Introduction to Time-to-Event

Mabel Carabali

Risk & Hazards

EBOH, McGill University

Updated: 2025-09-13

Doctor: "Sir, the
surgery has a 50% survival rate,
but don't worry, my last
20 patients have all survived"

What to do, if you want to:

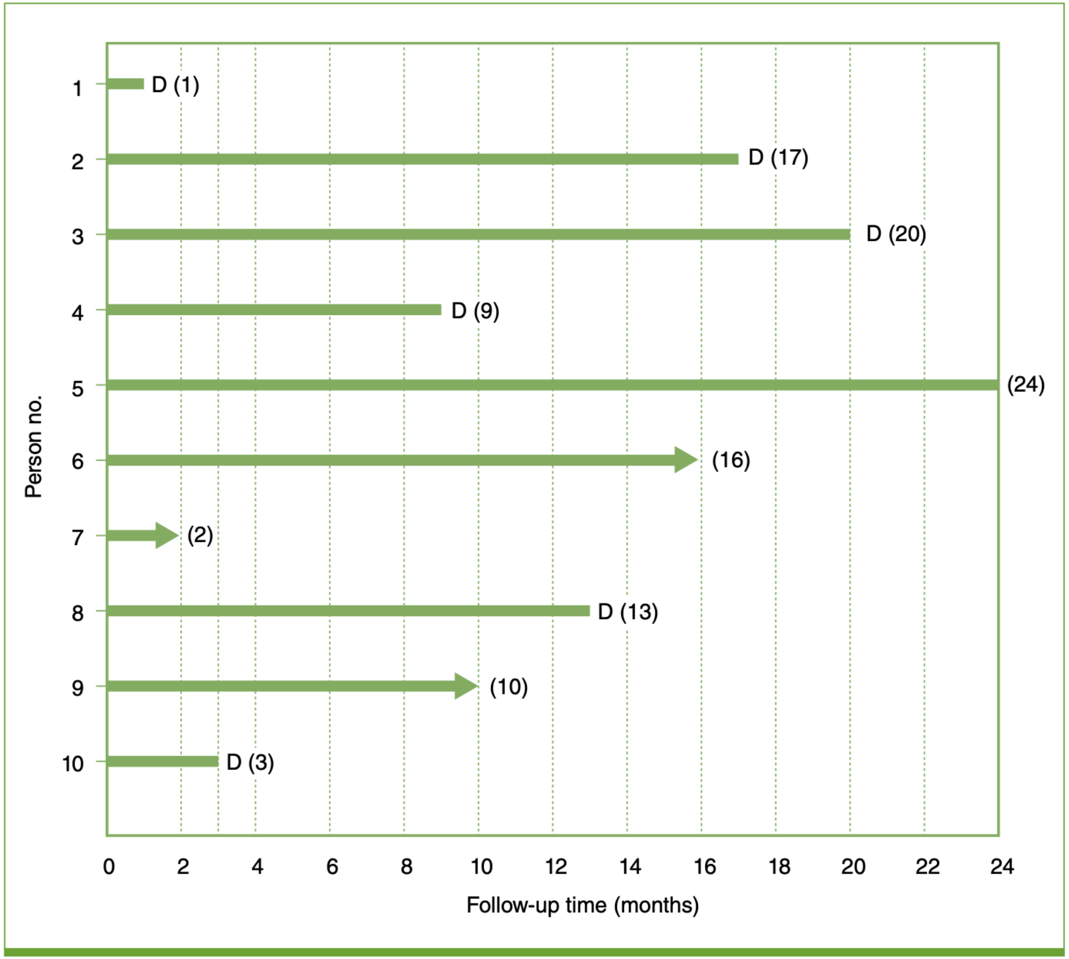
- Estimate median survival times, plot survival over time after treatment, or estimate the probability of surviving beyond a prespecified time interval (eg, 5-year survival rate)?
- Assess whether survival times are related to covariates and/or adjust for potential confounders.
- Account for censoring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



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

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.

