

# Introduction to Time-to-Event

Mabel Carabali

Risk & Hazards

EBOH, McGill University

01-08-2023 Updated: ( 2024-09-23)

## What to do, if you want to:

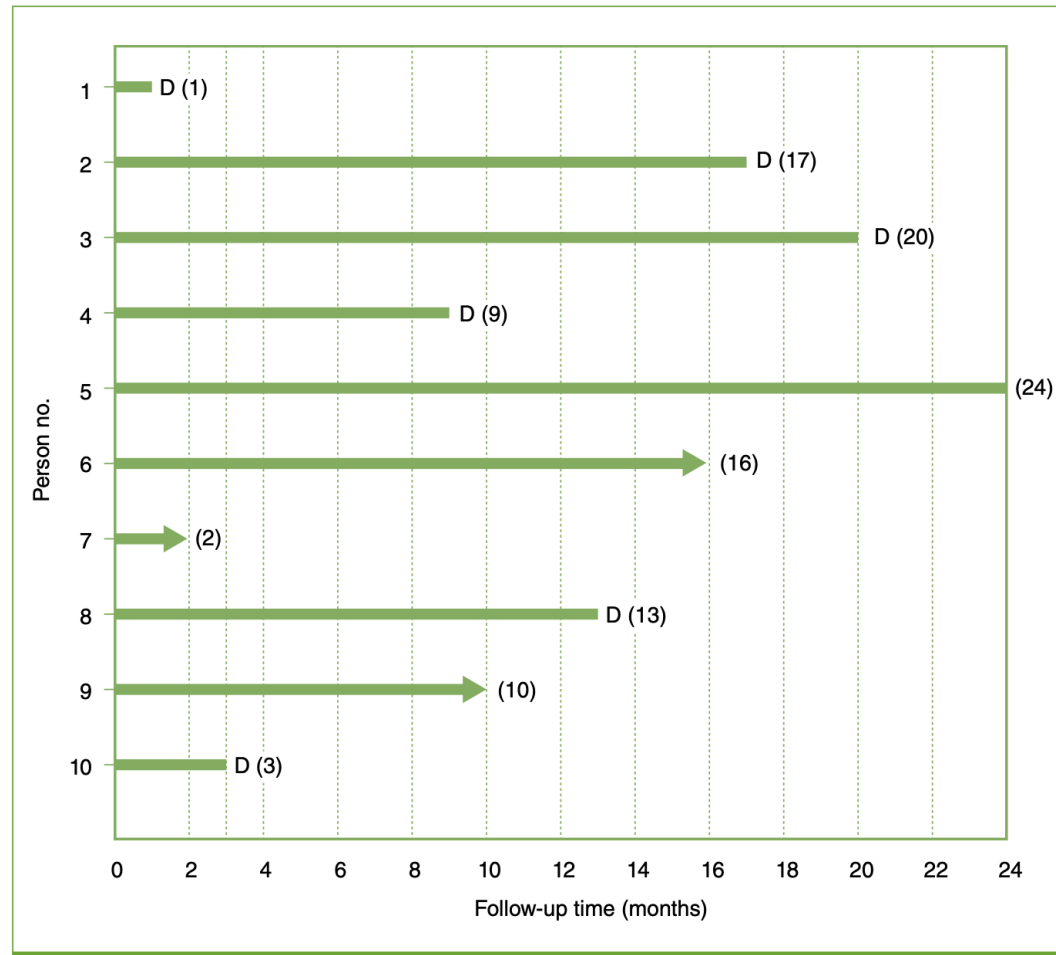
- Estimate median survival times, plot survival over time after treatment, or estimate the probability of surviving beyond a prespecified time interval (eg, 5-year survival rate)?
- Assess whether survival times are related to covariates and/or adjust for potential confounders.
- Account for censoring and avoid **lead time bias**

## Objectives

1. Review the concept of time-to-event a.k.a. "survival" analysis
2. Provide an introduction to epidemiological and statistical methods for the appropriate analysis of time-to-event data

More about this on 705

# RECALL



# Relationship between Incidence Rate and Incidence Proportion

- The method of calculating **risks over a time period with changing incidence rates** is known as **survival analysis**.
- *"The cumulative probability of the event during a given interval lasting  $m$  units of time and beginning at time  $x$ , is the proportion of new events during that period of time in which the denominator is the initial population corrected for losses".*

(Szklo M, Nieto FJ. Epidemiology : Beyond the Basics. Fourth ed.)

