3.2 Survival, hazard, Cox regression

Time-to-event data

- Survival analysis concerns analysing the time to the occurrence of an event, e.g. time until a patient dies.

- Such analysis is used for cohort studies and randomized clinical trials (RCTs), where study participants are followed from a start time to an endpoint.

- The outcome has two components
  - the time the individual was followed for
  - an event indicator to distinguish between events (usually coded 1) and non non-events (“censorings”) coded 0.

Examples of time-to-event data

- time (years) from diagnosis of cancer to death
- time (months) from delivery to next pregnancy
- time (weeks) from birth to infant vaccination.
- time (days) from admission to discharge of hospital patients

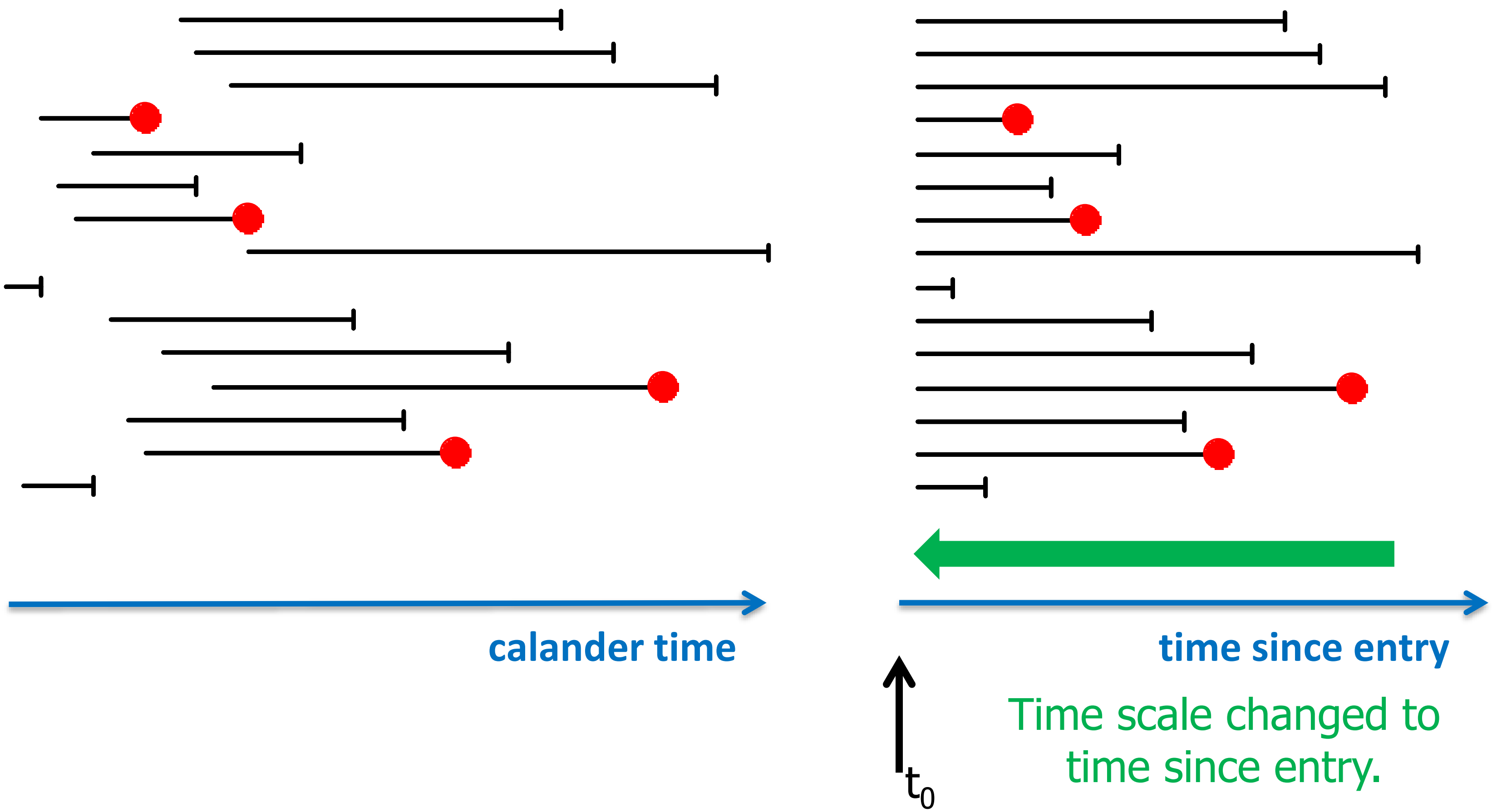
How to describe the pattern of the incidence rate over time

- We have seen how “time-to-event” information can be used to calculate the incidence rate over the follow-up period.
- The events (e.g. deaths) that we observe are only among those individuals still being followed.
- Need to take time-at-risk (follow-up time) into account if we wish to describe the risk at specific time points and not just an overall incidence:

This is what **survival analysis** achieves.

