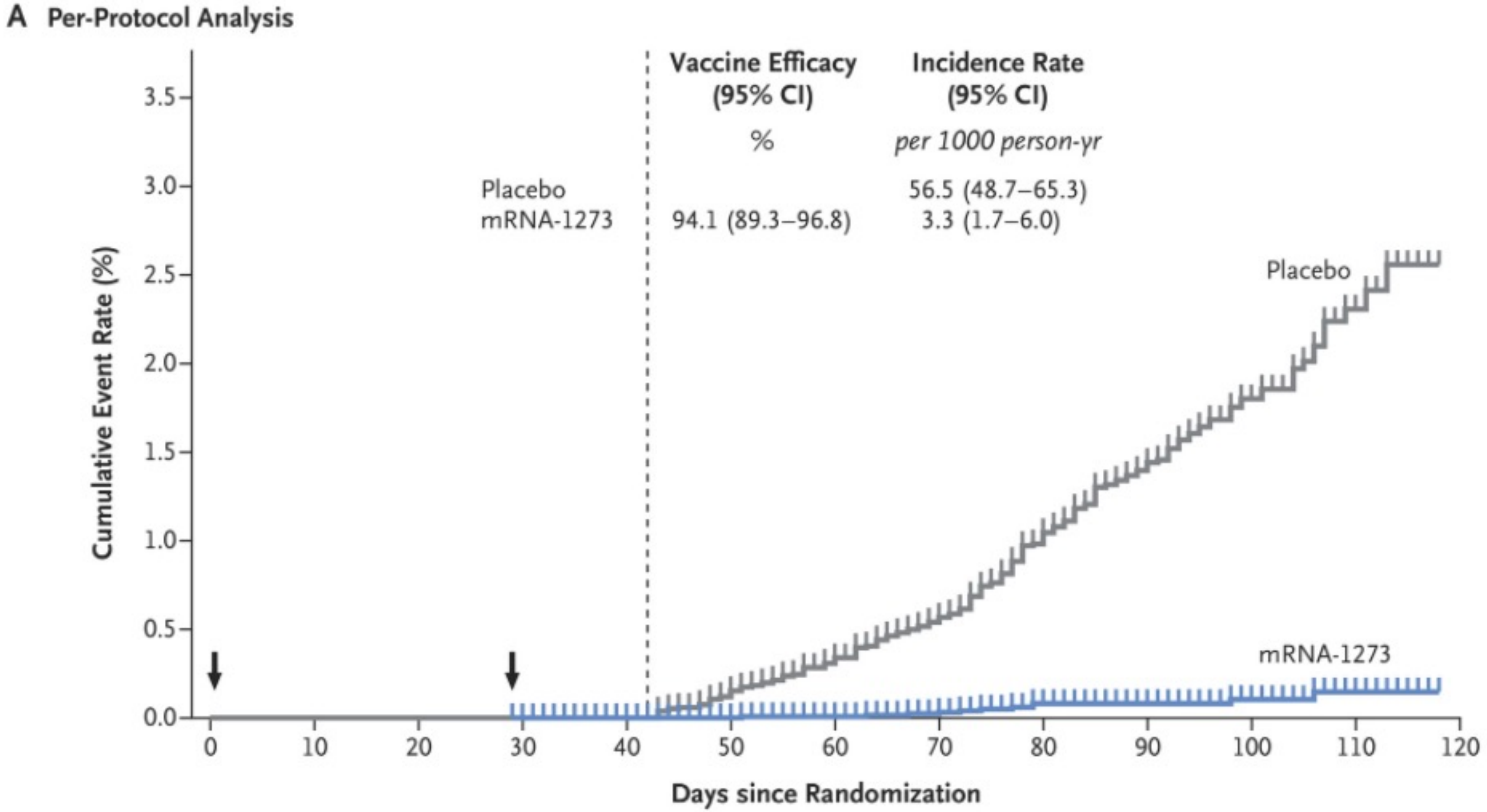
Survival Analysis

February 16, 2022

Why survival analysis?