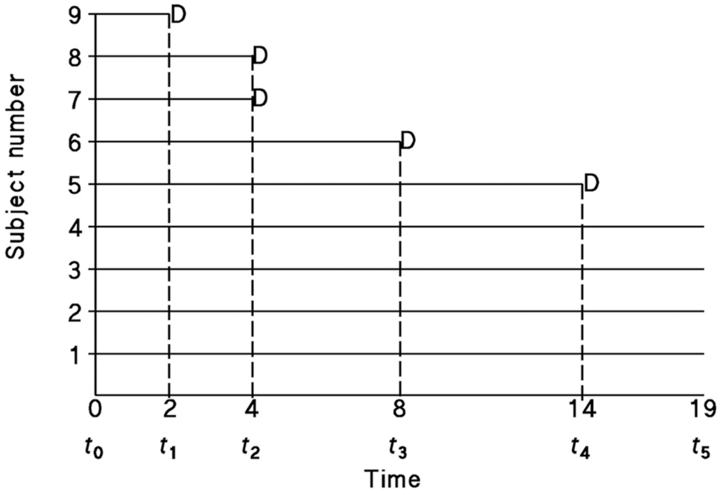
- At time 0, the survival probability is 1, i.e. $S(t_0) = 1$

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)

$$Incidence\ Proportion = 1 - S$$

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

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 (ΔA_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
% Surviving (S_k)		8/9	6/8	5/6	4/5	4/4
Interval length ($N_k \Delta A_k$)		2	2	4	6	5
Person-time (ΔA_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

$Incidence\ Proportion = 1 - S$ = 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" $\longrightarrow$ 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 $(1 - S_i) \rightarrow 1 - S_{24} = 1 - 0.18 = 0.82$

