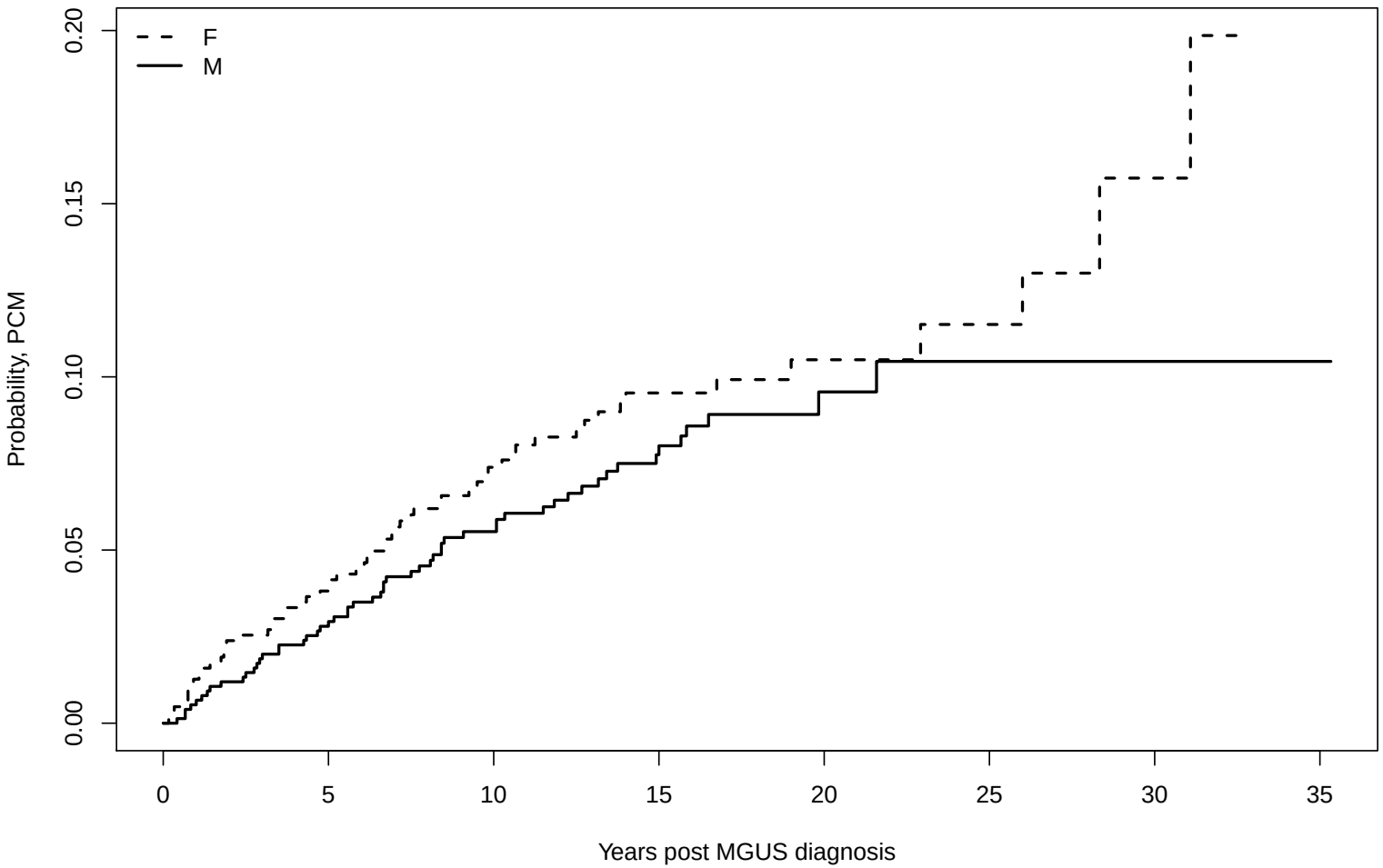
Example: MGUS - Aalen-Johansen estimate

```
> fit[,2] # plot PCM
Call: survfit(formula = Surv(etime, event) ~ sex, data = mgus2)

              n nevent   rmean*
sex=F, pcm 631     59 42.77078
sex=M, pcm 753     56 31.82962
   *mean time in state, restricted (max time = 424 )
```

```
> plot(fit[,2], col=1, lty=2:1,
      xscale=12, lwd=2,
      xlab="Years post MGUS diagnosis", ylab="Probability, PCM")
> legend("topleft", c('F','M'),
       col=1, lty=2:1, lwd=2, bty='n')
```