- "time to event" data

- commonly used in clinical trial settings - time to outcome (usually something bad)



No. at Risk													
Placebo	14,073	14,073	14,073	14,072	13,416	12,992	12,361	11,147	9474	6563	3971	1172	0
mRNA-1273	14,134	14,134	14,134	14,133	13,483	13,073	12,508	11,315	9684	6721	4094	1209	0

Baden et al. (2020) NEJM "Efficacy and safety of the mRNA-1273 SARS-CoV2 vaccine"

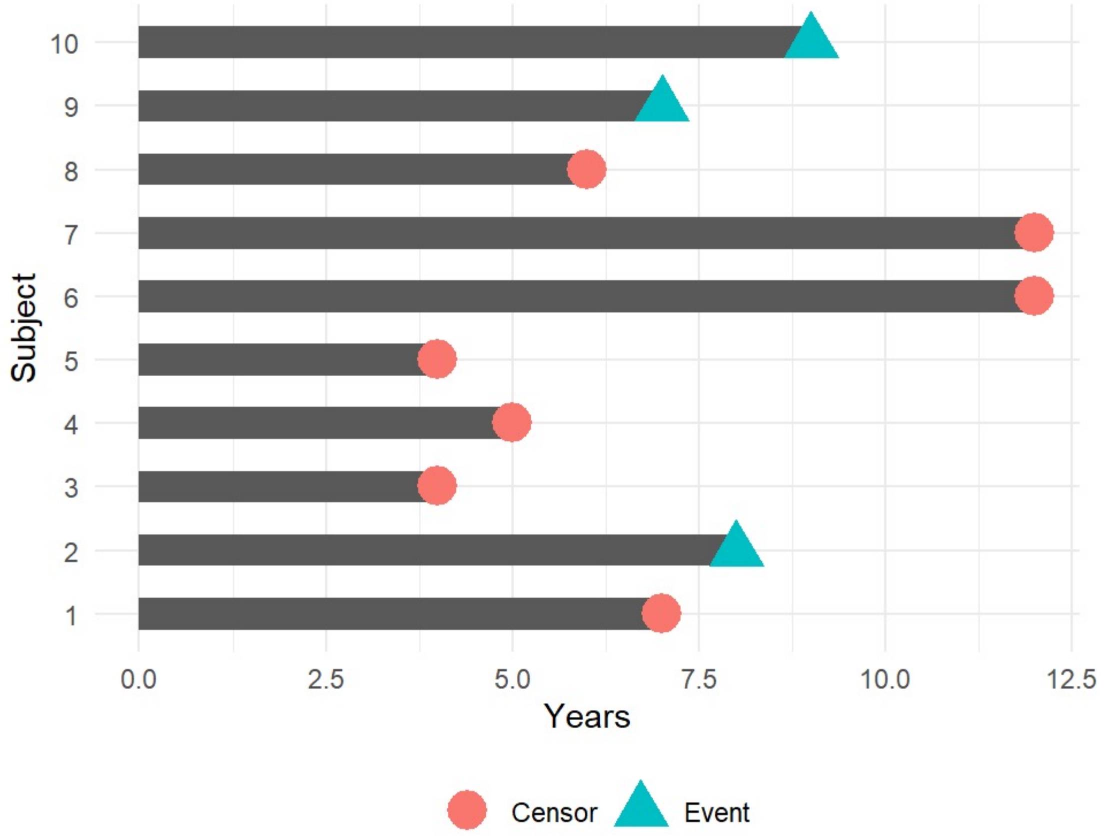
Why survival analysis?

- Why do we need a special method for time to event data?

- **censoring** - we may end observation before we determine whether an event has occurred or not. Ignoring censoring generates inaccurate estimates of probability of an outcome

- typically data are **right-censored** - outcome not observed by end of observation period. (Data can also be left-censored or interval-censored)

Censored survival data



How to compute proportion of subjects that are event-free at 10 years?

6, 7 were event-free at 10 years

2, 9, and 10 had the event before 10 years

1, 3, 4, 5, 8 were censored before 10 years. Data collection ended, but we don't know whether or not they had the event – how to incorporate into the estimate?