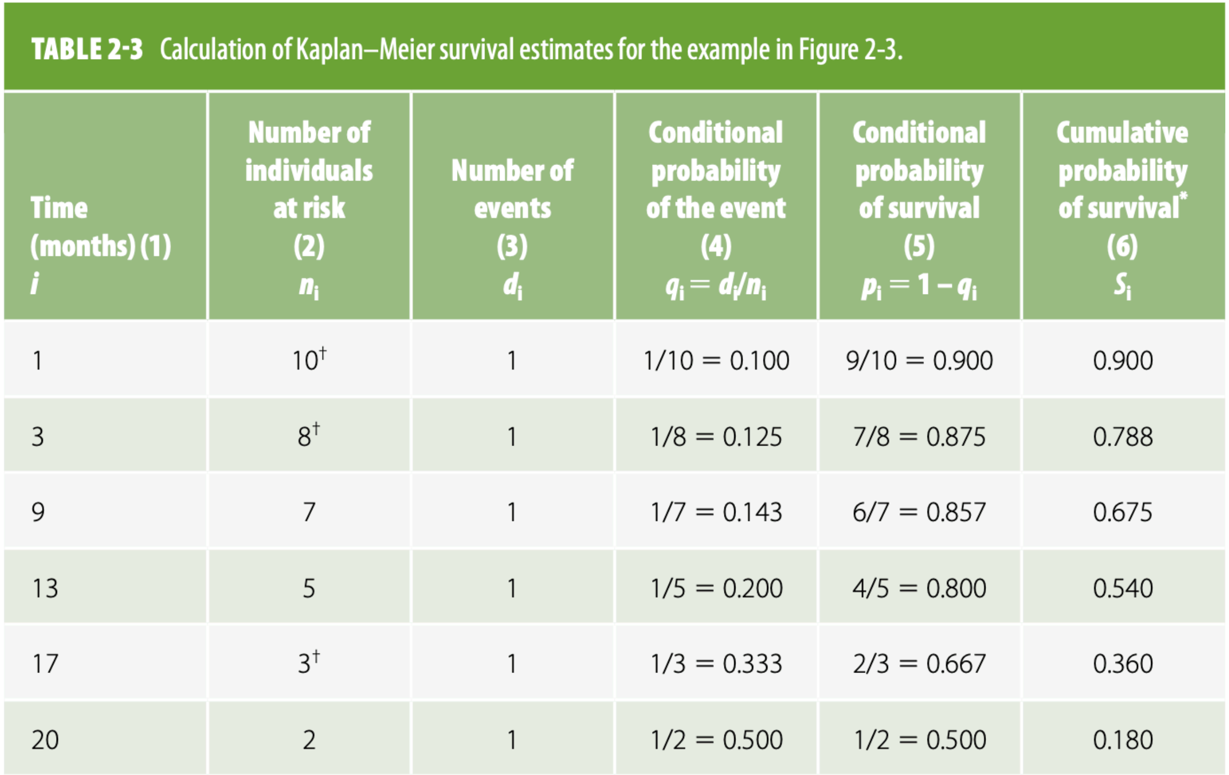


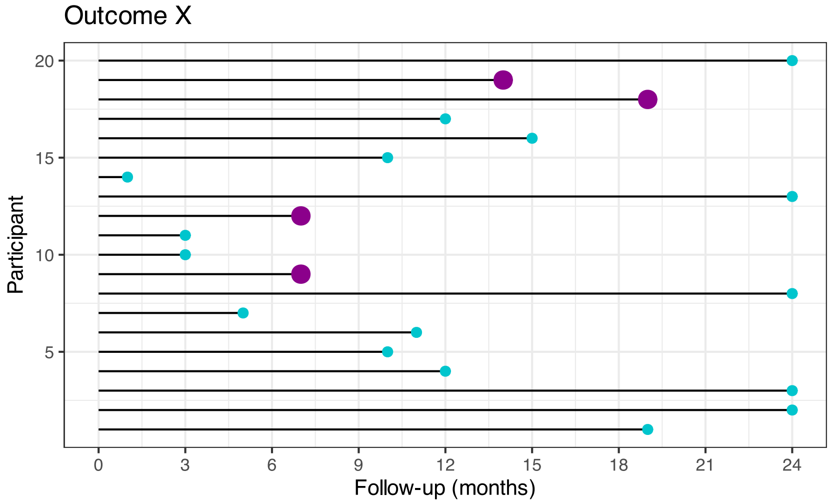
TABLE 2-3 Calculation of Kaplan–Meier survival estimates for the example in Figure 2-3.

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†	1	1/10 = 0.100	9/10 = 0.900	0.900
3	8†	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†	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

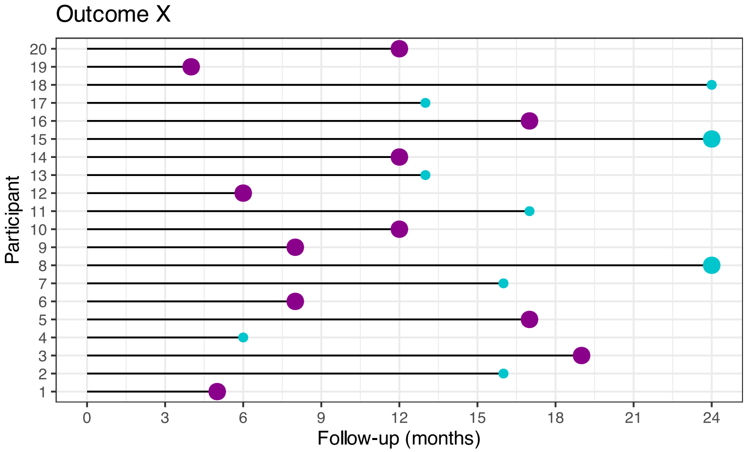
```
surv_object<-Surv(time = dat$t,
                  event = dat$O2)

eventKM<- survfit(surv_object ~ 1,
                  data = dat,
                  type="kaplan-meier")
```

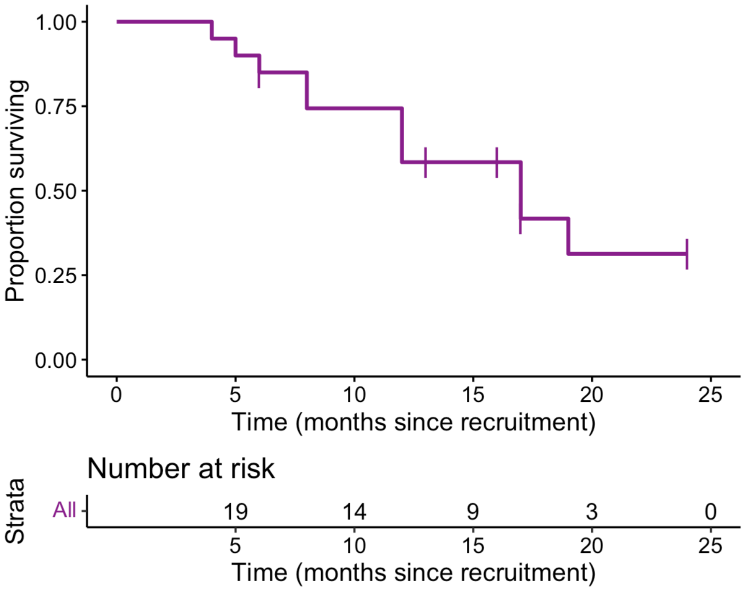
Output

```
## 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
##     7     16       2    0.875  0.0827        0.727            1
##    14      9       1    0.778  0.1175        0.578            1
##    19      7       1    0.667  0.1440        0.437            1
```