Visualizing individual survival data (open cohort)

Each line a person ● Event | Censored

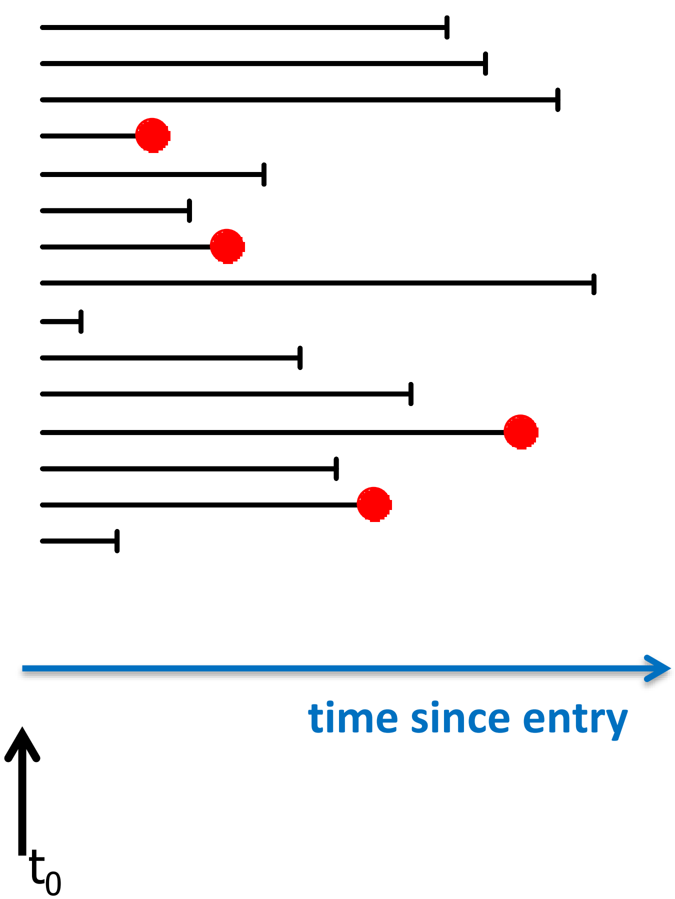


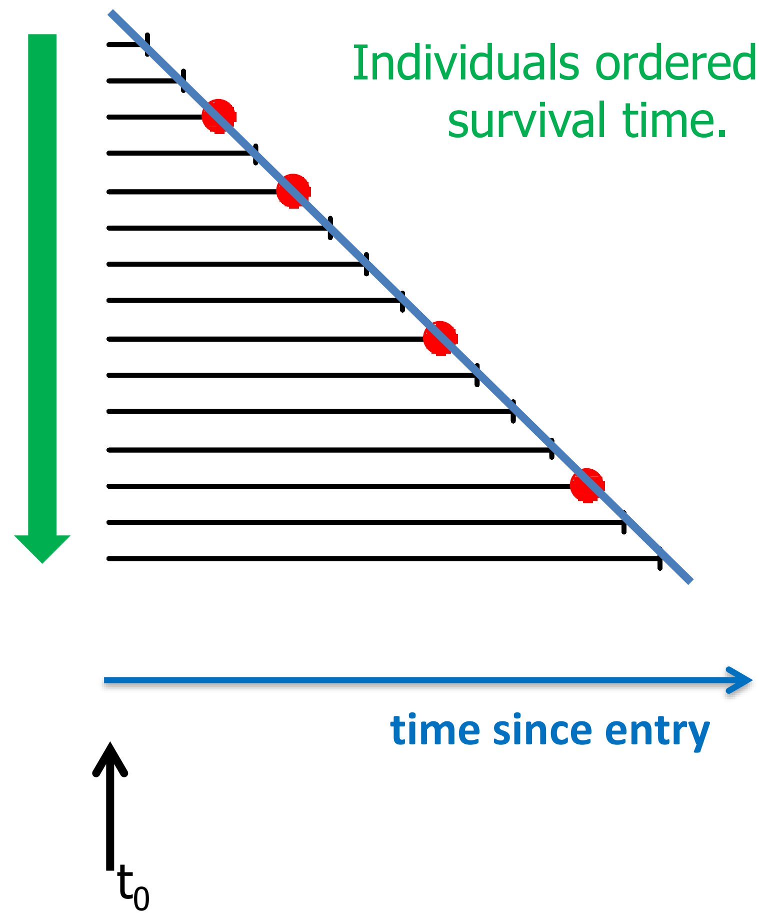
Visualizing individual survival data

Each line a person

Event

Censored

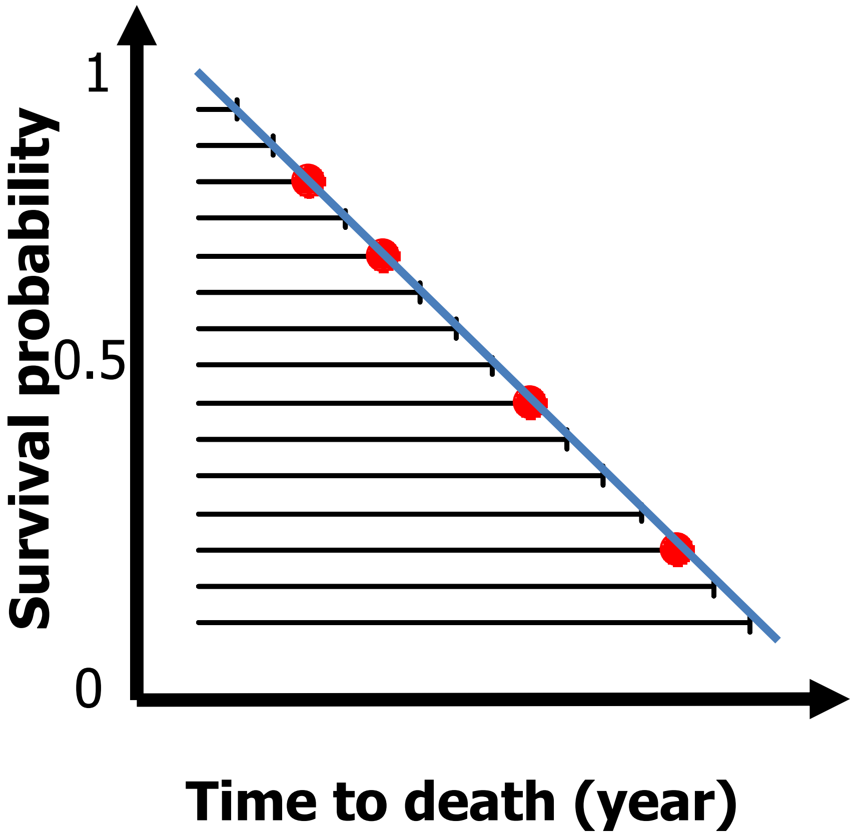