Survival data

- "time to event" data have 2 components:
  - observed time (minimum of event time and censoring time)
  - event indicator: convention is 1 if event observed, 0 if censored

- R uses `Surv` function (CAPITAL S) from the `survival` package to compose a "survival object"
  - 1 if event observed, 0 if censored
  - will also accept TRUE/FALSE (TRUE = event) or 1/2 (2 = event)
  - `Surv(time, event)`

Survival data

- `Surv` object combines times and censoring information for use in specialized functions to deal with time-to-event data

```
> Surv(fkdt$Years, fkdt$censor_01)
 [1]  7+  8   4+  5+  4+ 12+ 12+  6+  7   9
```

20220216_survival.R

Kaplan-Meier curve

- Nonparametric function (does not depend on any theoretical distribution function)
- Product of proportions known to survive at times up to and including time of event *i*

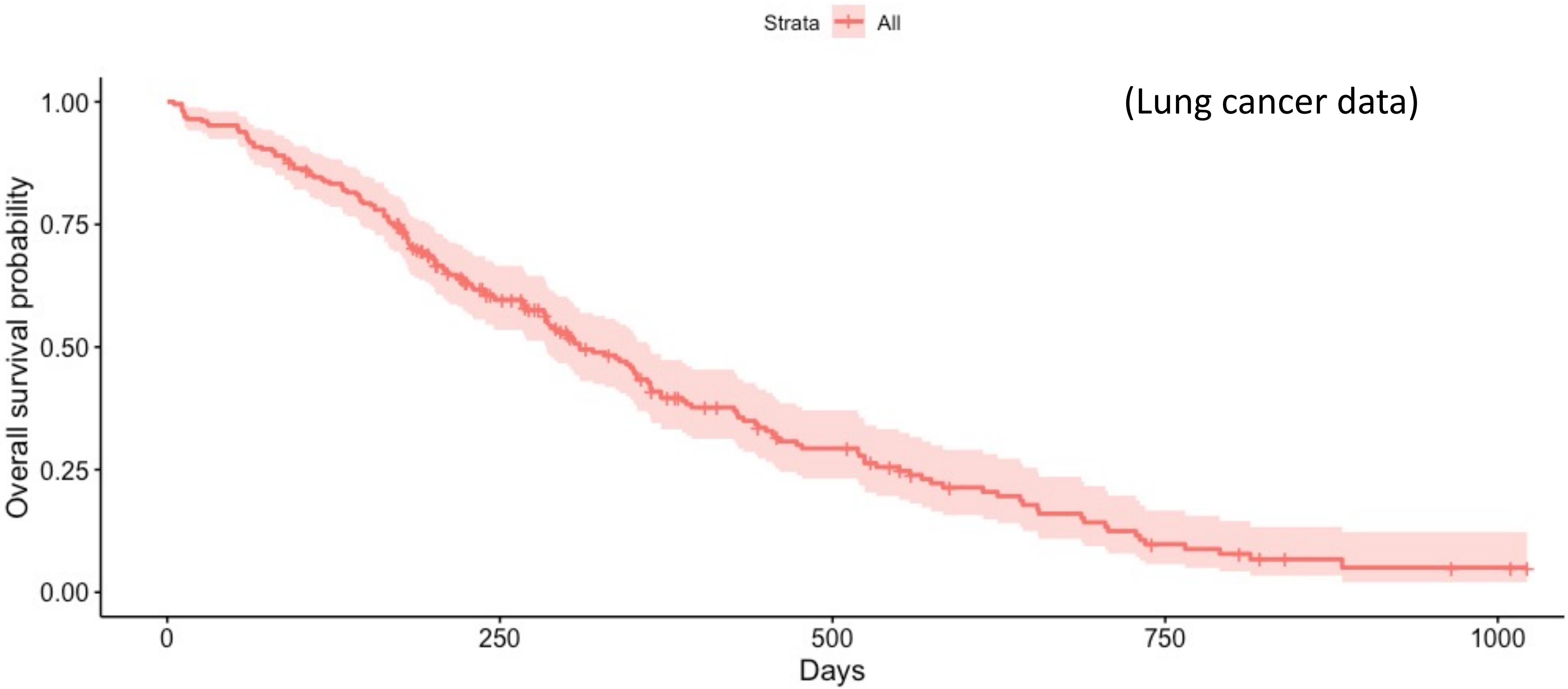
$$\text{estimate of survival at time } t = \prod_{i:\, t_i \leq t} \left(1 - \frac{\text{number of events at time } i}{\text{number known not to have had an event at time } i}\right)$$

Kaplan-Meier curve

$$\text{estimate of survival at time } t = \prod_{i:\, t_i \leq t} \left(1 - \frac{\text{number of events at time } i}{\text{number known not to have had an event by time } i}\right)$$