# Assumptions in the Estimation of Cumulative Incidence Based on Survival Analysis

- **Uniformity of Events and Losses** Within Each Interval (Classic Life Table).
- Events and losses are approximately uniform during each defined interval.
- If risk changes rapidly within a given interval, then calculating a cumulative risk over the interval is not very informative.
- The rationale underlying the method to correct for losses—that is, subtracting one-half of the losses from the denominator also depends on the assumption that losses occur uniformly.
- **Independence of censoring AND Survival**
- **No secular trends!**

# Incidence Proportion & Survival Proportion

Survival proportion: complementary to Incidence Proportion

- "Proportion of a closed population at risk that does not become diseased within a given period of time."
- $S = 1 - R$ ,

where  $R$  = incidence proportion;  $S$  = survival proportion

- Equivalently, the proportion of remaining disease free alive individuals by the end of the follow-up period.

Lash, T, et al. Modern Epidemiology 4th, Wolters Kluwer Health, 2021

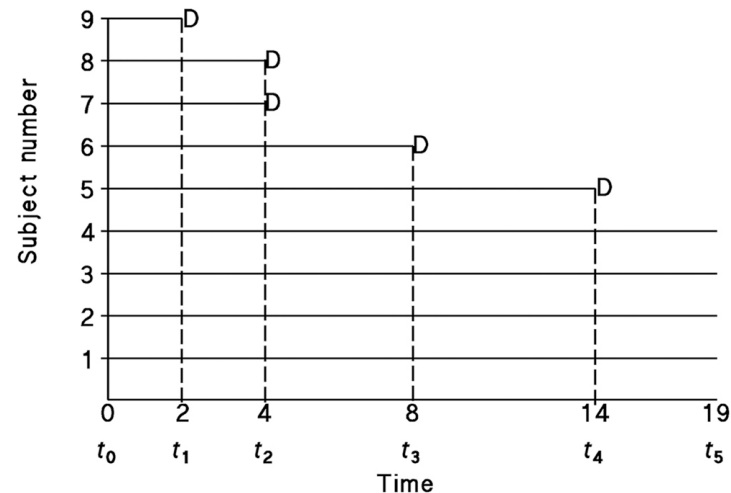
# Survival probability

- **Survival probability** at a certain time,  $S(t)$ , is a conditional probability of surviving beyond that time, given that an individual has survived just prior to that time.
- Can be estimated as the number of individuals who are alive without loss to follow-up at that time, divided by the number of individuals who were alive just prior to that time.
- The **Kaplan-Meier** estimate of survival probability is the product of these conditional probabilities up until that time.
- At time 0, the survival probability is 1, i.e.  $S(t_0) = 1$

# Product Limit Formula

## Kaplan-Meier Formula

- The product limit formula shows us how to calculate the **survival proportion** (and thereby the incidence proportion) over a period of time when the risk changes.
- It shows that we need to multiply the survival proportions over all the intervals to calculate the overall survival proportion.





# Product Limit Formula

Example:

	Start (0)	2	4	8	14	19 (End)
Index ( $k$ )		1	2	3	4	5
Nb Outcomes ( $A_k$ )	0	1	2	1	1	0
Nb at Risk ( $N_k$ )						
% Surviving ( $S_k$ )						

Nine people over 20 years

## Product Limit Formula (Kaplan-Meier Formula)

$$\text{Incidence Proportion} = 1 - S$$

$$S = \prod_{k=1}^v \left( \frac{N_k - A_k}{N_k} \right)$$

where  $v$  are sub-intervals

PLF = We need to multiply the survival proportions over all the intervals to calculate the overall survival proportion.