Survival function, $S(t)$

Describes the probability of ”surviving” to time t, $S(t)$.

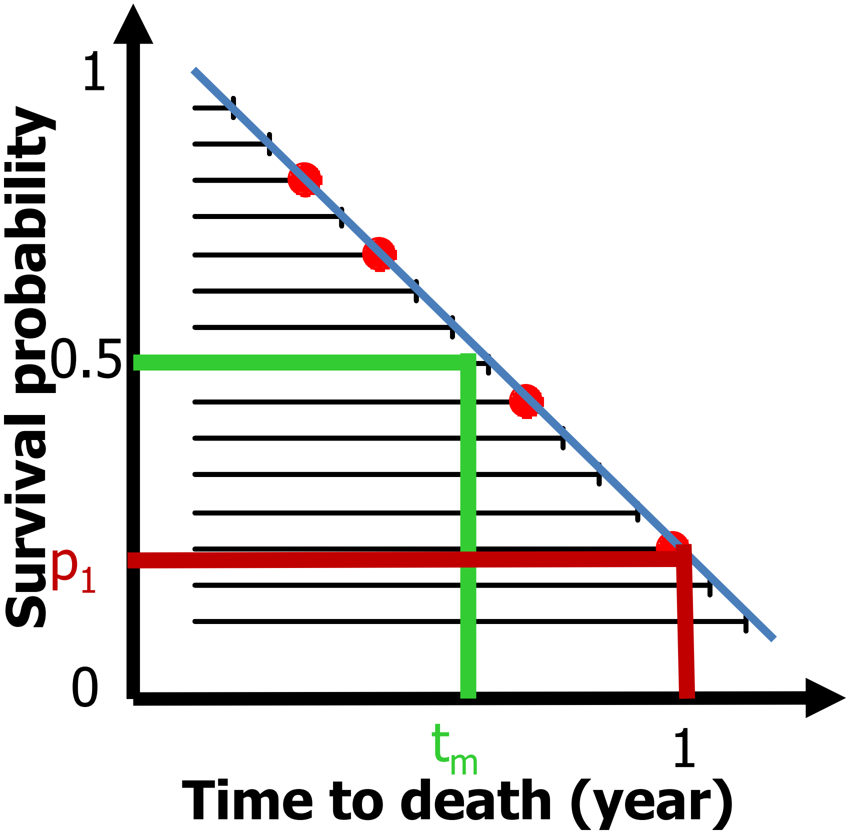
- Properties:
  - Value between 0 and 1.
  - All (100%) “alive” at start.
  - Decreasing over time



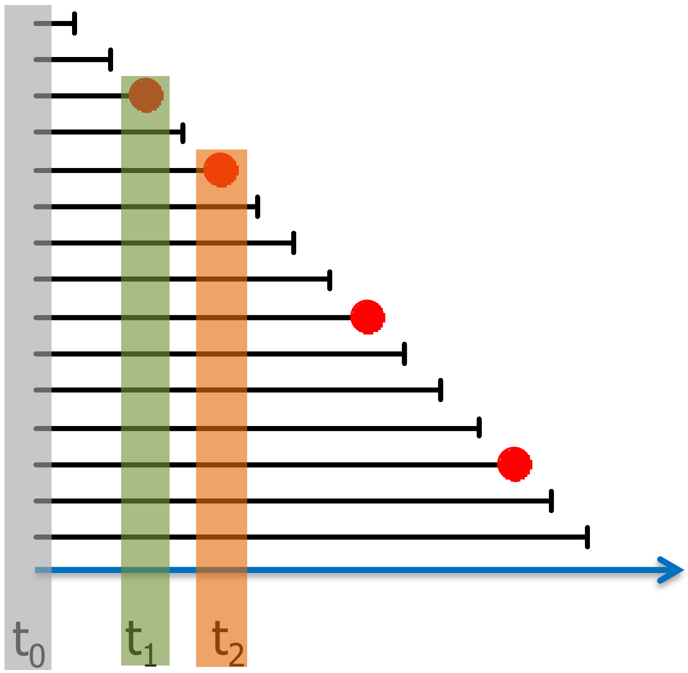
Survival function, S(t)