MAYO CLINIC

Example: MGUS - Aalen-Johansen estimate



MAYO CLINIC

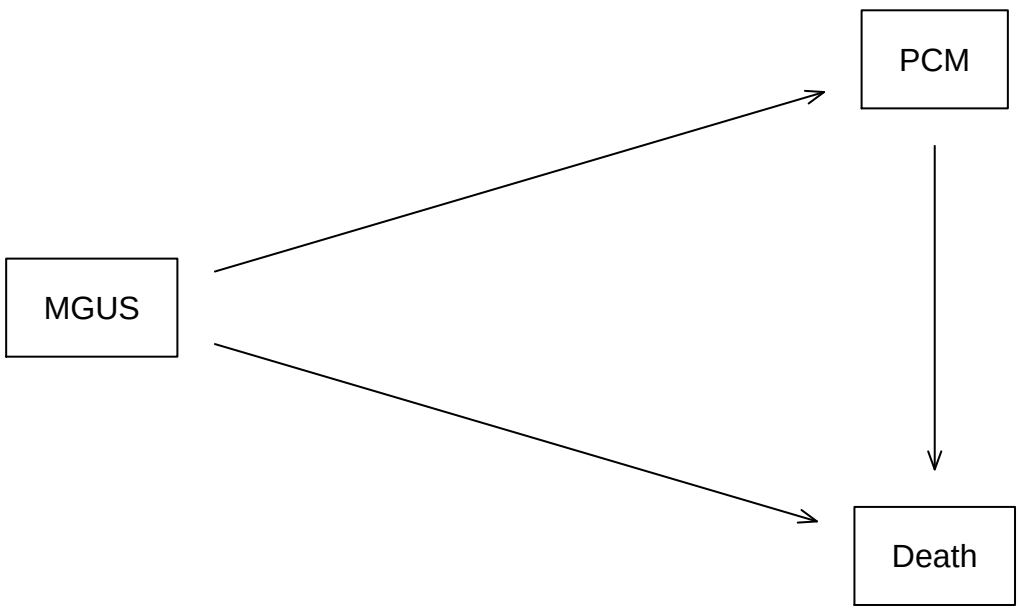
Example: MGUS data - Cox Model

- ID is required

```
> cfit <- coxph(Surv(etime, event) ~ sex, data=mgus2, id=id)
> cfit
Call:
coxph(formula = Surv(etime, event) ~ sex, data = mgus2, id = id)


1:2         coef exp(coef) se(coef) robust se      z     p
  sexM -0.05938   0.94235  0.18723   0.18733 -0.317 0.751


1:3         coef exp(coef) se(coef) robust se     z        p
  sexM 0.22800   1.25608  0.06900   0.06869 3.319 0.000902

 States:  1= (s0), 2= pcm, 3= death
```

MGUS: Modify the Diagram

	MGUS	PCM	Death
MGUS	0	1	1
PCM	0	0	1
Death	0	0	0



MGUS: Multistate

- Are there subjects where PCM and death occur at the same time?

	id	ptime	futime	pstat	death
1	190	101	101	1	1
2	383	147	147	1	1
3	619	81	81	1	1
4	780	39	39	1	1
5	1013	128	128	1	1
6	1037	16	16	1	1
7	1098	74	74	1	1
8	1104	8	8	1	1
9	1262	67	67	1	1

MAYO CLINIC

- What to do with 9 subjects who have PCM & death at the same time?
  - Cannot have a time of length 0, so push progression back by 0.1 month.

```
> # if subject progresses and death occurs on the same day,
> # subtract .1 month from the progression time
> ptemp <- with(mgus2, ifelse(ptime==futime & pstat==1,
                              ptime-.1, ptime))
```

MAYO CLINIC

```
> # the first call to tmerge sets the time range,
> # so start with the longest times, which are for death
> newdata <- tmerge(mgus2, mgus2, id=id,
                    death=event(futime, death))
>
> # now add additional observations for progressions
> newdata <- tmerge(newdata, mgus2, id,
                    pcm = event(ptemp, pstat))
```