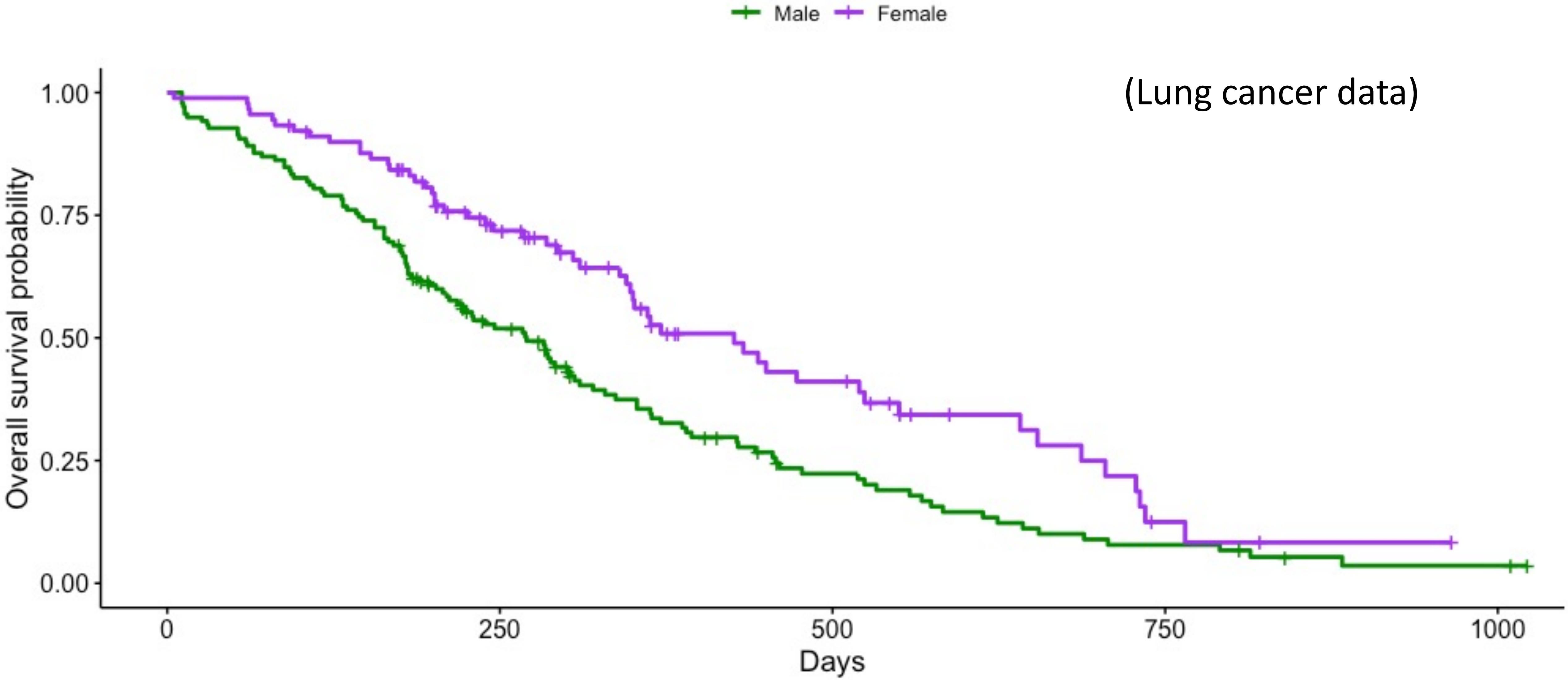
```
> Surv(fkdt$Years, fkdt$censor_01)
 [1]  7+  8   4+  5+  4+ 12+ 12+  6+  7   9
```

- First <u>event</u> is at time point 7 in our sample data
- K-M estimate at time 7 is $1 \times \left(1 - \frac{1}{6}\right)$ = 0.833
  - 4 cases censored before time 7
- K-M estimate at time 8 is $1 \times \left(1 - \frac{1}{6}\right)\left(1 - \frac{1}{4}\right)$ = 0.625
  - 2 more cases before time 8 (at time 7): 1 censored, one had an event
- K-M estimate at time 9 is $1 \times \left(1 - \frac{1}{6}\right)\left(1 - \frac{1}{4}\right)\left(1 - \frac{1}{3}\right)$ = 0.4167
  - 1 more case lost before time 9 (at time 8): 1 event
- Estimate does not change after this point because there are no more events, only censored cases



(Lung cancer data)