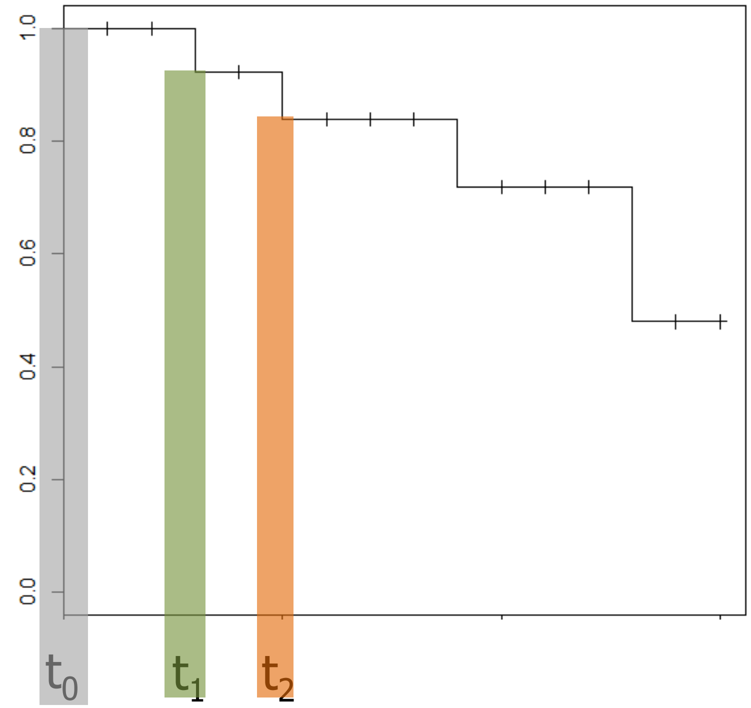
- S(t) contains all information about the survival probability changes *over time*.
- Provides estimates of:
  - Median survival time (t_m).
  - 1 year survival probability (p_1).

We estimate of S(t) using a "Kaplan-Meier" curve ....