MAYO CLINIC

MGUS: Multistate

```
> attr(newdata, "tcount")
```

```
      early late gap within boundary lead trail tied
death     0    0   0      0        0    0  1384    0
pcm       0    0   0    115        0    0  1269    0
```

MGUS: Multistate

```
> with(newdata, table(death, pcm))
      pcm
death   0   1
    0 421 115
    1 963   0
> # first create a 0,1,2 event variable and make it a factor
> temp <- with(newdata, ifelse(death==1, 2, pcm))
> newdata$event <- factor(temp, 0:2,
                          labels=c("censor", "pcm", "death"))
```

MAYO CLINIC

```
> survcheck(Surv(tstart, tstop, event) ~ sex,
            data=newdata, id=id)
```

```
Call:
survcheck(formula = Surv(tstart, tstop, event) ~ sex, data = new
    id = id)
1384 subjects available for analysis
Transitions table:
       to
from    pcm death (censored)
  (s0)  115   860        409
  pcm     0   103         12
  death   0     0          0
```

```
Number of subjects with 0, 1, ... transitions to each state:
       count
state      0   1   2
  pcm   1269 115   0
  death  421 963   0
  (any)  409 872 103
```

MGUS: Multistate

```
> cfit <- coxph(Surv(tstart,tstop,event)~sex, data=newdata, id=i

Call:
coxph(formula = Surv(tstart, tstop, event) ~ sex, data = newdata
    id = id)


1:2        coef exp(coef) se(coef) robust se      z     p
  sexM -0.05934   0.94238  0.18723   0.18733 -0.317 0.751


1:3        coef exp(coef) se(coef) robust se    z        p
  sexM 0.22802   1.25610  0.06900   0.06869 3.32 0.000901


2:3        coef exp(coef) se(coef) robust se     z   p
  sexM 0.02822   1.02862  0.20408   0.22429 0.126 0.9
```

NAFLD

- A. Allen, Non-alcoholic fatty liver disease incidence and impact on metabolic burden and death, a 20 year community study. Hepatology 2018, 67:1726–1736.
- The prevalence of non-alcoholic fatty liver disease (NASH) has risen to 24%.
- Now the most common cause of chronic liver disease.
- Diagnosed with abdominal MRI.
- NASH = NAFLD + inflammation requires biopsy for diagnosis.