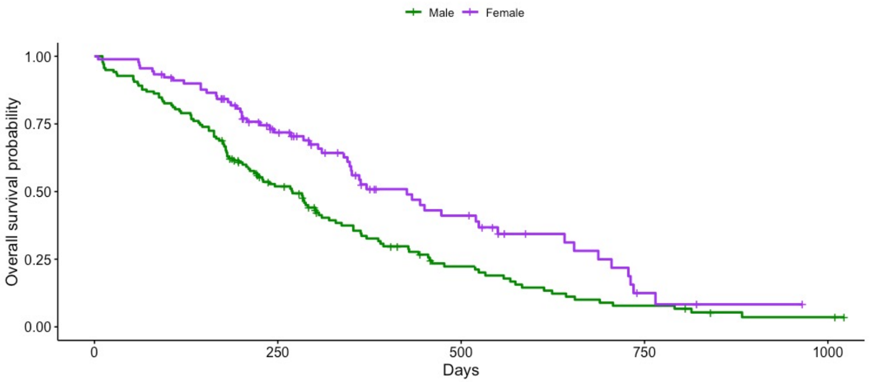
20220216_survival.R



20220216_survival.R

Comparison of survival curves

- "Log-rank" test works for comparing on single factor

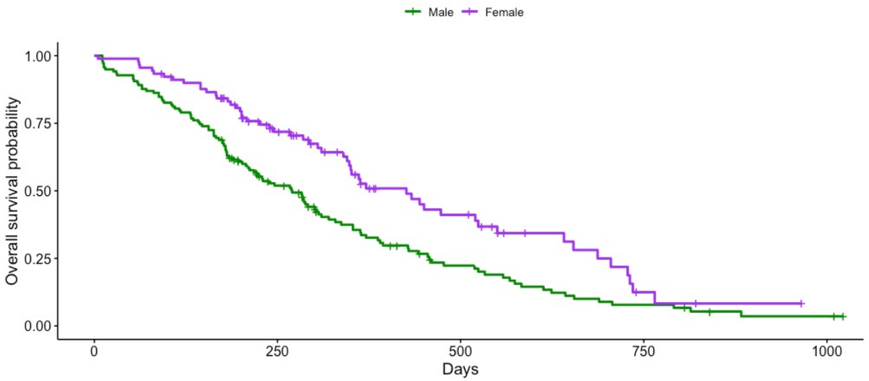



```
> survdiff(Surv(time, status) ~ sex, data = lung)
Call:
survdiff(formula = Surv(time, status) ~ sex, data = lung)

        N Observed Expected (O-E)^2/E (O-E)^2/V
sex=1 138      112     91.6      4.55      10.3
sex=2  90       53     73.4      5.68      10.3

 Chisq= 10.3  on 1 degrees of freedom, p= 0.001
```

- Cox proportional hazards regression allows more complex designs
  - Assumes ratio of hazards for any 2 individuals at any time point is constant
  - Does not allow K-M curves that cross or have different shapes (one stops but other does not)

20220216_survival.R

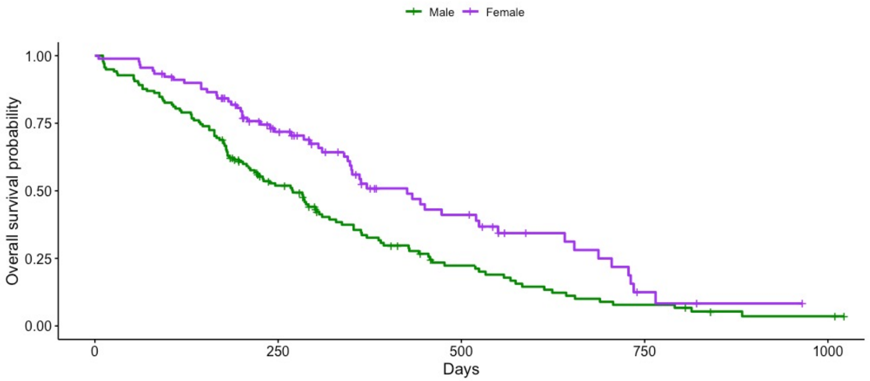
Cox proportional hazards regression




```
> coxph(Surv(time, status) ~ sex, data = lung) %>% summary()
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

Concordance= 0.579  (se = 0.021 )
Likelihood ratio test= 10.63  on 1 df,   p=0.001
Wald test            = 10.09  on 1 df,   p=0.001
Score (logrank) test = 10.33  on 1 df,   p=0.001
```

