not only valid for time-to-death data
but for all types of time-to-event data.

Basics of Survival Analysis (with R)

CRUK Bioinformatics Summer School 2019

Dominique-Laurent Couturier & Serigne Lo

Outline of presentation

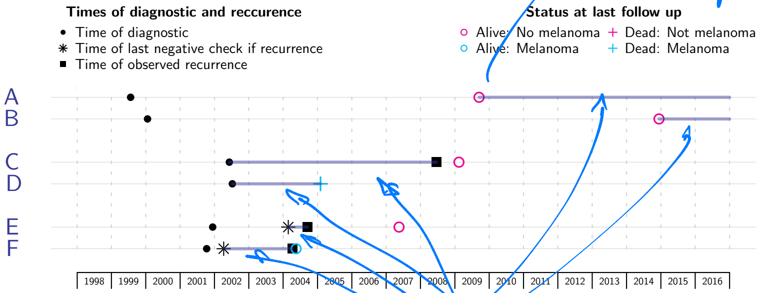
- Properties of time-to-event data,
- Kaplan-Meier curves,
- Log-rank test
- Cox proportional hazard regression models
  and interpretation of hazards ratios

Example of time-to-event data:

Time to relapse for melanoma patients

loss of follow-up, drop-out

~10y
~14y

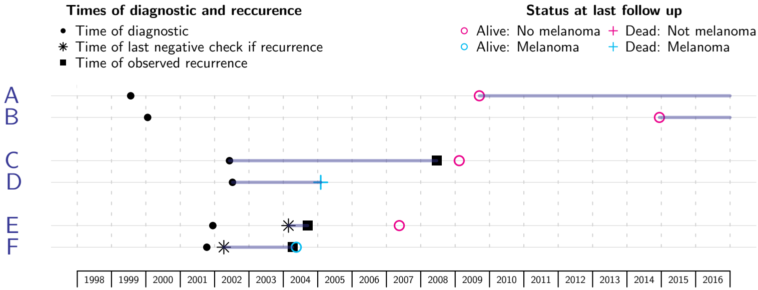


~2-5y

different follow-up time

time-to-event often not precisely observed.

Example of time-to-event data:
Time to relapse for melanoma patients



3 types of censoring:

- Right censoring: event did not occur before time of last follow-up,
- Left censoring: event occurred before a certain time,
- Interval censoring: event occurred during a specific interval of time.

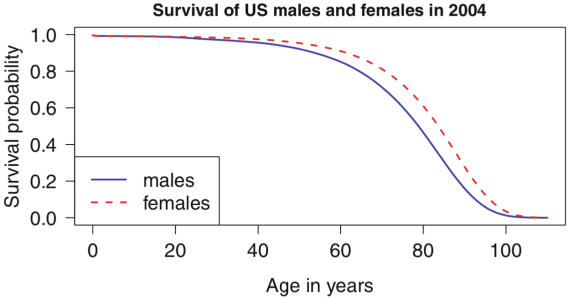
Quantities/Functions of interest in survival analyses

Let T denote the time-to-event of interest with

- $0 < T < \infty$,
- $T \sim f_T(t)$. Exponential, Weibull, Gamma distributions

Then, the survival function, $S(t)$, is defined as the probability of surviving until point t, i.e.,

$$S(t) = \text{Prob}(T > t)$$



Source: Moore (2006)

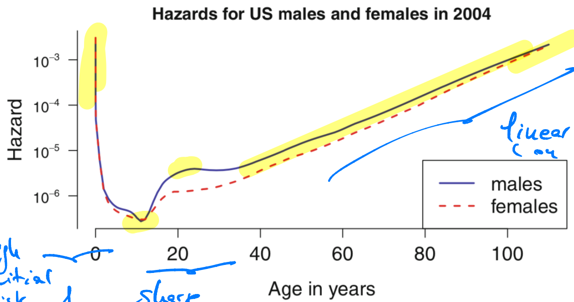
Quantities/Functions of interest in survival analyses

Let T denote the time-to-event of interest with

- $0 < T < \infty$,
- $T \sim f_T(t)$.