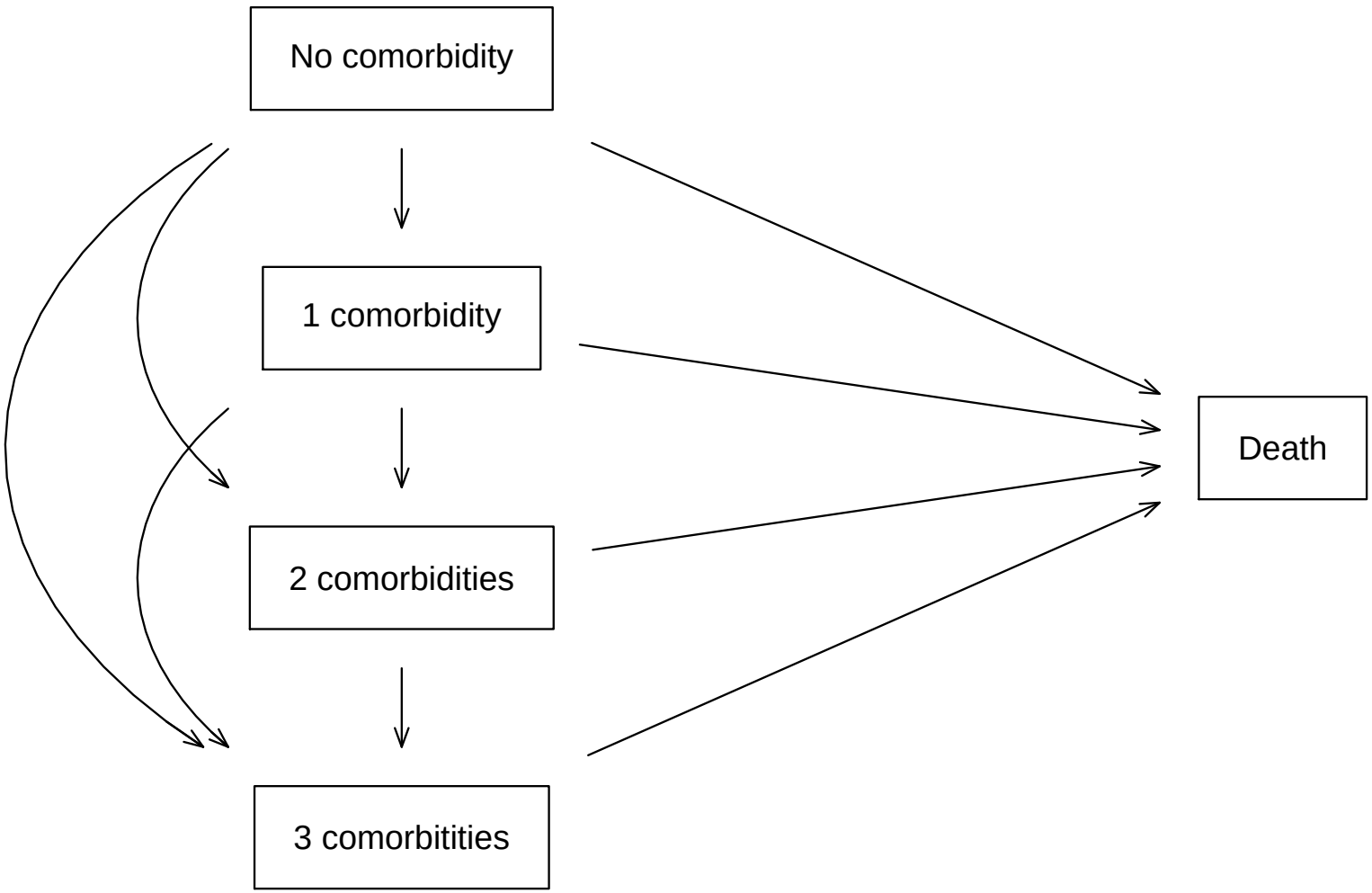
MAYO CLINIC

NAFLD Study

- All NAFLD diagnosis from 1997 to 2014 in Olmsted County, Minnesota.
- Utilize the Rochester Epidemiology Project
- One year delay.
- 4 controls matched on age and sex, then followed forward until the analysis date.
- 3864 cases of NAFLD and 14016 controls, 331 overlap.