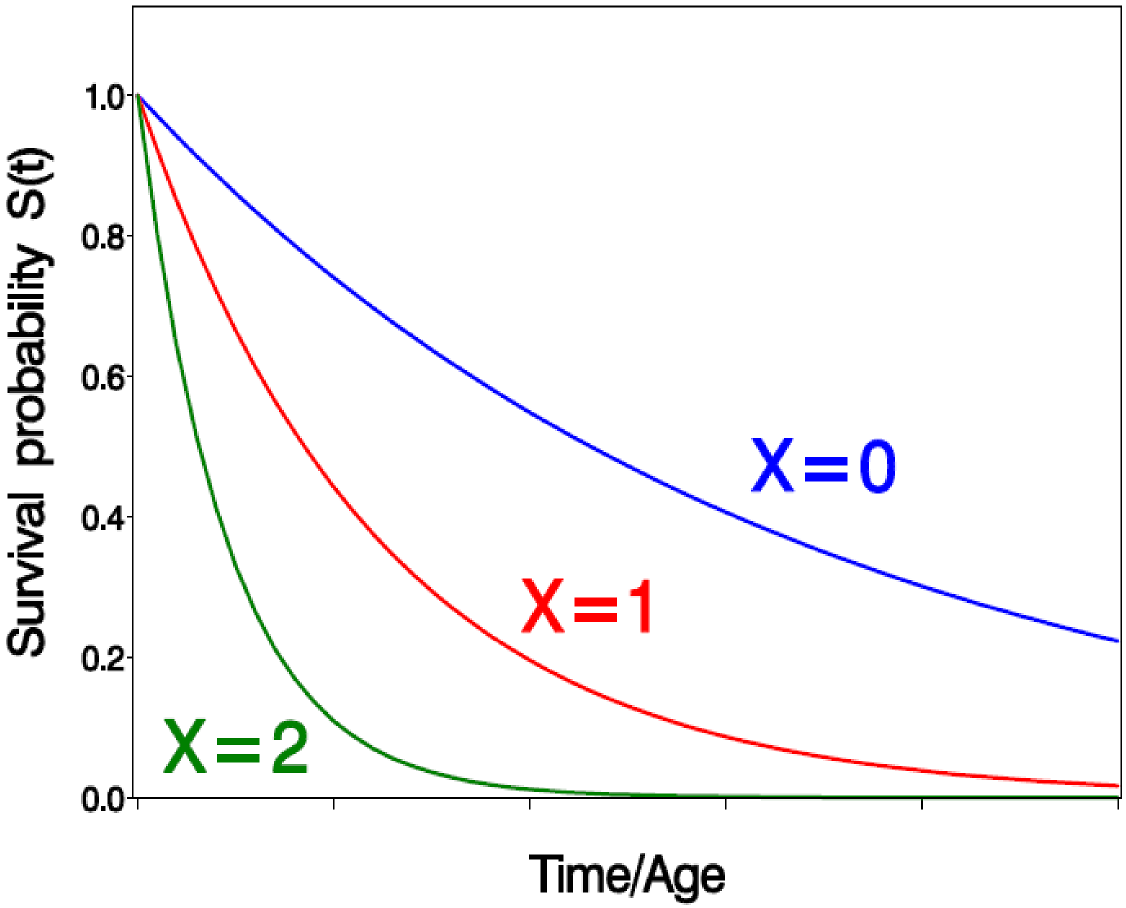
Kaplan-Meier curve



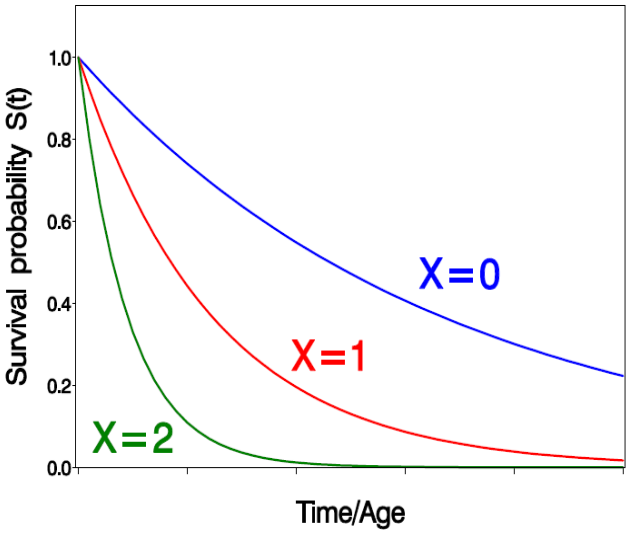


Time:	t_0	t_1	t_2	t_i
At risk:	$n_0 = 15$	$n_1 = 13$	$n_2 = 11$	n_i
Probability of surviving at t:	$P_0=1-d_0/n_0$	$P_1=1-d_1/n_1$	$P_2=1-d_2/n_2$	$P_i=1-d_i/n_i$
Estimate of S(t):	1	$P_1= 12/13$	$P_1P_2 = (12/13)(10/11)$	$P_1P_2\ldots P_{i-1}P_i$

We usually want to assess how survival depends on an exposure X.



Comparing survival curves



- Individuals with X=0 have better survival compared to those with X=1 or X=2
- Survival (Kaplan-Meier) curves are compared formally using the log-rank test

Often, we want to study how survival depends on exposure and confounders, as we did for binary outcomes (using logistic models)

So we need to model the survival

Cox regression model

Usual regression model for survival data is the

Cox proportional hazards model which:

- models the *hazard*, *h(t)*, i.e. the instantaneous rate (events per unit time) at time t.
- assumes the hazard for an individual with exposure X is:

$$h(t|X) = h_0(t)\exp^{\beta X} \quad \text{i.e. } \ln\{h(t)\} = \ln\{h_o(t)\} + \beta X$$

where $h_0(t)$ is the "baseline" hazard (if $X = 0$)

$$\frac{h(t|X)}{h_0(t)} = \exp^{\beta X}$$ is the hazard ratio, HR

Note the similarity to the logistic model and the OR

Compare models

Models	Linear Predictors	Measure of Associations
Linear Regression	$Y[X] = \alpha + \beta X$	Slopes
Logistic Regression	$\ln(P[Y=1\|X]/P[Y=0\|X]) = \alpha + \beta X$	Odds ratios
Cox Regression	$\ln\{h(t\|X)\} = \ln\{h_0(t)\} + \beta X$	Hazard ratios

Hazard and survival functions

Mathematical connection between $\mathrm{h}(t|X)$ and $S(t|X)$:

$$\mathrm{h}(t|X) = h_0(t)exp^{\beta X}$$

equivalent to

$$S(t|X) = [S_0(t)]^{exp^{\beta X}}$$

Large hazard implies a *rapid rate of decline* in survival $S(t|X)$

Hazard and survival functions

$$\boxed{S(t|X) = [S_0(t)]^{exp^{\beta X}}}$$

- In case $\beta > 0$:

$$X \nearrow \quad \Longrightarrow \quad exp^{\beta X} \nearrow \quad \Longrightarrow \quad S(t|X) < S_0(t)$$

Higher X-values associated with increased risk for event

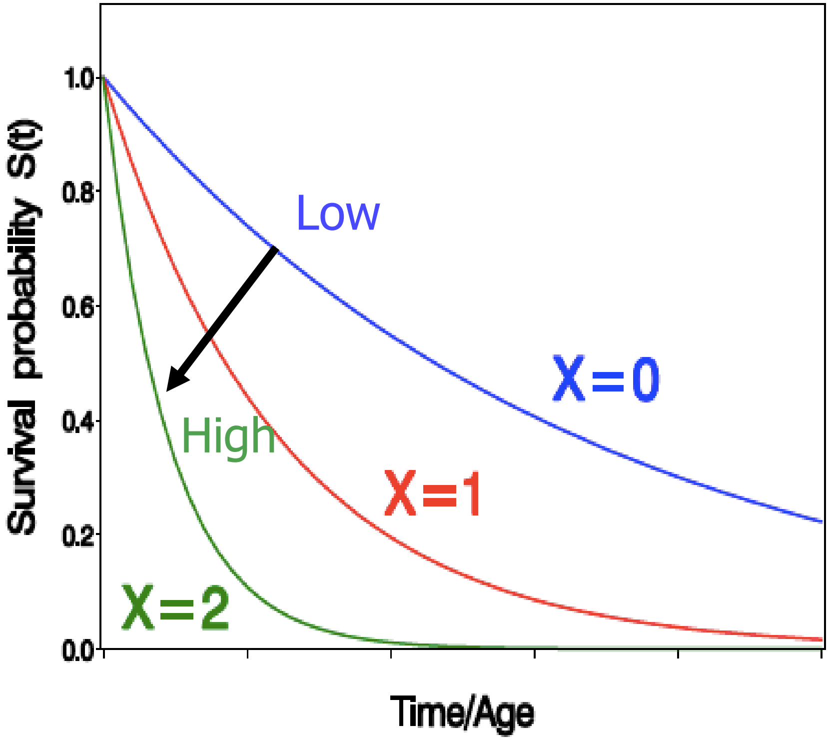
- In case $\beta < 0$:

$$X \nearrow \quad \Longrightarrow \quad exp^{\beta X} \searrow \quad \Longrightarrow \quad S(t|X) < S_0(t)$$

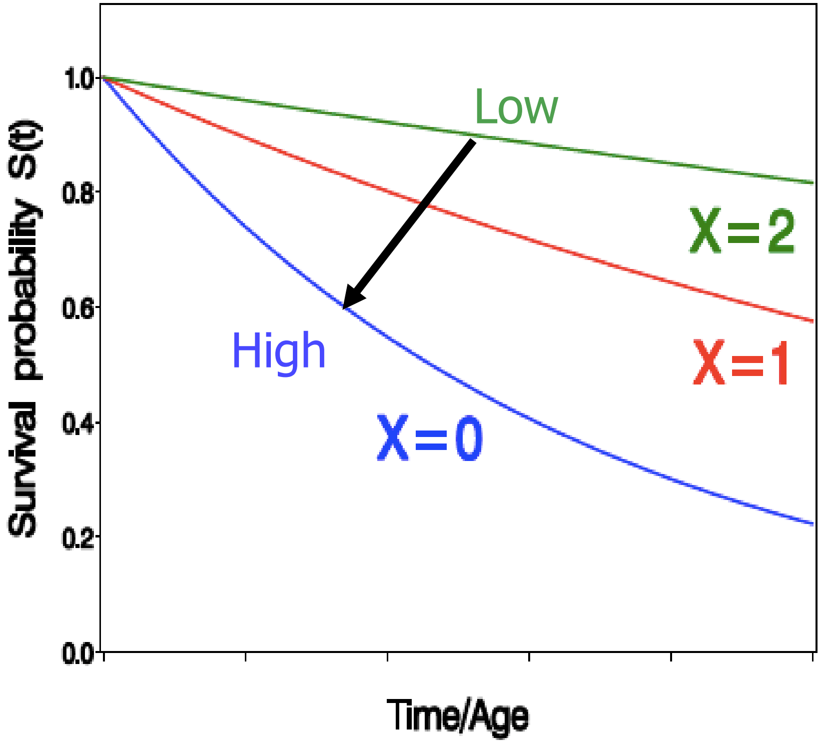
Higher X-values associated with reduced risk for event