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:

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

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 = $S_k = 1 - IR_k \Delta_{tk}$

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

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
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

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 = $1 - S$ = **1 - 0.444**

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

$Risk_c =$ **0.556** $\cong Risk_{ExpForm}$ = **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 `myeloma` 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. 2 year) use `summary()`.
- Produce survival curve including this information

Get a Table 1

```
myeloma<- survival::myeloma

tab1<- myeloma %>%
  tbl_summary()
```

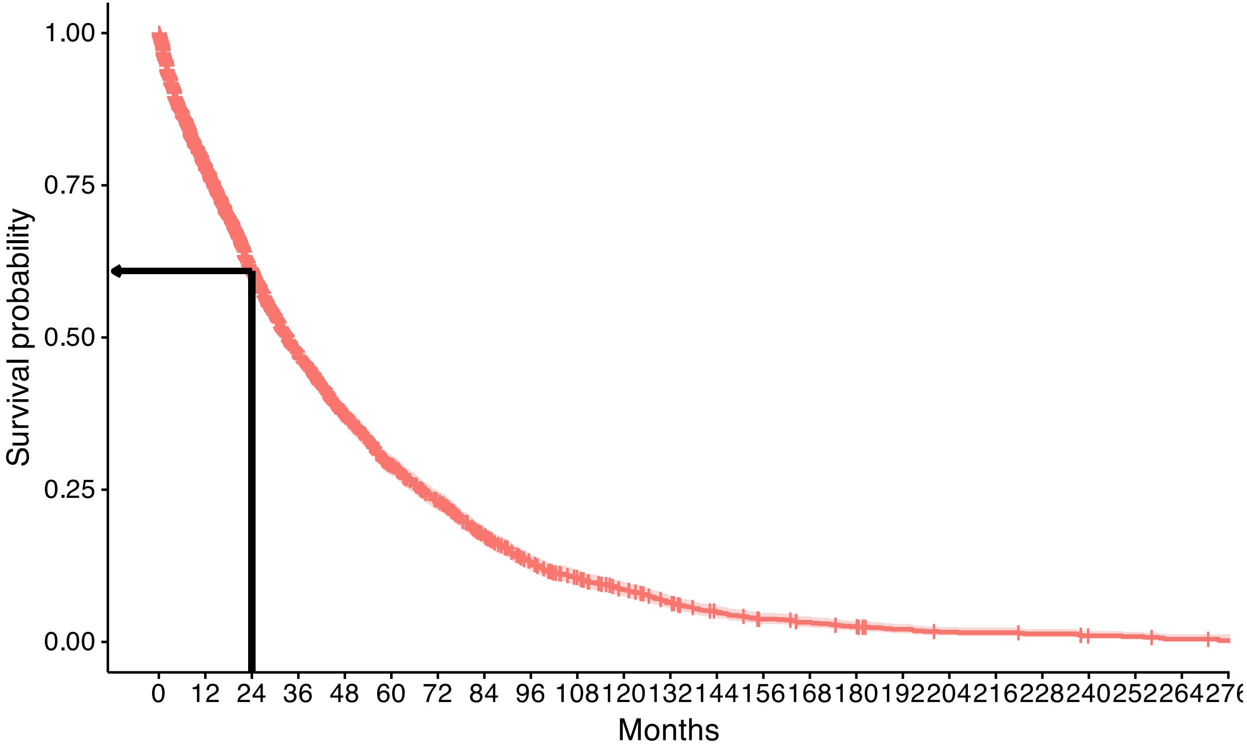
Check the package `gtsummary` for more information and details on formatting