MAYO CLINIC

NAFLD: Target



NAFLD: Data

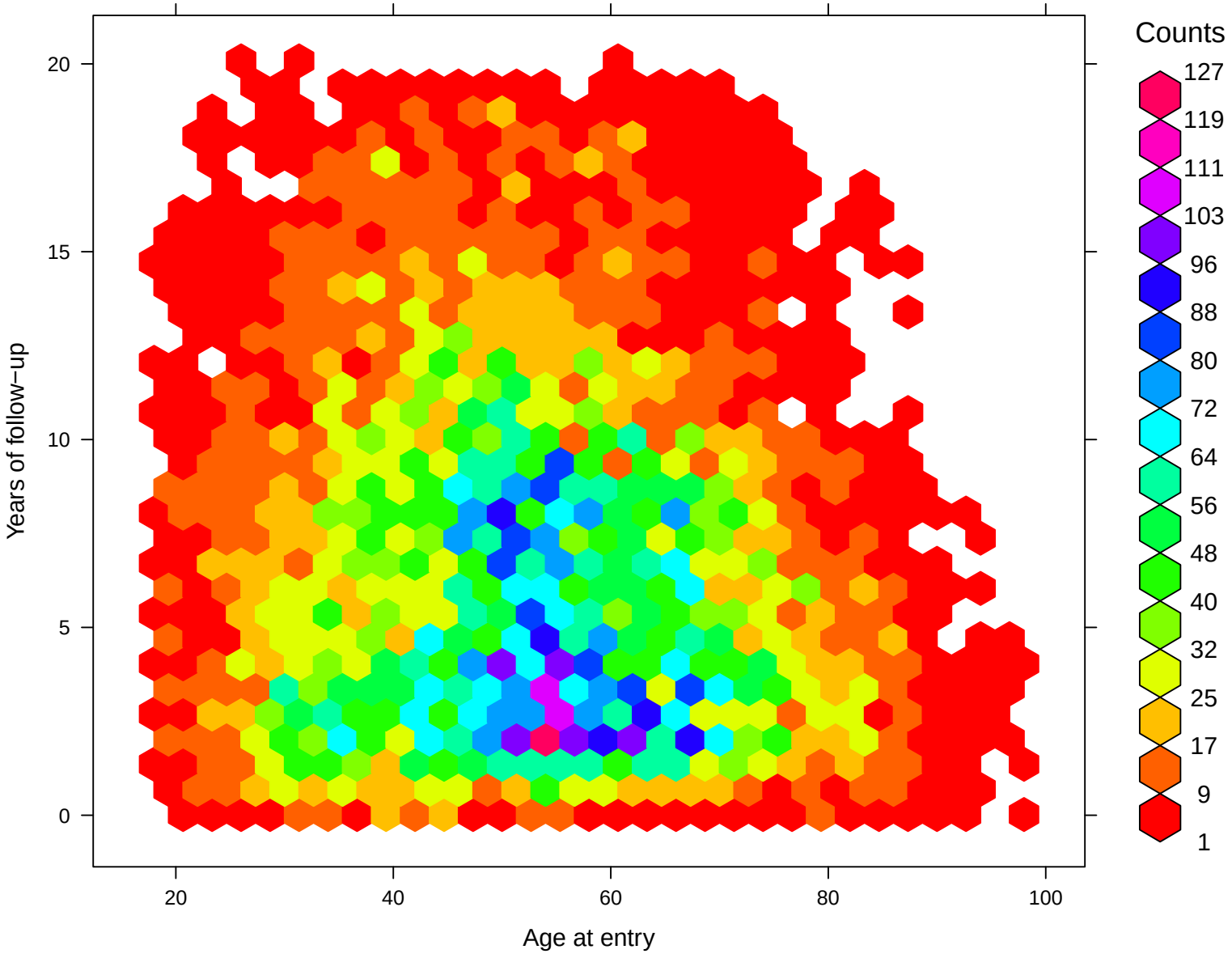
- `nafld1`: One observation per subject. Baseline covariates plus follow-up time and death.
- `nafld2`: Variables of id, days, test, and value. Contains selected tests and clinical observations.
- `nafld3`: Variables of id, days, and event type. One observation for each outcome: occurrence of NASH, hypertension, diabetes, etc.
- To anonymize patients, all dates have been replaced with "days since first enrollment".

MAYO CLINIC

NAFLD: Data

- Metabolic comorbidities are diabetes, hypertension, and dyslipidemia
- Focus on a model with 0, 1, 2, 3, of these + death
- The real work is in building and checking a dataset; the fits will be easy.

MAYO CLINIC


Years of follow−up
0
5
10
15
20
20
40
60
80
100
Age at entry
Counts
127
119
111
103
96
88
80
72
64
56
48
40
32
25
17
9
1

MAYO CLINIC

NAFLD: tmerge

```
> keep <- c("id", "age", "male", "bmi", "ntime")
> data1 <- tmerge(nafld1[, keep], nafld1, id,
                  death= event(futime, status))
> data1 <- tmerge(data1, subset(nafld3, event=="nafld"), id,
                  nafld = tdc(days))
> data1 <- tmerge(data1, subset(nafld3, event=="diabetes"), id,
                  diab= tdc(days), e1= event(days))
> data1 <- tmerge(data1, subset(nafld3, event=="htn"), id,
                  htn= tdc(days), e2= event(days))
> data1 <- tmerge(data1, subset(nafld3, event=="dyslipidemia"),
                  dyslip = tdc(days), e3= event(days))
```

MAYO CLINIC

```
> attr(data1, "tcount")
```

```
        early late gap within boundary lead trail tied
death       0    0   0      0        0    0 17549    0
nafld       0   13   0    318        0 3533     0    0
diab     2393    0   0   1058        0    1     0    0
e1       2393    0   0      0     1058    1     0    0
htn      5022    0   0   2045       24    1     5    0
e2       5022    0   0      0     2069    1     5    0
dyslip   8663    0   0   1713       82    2     2    0
e3       8663    0   0      0     1795    2     2    0
```