# Product Limit Formula (Kaplan-Meier Formula)

	Start (0)	2	4	8	14	19 (End)
Index ( $k$ )		1	2	3	4	5
Nb Outcomes ( $A_k$ )	0	1	2	1	1	0
Nb at Risk ( $N_k$ )	9	9	8	6	5	4
% Surviving ( $S_k$ )		8/9	6/8	5/6	4/5	4/4
Interval length ( $N_k \Delta A_k$ )						
Person-time ( $\Delta A_k$ )						
Incidence Rate ( $IR_k$ )						

# Product Limit Formula (Kaplan-Meier Formula)

Hand calculations for 9 people followed for 19 months.

	Start (0)	2	4	8	14	19 (End)
Index ( $k$ )		1	2	3	4	5
Nb Outcomes ( $A_k$ )	0	1	2	1	1	0
Nb at Risk ( $N_k$ )	9	9	8	6	5	4
<b>% Surviving ( <math>S_k</math> )</b>		8/9	6/8	5/6	4/5	4/4
Interval length ( $N_k \Delta A_k$ )		2	2	4	6	5
Person-time ( $\Delta A_k$ )		18	16	24	30	20
Incidence Rate ( $IR_k$ )		1/18	2/16	1/24	1/30	0/20

$$S = (8/9) \times (6/8) \times (5/6) \times (4/5) \times (4/4)$$

$$S = 0.444$$

$$\text{Incidence Proportion} = 1 - S = 0.556$$

# Kaplan-Meier Formula

- The Kaplan-Meier approach involves the calculation of the probability of each event at the time it occurs.
- The denominator for this calculation is the population at risk at the time of each event's occurrence
  - The probability of each event is a "conditional probability" → conditioned on being at risk (alive and not censored) at the event time.
- The advantage of the Kaplan-Meier Product-Limit method over the life table method is that the resulting estimates do not depend on the grouping of the data (into a certain number of time intervals).
  - However, the Product-Limit method and the life table method are identical if the intervals of the life table contain at most one observation.
- Regardless of the method used in the calculation (actuarial or Kaplan-Meier), the cumulative incidence is a proportion, unitless, and its value range from 0 to 1 (or 100%)

# Kaplan-Meier Formula

$Survival = S_i$ , Cumulative probability of the event  $(1 - S_i) \rightarrow 1 - S_{24} = 1 - 0.18 = 0.82$

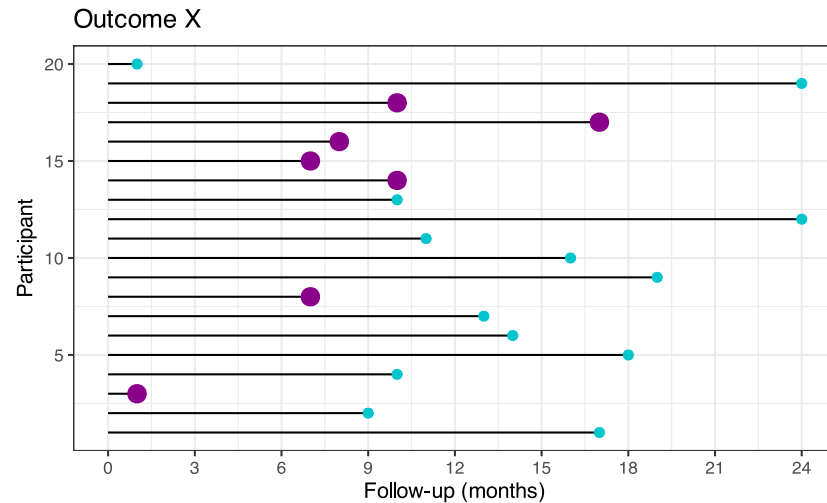
**TABLE 2-3** Calculation of Kaplan–Meier survival estimates for the example in Figure 2-3.