Hazard and survival functions

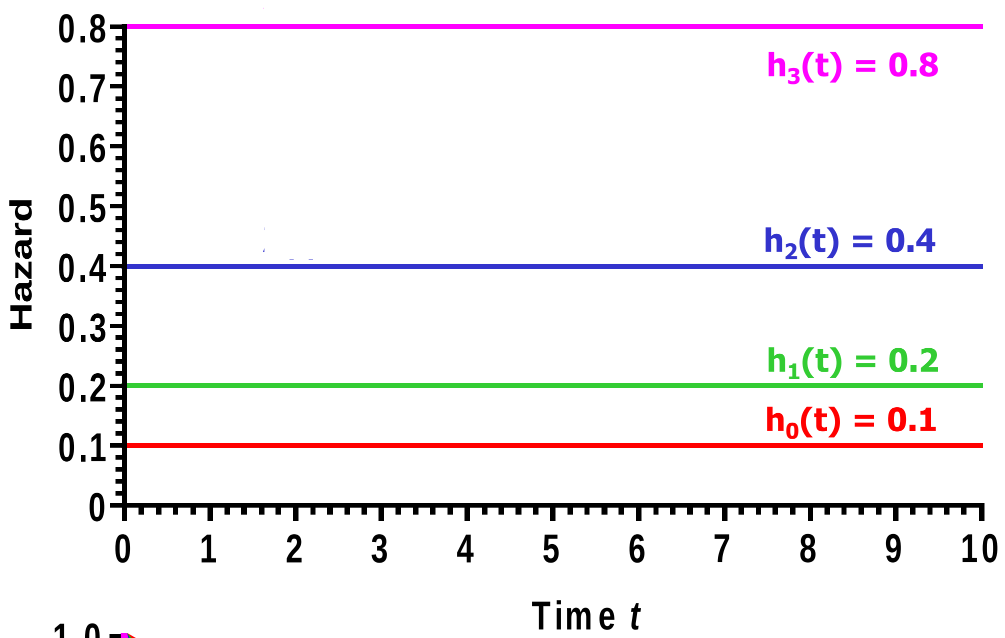
$\beta > 0$



$\beta < 0$



Example of 4 groups, each with constant hazard



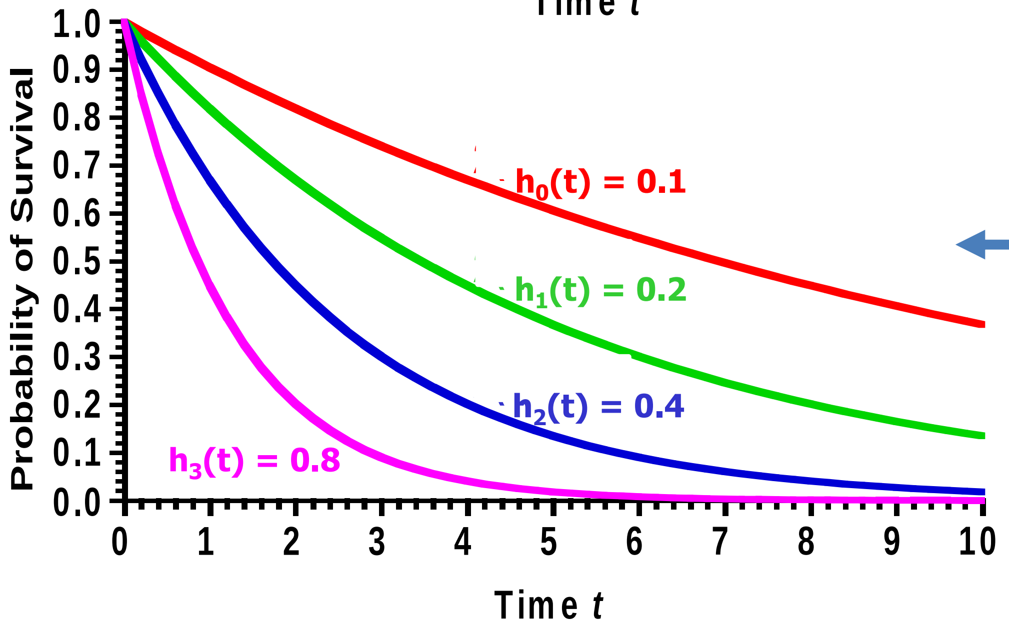
Using red as reference or "baseline hazard":

$h_0(t) = 0.1$

$HR_{1vs0} = h_1(t)/h_0(t) = 2$

$HR_{2vs0} = h_2(t)/h_0(t) = 4$

$HR_{3vs0} = h_3(t)/h_0(t) = 8$



Survival curves look like this

Proportional hazards (PH) assumption

- Means the ratio of the hazards for the two groups is constant over time, $\exp^{\beta}$ does not depend on time.

- Places no restrictions on the shape of the baseline hazard, $h_0(t)$, but requires $h(t|X)/h_0(t) = \exp^{\beta X}$.

- In previous example, the 4 hazards were constants.

Cox regression model

Finds the β that gives best fit of the hazard $h(t|X) = h_0(t)\exp(\beta X)$ to the data

or equivalently,

exp^{β} =HR

$$\ln\{h(t|X)\} = \ln\{h_0(t)\} + \beta X$$

Note similarity to logistic regression where we find β that gives best fit of the logistic model to the data

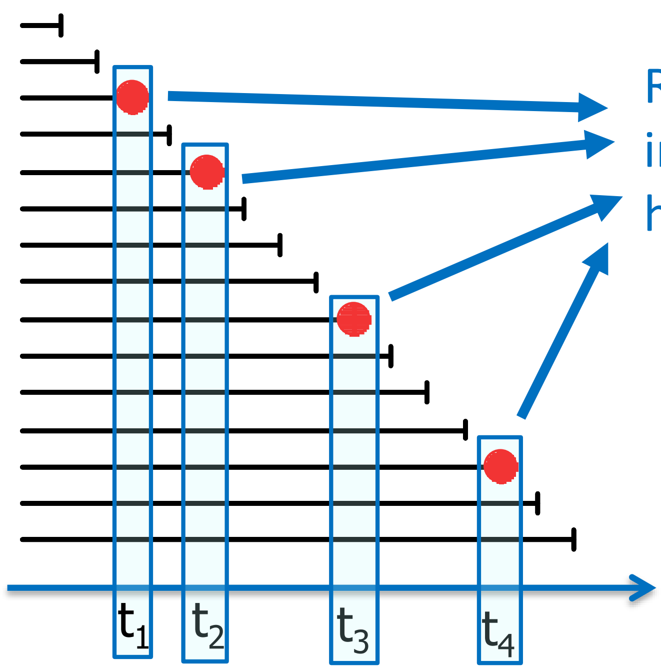
exp^{β} =OR

$$\text{logit}(P[Y=1]) = \boldsymbol{\alpha} + \beta X$$

Cox regression model:
estimates β by maximum (partial) likelihood

At each event time, individuals at risk of the event are called the “risk set”

But only one individual actually has the event (if time is precise)



Risk sets $\boldsymbol{R_i}$**’s** consist of individuals at risk of having the event at time t_i