MAYO CLINIC

NAFLD: Four row subject from data1

	id	age	tstart	tstop	nafld	htn	diab	dyslip	death
159	135	40	0	355	1	0	0	0	0
160	135	40	355	2133	1	0	0	1	0
161	135	40	2133	3220	1	1	0	1	0
162	135	40	3220	5269	1	1	1	1	0

MAYO CLINIC

Same subject, nafld3

```
     id days         event
252 135    0         nafld
253 135  355 dyslipidemia
254 135 2133           htn
255 135 2343  sleep apnea
256 135 3220     diabetes
```

NAFLD: tmerge

```
> test <- tmerge(nafld1[, 1:2], nafld1, id,
                 death = event(futime, status))
> attr(test, "tcount")
      early late gap within boundary leading trailing tied
death     0    0   0      0        0       0    17549    0
>
> subset(test, id==135)
     id age tstart tstop death
135 135  40      0  5269     0
```

NAFLD: tmerge

```
> test <- tmerge(nafld1[, 1:2], nafld1, id,
                 death = event(futime, status))
> test <- tmerge(test, subset(nafld3, event=="nafld"), id,
                 nafld = tdc(days))
>
> attr(test, "tcount")
      early late gap within boundary leading trailing tied
death     0    0   0      0        0       0    17549    0
nafld     0   13   0    318        0    3533        0    0
>
> subset(test, id==135)
     id age tstart tstop death nafld
138 135  40      0  5269     0     1
```

MAYO CLINIC

NAFLD: tmerge

```
> test <- tmerge(nafld1[, 1:2], nafld1, id,
                 death = event(futime, status))
> test <- tmerge(test, subset(nafld3, event=="nafld"), id,
                 nafl = tdc(days))
> test <- tmerge(test, subset(nafld3, event=="diabetes"), id,
                  diab= tdc(days), e1= event(days))
> attr(test, "tcount")
      early late gap within boundary leading trailing tied
death     0    0   0      0        0       0    17549    0
nafl      0   13   0    318        0    3533        0    0
diab   2393    0   0   1058        0       1        0    0
e1     2393    0   0      0     1058       1        0    0
>
> subset(test, id==135)
     id age tstart tstop death nafl diab e1
142 135  40      0  3220     0    1    0  1
143 135  40   3220  5269     0    1    1  0
```

NAFLD: tmerge

```
> test <- tmerge(test, subset(nafld3, event=="htn"), id,
                 htn= tdc(days))
> attr(test, "tcount")
      early late gap within boundary leading trailing tied
death     0    0   0      0        0       0    17549    0
nafl      0   13   0    318        0    3533        0    0
diab   2393    0   0   1058        0       1        0    0
e1     2393    0   0      0     1058       1        0    0
htn    5022    0   0   2045       24       1        5    0
>
> subset(test, id==135)
     id age tstart tstop death nafl diab e1 htn
155 135  40      0  2133     0    1    0  0   0
156 135  40   2133  3220     0    1    0  1   1
157 135  40   3220  5269     0    1    1  0   1
```

NAFLD: tmerge

```
> test <- tmerge(test, subset(nafld3, event=="dyslipidemia"), id
                 lip= tdc(days), e3= event(days))
> attr(test, "tcount")
      early late gap within boundary leading trailing tied
death     0    0   0      0        0       0    17549    0
nafl      0   13   0    318        0    3533        0    0
diab   2393    0   0   1058        0       1        0    0
e1     2393    0   0      0     1058       1        0    0
htn    5022    0   0   2045       24       1        5    0
lip    8663    0   0   1713       82       2        2    0
e3     8663    0   0      0     1795       2        2    0
>
> subset(test, id==135)
     id age tstart tstop death nafl diab e1 htn lip e3
159 135  40      0   355     0    1    0  0   0   0  1
160 135  40    355  2133     0    1    0  0   0   1  0
161 135  40   2133  3220     0    1    0  1   1   1  0
```

In any sufficiently large sample, any outrageous thing is likely to happen. P. Diaconis and Mosteller, Method of studying coincidences, JASA 1989.