Characteristic	N = 3,882[1]
id	1,948 (972, 2,925)
year	82 (72, 89)
entry	0 (0, 109)
futime	693 (243, 1,461)
death	2,769 (71%)

[1] Median (Q1, Q3); n (%)

Worked example: Survival at 24 months (2 Years)

```
f1 <- survfit(Surv(futime, death) ~ 1, data = myeloma)
```



Summary Results

```r
summary(survfit(Surv(futime, death) ~ 1,  data = myeloma), times = 730.5)
```

```
## Call: survfit(formula = Surv(futime, death) ~ 1, data = myeloma)
##
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##   730   1878    1340    0.609 0.00843        0.593        0.626
```

What happens if you use a naive estimate ?

1340 of the 3882 patients died by 2^{nd} year so:

$$\left(1 - \frac{1340}{3882}\right) \times 100 = 65.5\%$$

You get an **incorrect** estimate of the 2-year probability of survival when you ignore the fact that 664 patients were censored before 2 year.

What happens if you use a naive estimate ?

- Recall the **correct** estimate of the 2-year probability of survival was **61%.** which is $\neq$ 65%
- 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 `myeloma` data

Comparing survival times between groups

```r
set.seed(7042025) #0= males; 1= females
myeloma$sex<- ifelse(myeloma$death==1, rbinom(length(myeloma$id), 1, 0.3), 0)
survdiff(Surv(futime, death) ~ sex, data = myeloma)
```

```
## Call:
## survdiff(formula = Surv(futime, death) ~ sex, data = myeloma)
##
##          N Observed Expected (O-E)^2/E (O-E)^2/V
## sex=0 3070     1957     2141      15.8      69.9
## sex=1  812      812      628      53.8      69.9
##
##  Chisq= 69.9  on 1 degrees of freedom, p= <2e-16
```

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(futime, death) ~ sex, data = myeloma)
```

```
## Call:
## coxph(formula = Surv(futime, death) ~ sex, data = myeloma)
##
##       coef exp(coef) se(coef)     z      p
## sex 0.3477    1.4159   0.0418 8.318 <2e-16
##
## Likelihood ratio test=65.75  on 1 df, p=5.119e-16
## n= 3882, number of events= 2769
```

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(futime, death) ~ sex, data = myeloma), exp = TRUE) %>%
  kable()
```

term	estimate	std.error	statistic	p.value
sex	1.415857	0.0418027	8.318476	0

Or use `tbl_regression` from the `gtsummary` package

```
coxph(Surv(futime, death) ~ sex, data = myeloma) %>%
  gtsummary::tbl_regression(exp = TRUE)
```

Characteristic	HR[1]	95% CI[1]	p-value
sex	1.42	1.30, 1.54	<0.001

[1] 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 β (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