Then, the hazard function, $h(t)$, is defined as the probability of instantaneous death at time t, i.e.,

$$h(t) = \frac{f_T(t)}{S(t)}$$



linear increase (on the log-scale)

high initial risk of death

sharp increase of risk, especially for men

Source: Moore (2016)

Quantities/Functions of interest in survival analyses

Let T denote the time-to-event of interest with

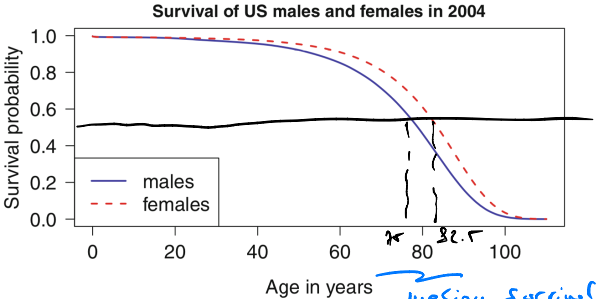
- $0 < T < \infty$,
- $T \sim f_T(t)$.

Then, the mean and median survival times, defined as the

$$\cancel{\mu = E[T]}$$

not always a suitable measure due to infinity values

$$\text{median}(T) = \text{time } t \text{ such that } S(t) = 0.5$$



median survival times

Source: Moore (2016)

Kaplan-Meier non-parametric survival function estimator

Product over the failure times of the conditional probabilities of surviving to the next failure time. Formally:

$$\widehat{S}(t) = \prod_{i:\ t_i \le t} (1 - \widehat{q}_i) = \prod_{i:\ t_i \le t} \left(1 - \frac{d_i}{n_i}\right)$$

where

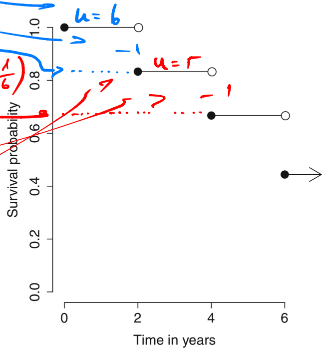
- d_i denotes the number of people who failed at that time,
- n_i denotes the number of subject at risk at time t_i.

t_i	n_i	d_i	q_i	$1 - q_i$	$S_i = \prod(1 - q_i)$
2	6	1	0.167	0.833	~~0.846~~
4	5	1	0.200	0.800	~~0.693~~
6	3	1	0.333	0.667	0.497

0.83

$0.665 = (1 - \frac{1}{5})(1 - \frac{1}{6})$

$\frac{1}{6}$ $\quad 1 - \frac{1}{6}$



Log-rank test to compare survival curves

Hypotheses of interest:

- **H0**: $S_1(t) = S_2(t)$ for all time points t,
- **H1**: $S_1(t) \neq S_2(t)$ for some time t.

	Control	Treatment	
Failure	d_{0i}	d_{1i}	d_i
Non-failures	$n_{0i} - d_{0i}$	$n_{1i} - d_{1i}$	$n_i - d_i$
	n_{0i}	n_{1i}	n_i

"It is constructed by computing the observed and expected number of events in one of the groups at each observed event time and then adding these to obtain an overall summary" (Wikipedia).

- Non-parametric test,
- Assumes that censoring is non-informative.