- Someone **will** die on the same day as their diabetes diagnosis, have first NAFLD and first hypertension on the same day, or any number of other overlaps.
- Be prepared to think through these cases.

MAYO CLINIC

NAFLD: Last additions

- age1, age2: age at start and end of interval
- cstate: number of metabolic conditions so far
- endpoint: censor, 1mc, 2mc, 3mc, death

```
> data1$age1 <- with(data1, age + tstart/365.25)
> data1$age2 <- with(data1, age + tstop/365.25)
> data1$cstate <- with(data1, diab + htn + dyslip)   # TD cov
```

MAYO CLINIC

NAFLD: Last additions

```
> tcount <- with(data1, e1 + e2 + e3)
> temp2 <- with(data1, ifelse(death, 4,
                ifelse(tcount ==0, 0, cstate + tcount)))
> data1$endpoint <- factor(temp2, 0:4,
          c("censored", "1mc", "2mc", "3mc", "death"))
> data1$cstate <- factor(data1$cstate, 0:3,
                  c("0mc", "1mc", "2mc", "3mc"))
> with(data1, table(cstate, endpoint))
      endpoint
cstate censored  1mc  2mc  3mc death
   0mc     5755 1829   70    4   263
   1mc     4650    0 1843   28   243
   2mc     3784    0    0 1048   417
   3mc     2308    0    0    0   441
```

MAYO CLINIC

NAFLD: Check data

```
> survcheck(Surv(tstart, tstop, endpoint) ~ male + nafld, data=d
            id=id, istate=cstate)
```

```
Call:
survcheck(formula = Surv(age1, age2, endpoint) ~ male + nafld,
    data = data1, id = id, istate = cstate)
17549 subjects available for analysis
Transitions table:
       to
from      1mc  2mc  3mc death (censored)
  0mc    1829   70    4   263       5705
  1mc       0 1843   28   243       4567
  2mc       0    0 1048   417       3687
  3mc       0    0    0   441       2220
  death     0    0    0     0          0
```

MAYO CLINIC

```
Number of subjects with 0, 1, ... transitions to each state:
       count
state       0    1   2   3  4
  1mc   15720 1829   0   0  0
  2mc   15636 1913   0   0  0
  3mc   16469 1080   0   0  0
  death 16185 1364   0   0  0
  (any) 12733 3673 938 183 22
```

MAYO CLINIC

NAFLD: Time scale

- Time since diagnosis
  - makes some sense for the NAFLD cases
  - Time since "your number was chosen out of a hat" for the controls?
  - Age and sex need to be in the model, and the model for them needs to be *correct*
  - The population death rate ranges from .03–500 /1000 over this age span; a small lack of fit in the age*sex modeling can dominate all other covariates.
- Age as a time scale:
  - Compares like with like. We can also stratify on sex if desired.
  - Age is not a covariate
- Time since index + case-control matching compares each subject to others of the same age and sex.