Time (months) (1) $i$	Number of individuals at risk (2) $n_i$	Number of events (3) $d_i$	Conditional probability of the event (4) $q_i = d_i/n_i$	Conditional probability of survival (5) $p_i = 1 - q_i$	Cumulative probability of survival* (6) $S_i$
1	10 <sup>†</sup>	1	1/10 = 0.100	9/10 = 0.900	0.900
3	8 <sup>†</sup>	1	1/8 = 0.125	7/8 = 0.875	0.788
9	7	1	1/7 = 0.143	6/7 = 0.857	0.675
13	5	1	1/5 = 0.200	4/5 = 0.800	0.540
17	3 <sup>†</sup>	1	1/3 = 0.333	2/3 = 0.667	0.360
20	2	1	1/2 = 0.500	1/2 = 0.500	0.180

\*Obtained by multiplying the conditional probabilities in column (5)—see text. † Examples of how to determine how many individuals were at risk at three of the event times (1, 3, and 17 months) are shown with vertical arrows in Figure 2-3 (Szklo M, Nieto FJ. Epidemiology : Beyond the Basics. Fourth ed.)

# Simulated Examples

## Recall from the previous scenarios



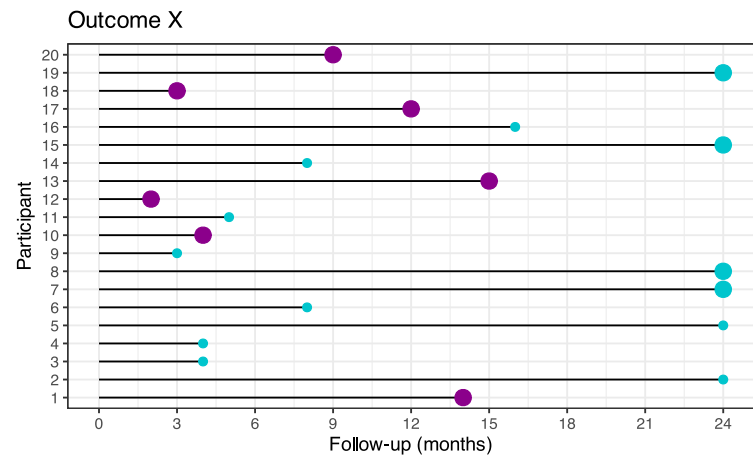
## Some coding

```
surv_object<-Surv(time = dat$t,  
                  event = dat$O2)  
  
eventKM<- survfit(surv_object ~ 1,  
                  data = dat,  
                  type="kaplan-meier")
```