Cox proportional hazard model

$$h_i(t|\mathbf{x}_i) = e^{\mathbf{x}_i^\top \boldsymbol{\beta}}\ h_0(t),$$

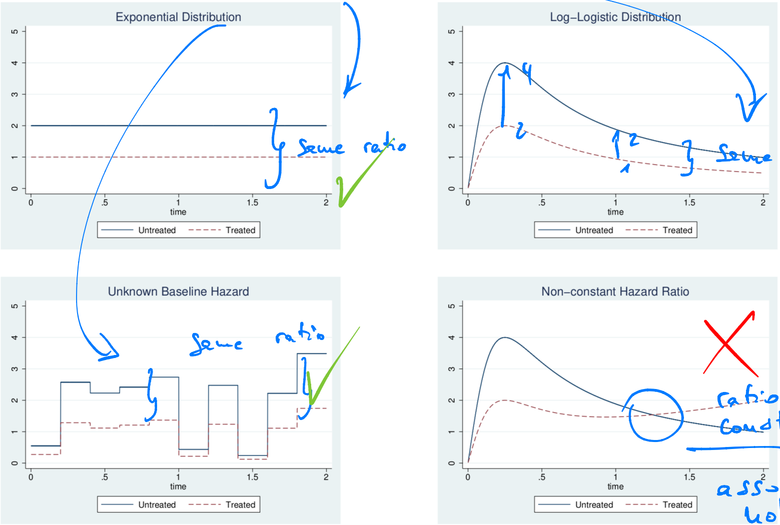
$$\frac{h_i(t|x_i)}{h_0(t)} = e^{x_i^\top \beta}$$

where:

- $h_i(t|\mathbf{x}_i)$ is the *hazard at time t for the ith individual* with covariates $\mathbf{x}_i = [x_1, ...x_p]^\top$,
- $h_0(t)$ is the *baseline hazard* at time t, i.e., the instantaneous probability of event for participants with $\mathbf{x}_i^\top \boldsymbol{\beta} = 0$,
- The ratio $\frac{h_i(t|\mathbf{x}_i)}{h_0(t)} = e^{\mathbf{x}_i^\top \boldsymbol{\beta}}$ is called the *hazard ratio*. It quantifies how much more individual with covariates $X_i = x_1, ...x_p$ is likely to experience the event of interest (death) as compared to the "baseline" individual.

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Cox model: Proportional hazard assumption



Source: Lunt (2018)

Same ratio

Same ratio

Same ratio

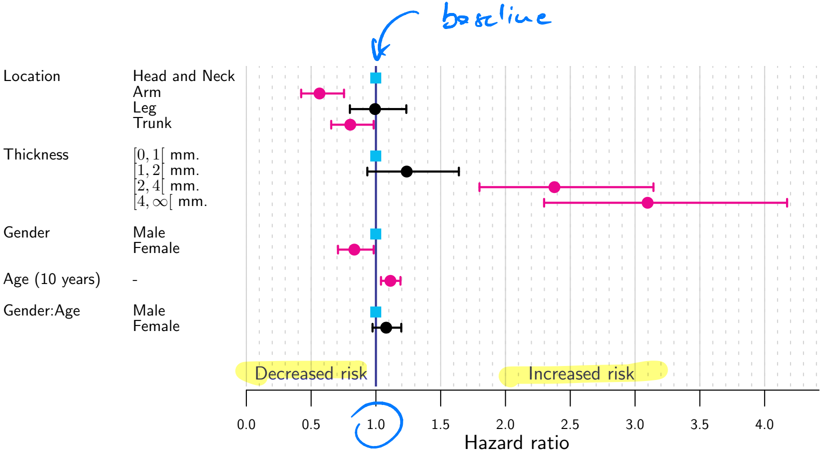
ratio not constant.

assumption not respected.

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Modelling of time of first recurrence of melanoma - MIA: $e^{\widehat{\beta}}$

- 1952 patients observed between 1998 and 2016, observed recurrence for 37% of the patients, 13% left censoring, 24% interval censoring, 63% right censoring
- $h_0(t)$ corresponds to instantaneous risk to have a melanoma recurrence at time t for men of average age with a diagnosed melanoma of small size (<1mm) located on the head/neck



CANCER RESEARCH UK | CAMBRIDGE INSTITUTE