MAYO CLINIC

NAFLD: Models

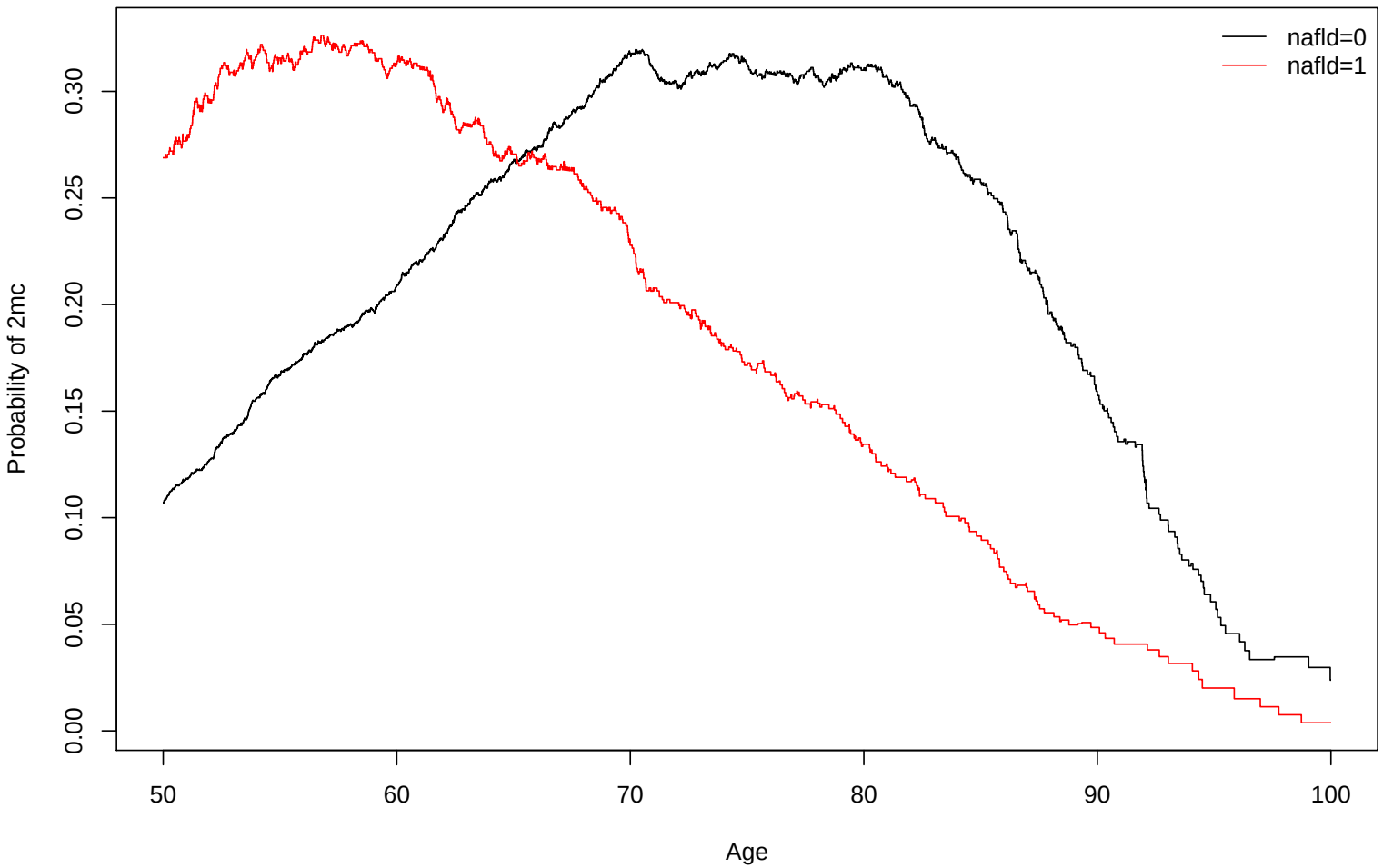
```
> nfit1 <- coxph(Surv(age1, age2, death) ~ male + nafld,
                 data=data1)
> nfit2 <- coxph(Surv(age1, age2, death) ~ male + nafld +
                   as.numeric(cstate),
                 data=data1)
> nfit3 <- coxph(Surv(age1, age2, death) ~ male +
                   strata(cstate)/nafld,  data= data1)
> nfit4a <- coxph(Surv(age1, age2, endpoint %in% c("1mc", "2mc",
                     strata(male) + nafld,
                 data=data1, subset= (cstate=="0mc"))
> nfit4b <- coxph(Surv(age1, age2, endpoint %in% c("2mc", "3mc")
                     strata(male) + nafld,
                 data=data1, subset= (cstate== "1mc"))
> nfit4c <- coxph(Surv(age1, age2, endpoint=="3mc") ~
                     strata(male) + nafld,
                 data=data1, subset= (cstate=="2mc"))
```

MAYO CLINIC

NAFLD: Aalen-Johansen curves

```
> multi <- survfit(Surv(age1, age2, endpoint) ~ nafld, data=data
                   istate=cstate, id=id, se=FALSE, start.time=50)
> multi$states
[1] "0mc"   "1mc"   "2mc"   "3mc"   "death"
> plot(multi[,3], col=1:2, xlab='Age', ylab='Probability of 2mc'
> legend("topright", legend=names(multi$strata),
         col=1:2, lty=1, bty='n')
```

MAYO CLINIC


nafld=0
nafld=1
Probability of 2mc
0.00
0.05
0.10
0.15
0.20
0.25
0.30
50
60
70
80
90
100
Age

MAYO CLINIC