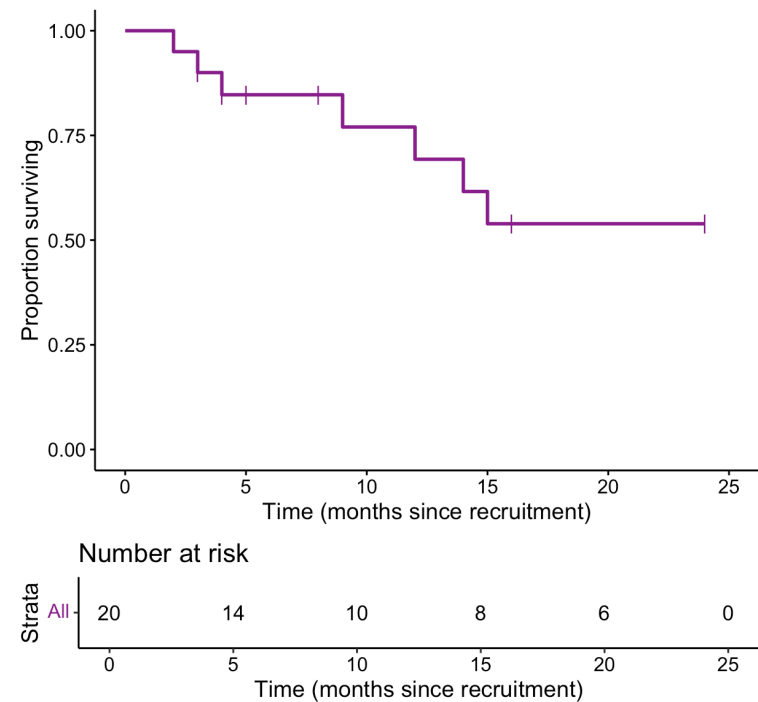
## Output

```
## Call: survfit(formula = surv_object ~ 1, data = dat, type = "kaplan-meier")  
##  
##   time n.risk n.event survival std.err lower 95% CI upper 95% CI  
##    1      20       1   0.950  0.0487    0.859    1.000  
##    7      18       2   0.844  0.0826    0.697    1.000  
##    8      16       1   0.792  0.0928    0.629    0.996  
##   10      14       2   0.679  0.1087    0.496    0.929  
##   17       6       1   0.565  0.1373    0.351    0.910
```

Other Example



Kaplan-Meier Curve - Outcome X





# Assumptions in the Estimation of Cumulative Incidence Based on Survival Analysis

- Uniformity of Events and Losses Within Each Interval (Classic Life Table).
- Events and losses are approximately uniform during each defined interval.
- If risk changes rapidly within a given interval, then calculating a cumulative risk over the interval is not very informative.
- The rationale underlying the method to correct for losses—that is, subtracting one-half of the losses from the denominator also depends on the assumption that losses occur uniformly.
- Independence of censoring AND Survival
- No secular trends!

# Exponential Formula

The exponential formula relates the incidence rate to the incidence proportion.

## Simplified:

$$Risk = 1 - e^{-Incidence\ rate \times Time}$$

**Elaborated:** Deriving the survival proportion as a function of the incidence rates for each interval:

- Total person-time at risk in the interval is  $N_k \Delta_{tk}$
- Number of Outcomes at time  $t_k$  is  $A_k$
- Number of people at Risk (alive by the end of follow-up)  $N_k$
- Incidence Rate in the time following =  $IR = \left( \frac{A_k}{N_k \Delta_{tk}} \right)$
- Incidence Proportion over same sub interval =  $IP_k = IR_k \Delta_{tk}$
- Survival Proportion for the sub interval =  $S_k = 1 - IR_k \Delta_{tk}$

# Exponential Formula

	0	2	4	8	14	19
Index ( $k$ )		1	2	3	4	5
Nb Outcomes ( $A_k$ )	0	1	2	1	1	0
Nb at Risk ( $N_k$ )	9	9	8	6	5	4
% Surviving ( $S_k$ )		8/9	6/8	5/6	4/5	4/4
<b>Interval length ( <math>N_k \Delta A_k</math> )</b>		2	2	4	6	5
Person-time ( $\Delta A_k$ )		18	16	24	30	20
<b>Incidence Rate ( <math>IR_k</math> )</b>		1/18	2/16	1/24	1/30	0/20

Survival Proportion for the sub interval =  $S_k = 1 - IR_k \Delta_{tk} \cong \exp(-IR_k \Delta_{tk})$

$$\exp(-0(5) - (\frac{1}{30})(6) - (\frac{1}{24})(4) - (\frac{2}{16})(2) - (\frac{1}{18})(2))$$

$$S_k = 0.483$$

# Exponential Formula

**Recall:**  $S = (8/9) \times (6/8) \times (5/6) \times (4/5) \times (4/4)$

$$S = 0.444$$

$$S_k = 0.483$$

$$Risk = 1 - e^{-Incidence\ rate \times Time}$$

$$Risk = 1 - e^{-IR_k \Delta_{tk}}$$

$$Risk_{ExpForm} = 1 - S_k = 0.517$$

$$Risk_c = 5/9 = 0.556 = 1 - S = \mathbf{1 - 0.444}$$

$$R = 1 - S \cong 1 - e^{-\sum_{k=1}^v IR_k \Delta_{tk}}$$

$$Risk_c = \mathbf{0.556} \cong Risk_{ExpForm} = \mathbf{0.517}$$

# Exponential Formula

## Assumptions

1. Closed population
2. Event under study is inevitable (no competing risk)
3. Number of events at each event time is a small proportion of the number at risk at that time (can be forced with fine measurement of time)