20220216_survival.R

Cox proportional hazards regression



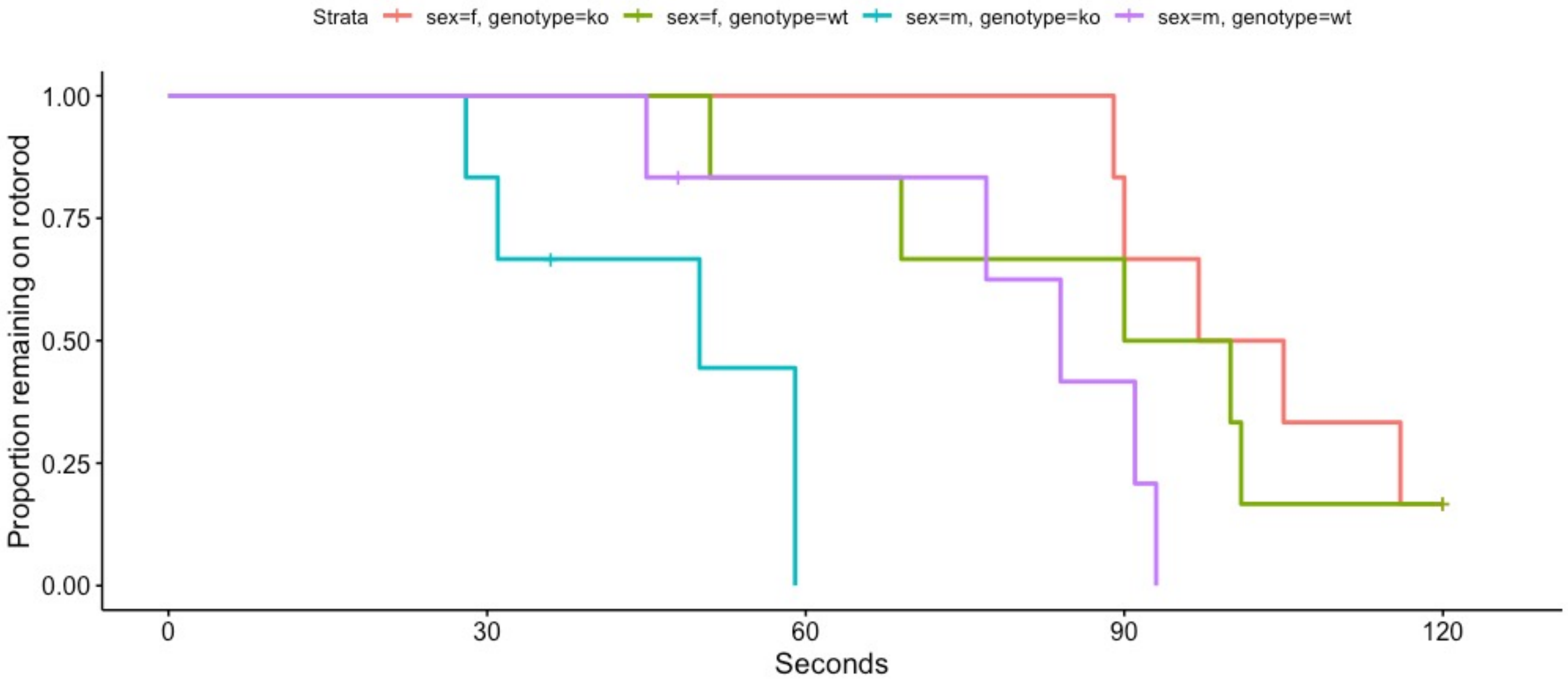

```
> coxph(Surv(time, status) ~ sex, data = lung) %>% summary()
Call:
coxph(formula = Surv(time, status) ~ sex, data = lung)

  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

Concordance= 0.579  (se = 0.021 )
Likelihood ratio test= 10.63  on 1 df,   p=0.001
Wald test            = 10.09  on 1 df,   p=0.001
Score (logrank) test = 10.33  on 1 df,   p=0.001
```