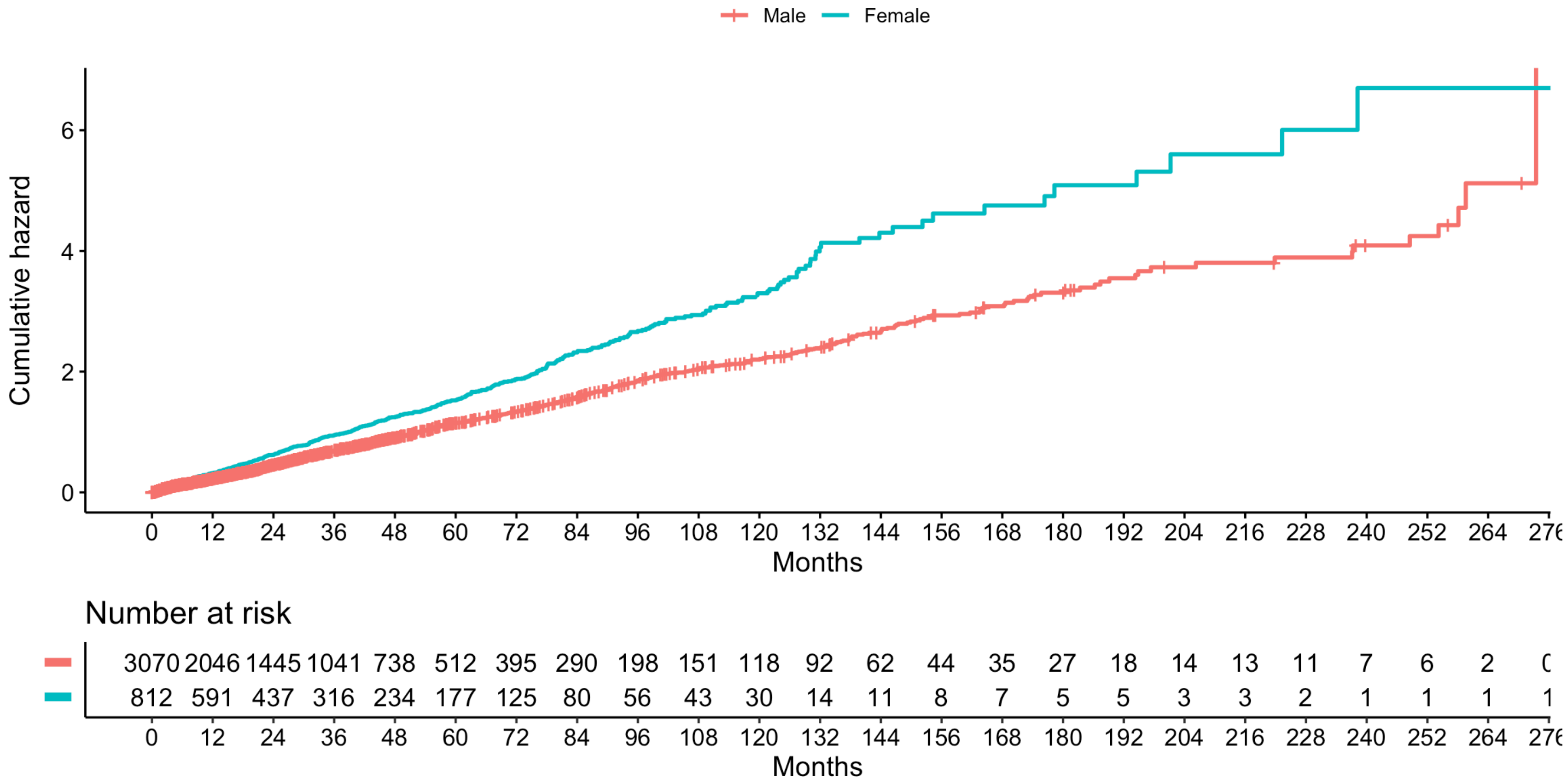
So our HR = 1.42 implies that around 1.42 times as many females are dying as males, at any given time.

```
fit4 <- survfit(Surv(futime, death) ~ sex, data = myeloma)

fitplot<- ggsurvplot(data = myeloma, fit = fit4, xlab = "Months", xscale = 30.4,
                     break.x.by = 364.8,   fun = "cumhaz",
                     legend.title = "", legend.labs = c("Male", "Female"),
           risk.table = TRUE, risk.table.y.text = FALSE)
```

Hazard ratios by Sex (assigned at birth):

From the Myeloma example



Number at risk

	0	12	24	36	48	60	72	84	96	108	120	132	144	156	168	180	192	204	216	228	240	252	264	276
🟥	3070	2046	1445	1041	738	512	395	290	198	151	118	92	62	44	35	27	18	14	13	11	7	6	2	0
🟦	812	591	437	316	234	177	125	80	56	43	30	14	11	8	7	5	5	3	3	2	1	1	1	1

Months

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.

Hernán,The Hazards of Hazard Ratios

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

1. Non-Collapsibility and Selection Bias
   - Cannot be directly interpreted as a population-level comparison of risk.
   - Built-in selection bias, even in randomized controlled trials.
2. Proportional Hazards Assumption Violation
3. Sensitivity to Follow-Up Time and Number of Events
   - use `survival::lung` and the `survival::colon` to work the example
4. Misleading Interpretation as Relative Risk and
5. Lack of absolute risk information

QUESTIONS?

COMMENTS?

RECOMMENDATIONS?