Assumptions 1 and 2 are also assumed for the product limit formula.

**Product-limit and exponential formulas:** Translate incidence-rate estimates from open populations into incidence-proportion estimates for a closed population of interest !

Lash, Timothy, L. et al. Modern Epidemiology. (4th Edition). 2020. (pg. 68

## Recall Survival probability

- **Survival probability** at a certain time,  $S(t)$ , is a conditional probability of surviving beyond that time, given that an individual has survived just prior to that time.
- Can be estimated as the number of patients who are alive without loss to follow-up at that time, divided by the number of patients who were alive just prior to that time.
- The **Kaplan-Meier** estimate of survival probability is the product of these conditional probabilities up until that time.
- At time 0, the survival probability is 1, i.e.  $S(t_0) = 1$

# Worked example

- The `survfit` function creates survival curves based on a formula.
- Let's generate the overall survival curve for the entire `lung` cohort from the `survival` package.
- Create `survfit` object and assign it to `f1`, (details about `f1` available via `names` or `str`).
- Often want to know probability of surviving a specific time (e.g. 1 year) use `summary()`.
- Produce survival curve including this information

# Get a Table 1

```
lung<- survival::lung
tab1<- lung %>%
  tbl_summary()
```

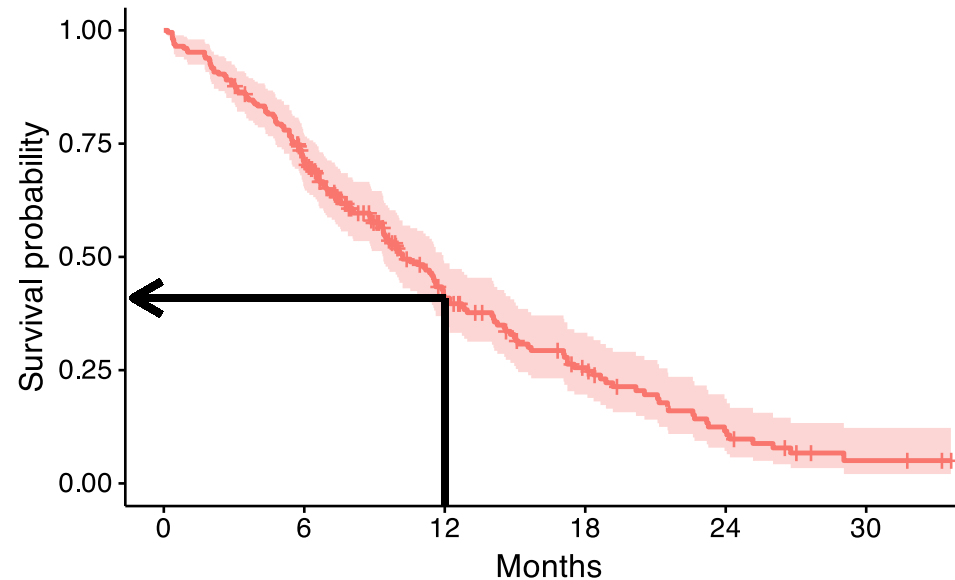
Check the package [gtsummary](#) for more information and details on formatting