exponentiated estimated coefficient is hazard ratio

at any one time, expect 0.588 females have died for every 1 male that has died

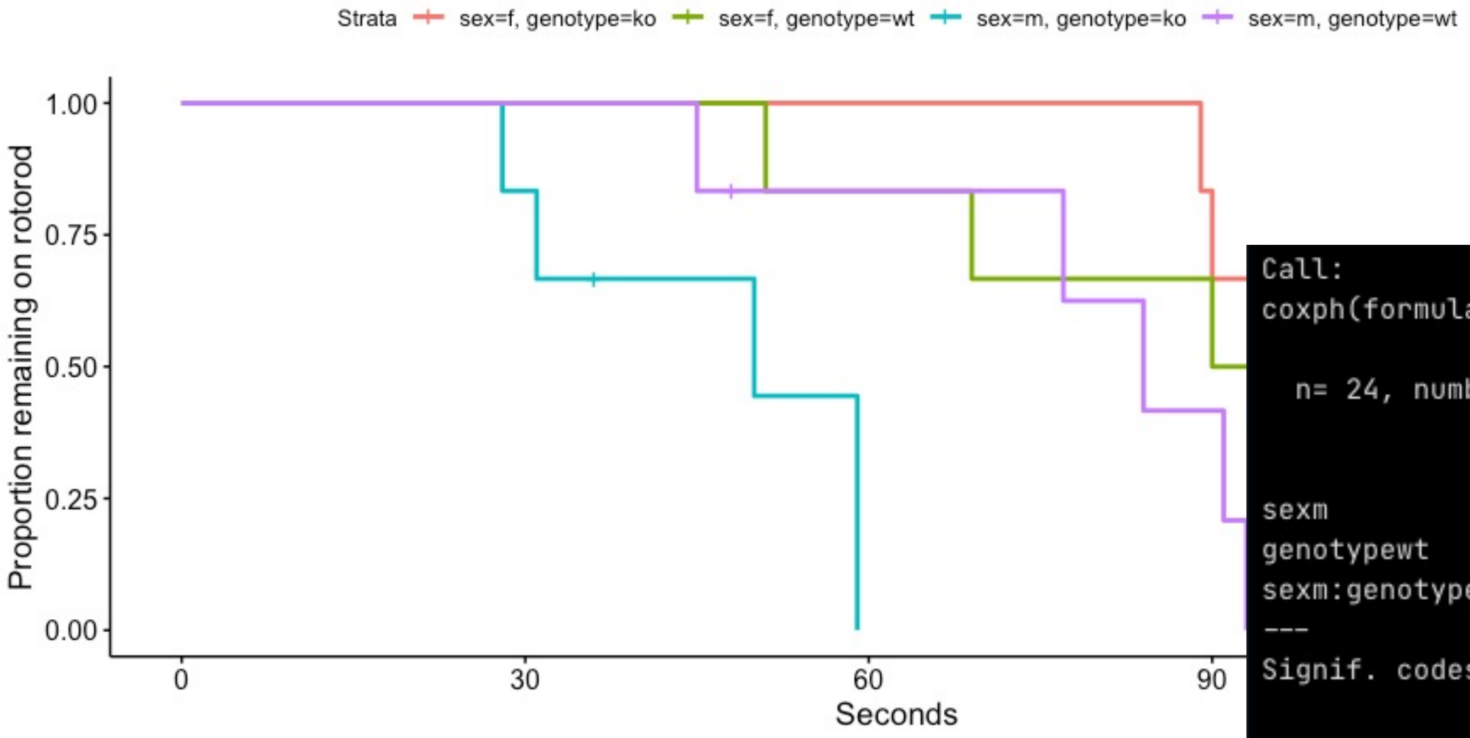
20220216_survival.R

Cox proportional hazards regression



20220216_survival.R

Cox proportional hazards regression




```
Call:
coxph(formula = Surv(time, outcome_cens) ~ sex * genotype, data = rotorod)

  n= 24, number of events= 20

                    coef exp(coef) se(coef)      z Pr(>|z|)
sexm             3.51034  33.45972  0.99699  3.521  0.00043 ***
genotypewt       0.37909   1.46095  0.63901  0.593  0.55302
sexm:genotypewt -2.43819   0.08732  1.10864 -2.199  0.02786 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                exp(coef) exp(-coef) lower .95 upper .95
sexm             33.45972    0.02989  4.741154   236.135
genotypewt        1.46095    0.68449  0.417553     5.112
sexm:genotypewt   0.08732   11.45235  0.009941     0.767

Concordance= 0.755  (se = 0.04 )
Likelihood ratio test= 15.72  on 3 df,   p=0.001
Wald test            = 13.22  on 3 df,   p=0.004
Score (logrank) test = 21.09  on 3 df,   p=1e-04
```

20220216_survival.R