Characteristic	<sup>1</sup> >N = 228 <sup>1</sup>
inst	11 (3, 16)
Unknown	1
time	256 (167, 399)
status	
1	63 (28%)
2	165 (72%)
age	63 (56, 69)
sex	
1	138 (61%)
2	90 (39%)
ph.ecog	
0	63 (28%)
1	113 (50%)
2	50 (22%)
3	1 (0.4%)
Unknown	1



## Worked example

```
f1 <- survfit(Surv(time, status) ~ 1, data = lung)
```



## Summary Results

```
summary(survfit(Surv(time, status) ~ 1, data = lung), times = 365.25)
```

```
## Call: survfit(formula = Surv(time, status) ~ 1, data = lung)
##
##   time n.risk n.event survival std.err lower 95% CI upper 95% CI
##   365     65   121    0.409  0.0358    0.345    0.486
```

# What happens if you use a naive estimate ?

121 of the 228 patients died by 1 year so:

$$\left(1 - \frac{121}{228}\right) \times 100 = 47\%$$

- You get an **incorrect** estimate of the 1-year probability of survival when you ignore the fact that 42 patients were censored before 1 year.
- Recall the **correct** estimate of the 1-year probability of survival was 41%.
- Ignoring censoring leads to an **overestimate** of the overall survival probability, because the censored subjects only contribute information for part of the follow-up time, and then fall out of the risk set, thus pulling down the cumulative probability of survival

# Comparing survival times between groups

- We can conduct between-group significance tests using a log-rank test.
- The log-rank test equally weights observations over the entire follow-up time and is the most common way to compare survival times between groups.
- Other methods weight according to early or late follow-ups (see `?survdiff` for different test options).

The `survdiff` function provides the log-rank p-value. For example, we can test whether there was a difference in survival time according to sex (SAB) in the `lung` data

```
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
```

# The Cox regression model

We may want to quantify an effect size for a single variable, or include more than one variable into a regression model to account for the effects of multiple variables.

The Cox regression model is a semi-parametric model that can be used to fit univariable and multivariable regression models that have survival outcomes.

$$h(t|X_i) = h_0(t) \exp(\beta_1 X_{i1} + \cdots + \beta_p X_{ip})$$

$h(t)$ : hazard, or the instantaneous rate at which events occur

$h_0(t)$ : underlying baseline hazard

Some key assumptions of the model:

- Non-informative censoring
- Proportional hazards
- Semi-parametric

# The Cox regression model

We can fit regression models for survival data using the `coxph` function, which takes a `Surv` object on the left hand side and has standard syntax for regression formulas in `R` on the right hand side.

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

```
## Call:
## coxph(formula = Surv(time, status) ~ sex, data = lung)
##
##           coef exp(coef) se(coef)      z      p
## sex -0.5310      0.5880   0.1672 -3.176 0.00149
##
## Likelihood ratio test=10.63  on 1 df, p=0.001111
## n= 228, number of events= 165
```

# Formatting Cox regression results

We can see a tidy version of the output using the `tidy` function from the `broom` package:

```
broom::tidy(coxph(Surv(time, status) ~ sex, data = lung), exp = TRUE) %>%  
  kable()
```

term	estimate	std.error	statistic	p.value
sex	0.5880028	0.1671786	-3.176385	0.0014912

Or use `tbl_regression` from the `gtsummary` package

```
coxph(Surv(time, status) ~ sex, data = lung) %>%  
  gtsummary::tbl_regression(exp = TRUE)
```

Characteristic	<sup>1</sup> >HR <sup>1</sup>	<sup>1</sup> >95% CI <sup>1</sup>	p-value
sex	0.59	0.42, 0.82	0.001
<sup>1</sup> HR = Hazard Ratio, CI = Confidence Interval			

# Hazard ratios

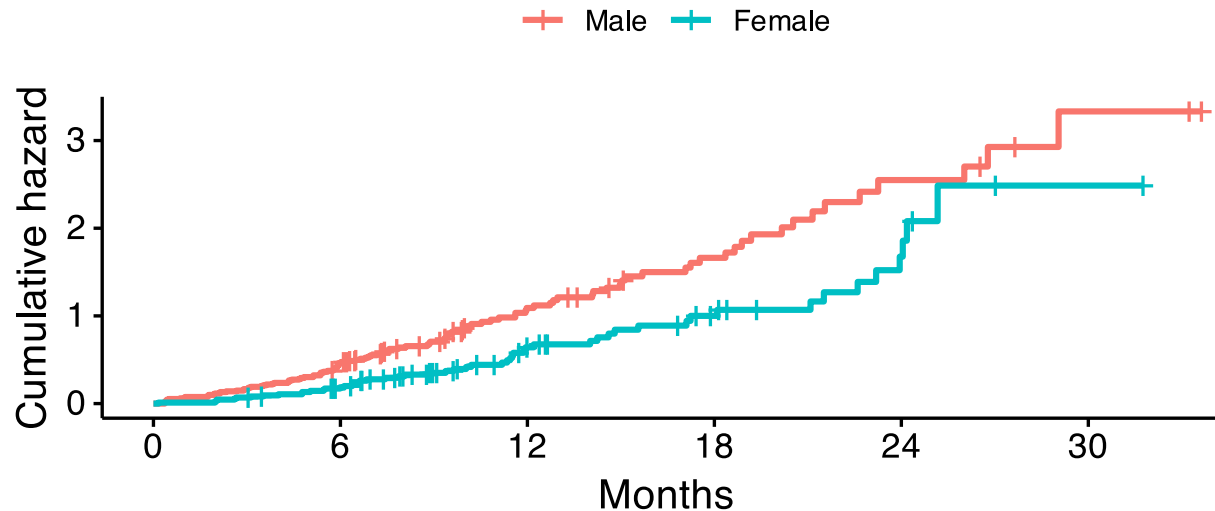
- The quantity of interest from a Cox regression model is a **hazard ratio (HR)**. The HR represents the ratio of hazards between two groups at any particular point in time.
- The HR is interpreted as the instantaneous rate of occurrence of the event of interest in those who are still at risk for the event. It is **not** a risk, though it is commonly interpreted as such.
- If you have a regression parameter  $\beta$  (from column **estimate** in our **coxph**) then  $HR = \exp(\beta)$ .
- A  $HR < 1$  indicates reduced hazard of death whereas a  $HR > 1$  indicates an increased hazard of death.

# Hazard ratios

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

```
fit4 <- survfit(Surv(time, status) ~ sex, data = lung)

ggsurvplot(data = lung, fit = fit4, xlab = "Months", xscale = 30.4,
            break.x.by = 182.4, fun = "cumhaz", legend.title = "",
            legend.labs = c("Male", "Female"),
            risk.table = TRUE, risk.table.y.text = FALSE)
```



Number at risk



# Hazard ratios

Are there potential issues with the hazard ratio?

Yes, the average HR ignores the distribution of events during the follow-up.

But a possible solutions of time specific HRs poses another problem

- selection bias due to depletion of susceptibles.

Possible solutions

- i) report survival curves
- ii) accelerated survival models.

# What are competing risks?

When subjects have multiple possible events in a time-to-event setting (e.g recurrence, death from disease, death from other causes).

So what's the problem? Why not just use KM approach and treat competing events as censored events?

Remember basic KM assumption - censored patients have same risk as those remaining under observation.

Unobserved dependence among event times is the fundamental problem that leads to the need for special consideration.

Two approaches to analysis in the presence of competing risks.

1. **Cause-specific hazards** instantaneous rate of occurrence of the given type of event in subjects who are currently event-free estimated using Cox regression (coxph function)
2. **Subdistribution hazards** instantaneous rate of occurrence of the given type of event in subjects who have not yet experienced an event of that type estimated using Fine-Gray regression (`crr()` in `cmprsk` package)

# Hazard of the Hazards

**QUESTIONS?**

**COMMENTS?**

**RECOMMENDATIONS?**