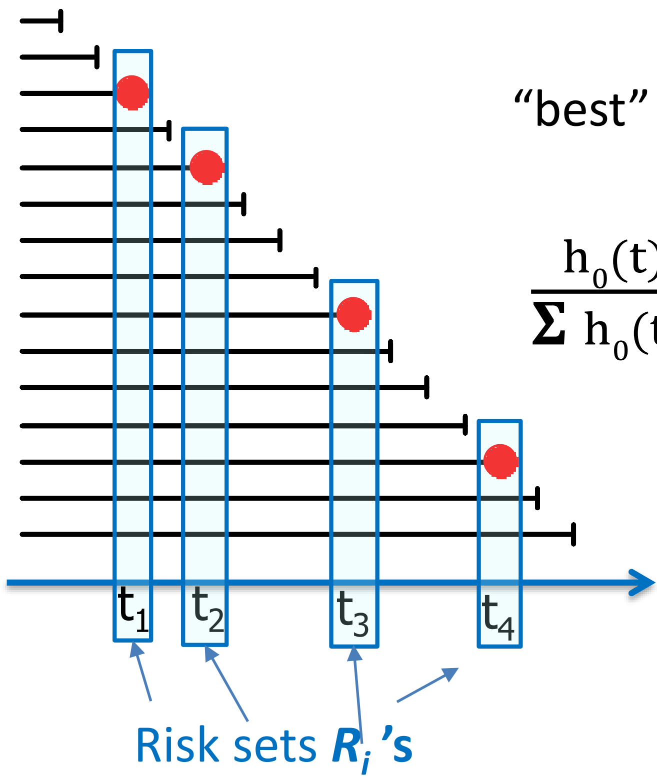
likelihood/hazard for the case that occurred = $\mathrm{h}_0(\mathrm{t})\exp^{\beta X_i}$

The total hazard of all individuals at risk at that time = $\boldsymbol{\Sigma}\ \mathrm{h}_0(\mathrm{t})\exp^{\beta X_k}$

Cox regression model:
estimates β by maximum (partial) likelihood

At each event time   hazard for the case = $\mathrm{h}_0(\mathrm{t})\exp^{\beta X_i}$

total hazard of risk set = $\boldsymbol{\Sigma}\, \mathrm{h}_0(\mathrm{t})\exp^{\beta X_k}$



"best" β maximises the ratio

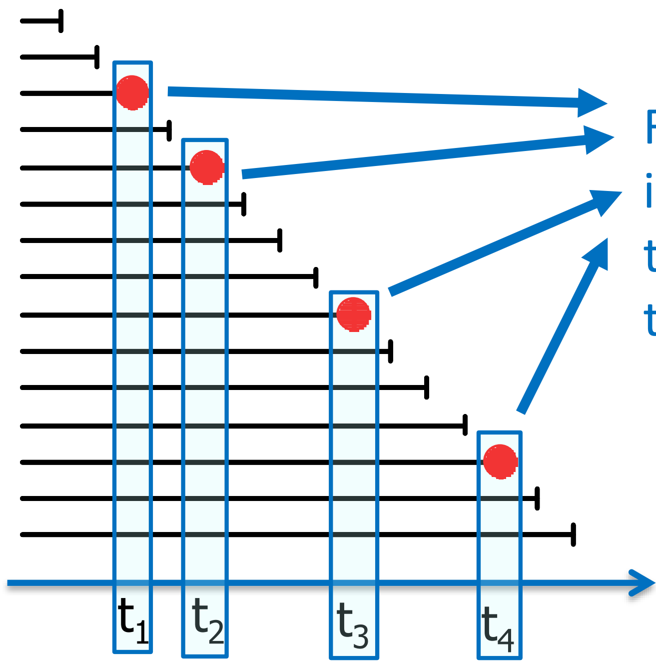
$$\frac{\mathrm{h}_0(\mathrm{t})\exp^{\beta X_i}}{\boldsymbol{\Sigma}\, \mathrm{h}_0(\mathrm{t})\exp^{\beta X_k}} = \frac{\exp^{\beta X_i}}{\boldsymbol{\Sigma}\, (\mathrm{t})\exp^{\beta X_k}}$$

Baseline
hazard
cancels

… at all event times….

Cox regression model: estimates β by maximum (partial) likelihood

Cox partial likelihood: $$L(\beta) = \prod_{t_i} \frac{\exp^{\beta X_i}}{\sum_{k \in R_i} \exp^{\beta X_k}}$$



Risk sets ***R_i* ’s** consist of individuals at risk of having the event at time t_i where t_i is the i-th event time.

Example*

Question: Is the survival of HIV+ individuals with no drug use history different from those with drug use history **after adjusting for age**?

Cox regression model:

$$\ln\{h(t|\mathrm{Drug}_i, \mathrm{Age}_i)\} = \ln\{h_0(t)\} + \beta_1 \mathrm{Drug}_i + \beta_2 \mathrm{Age}_i$$

H_0: $\beta_1 = 0$ (or hazards same: $\exp^{\beta_1} = 1$).

H_1: $\beta_1 \neq 0$ (or hazards different: $\exp^{\beta_1} \neq 1$)

* Data from Hosmer & Lemeshow, *Applied Survival Analysis*, 2nd ed, Wiley 2008

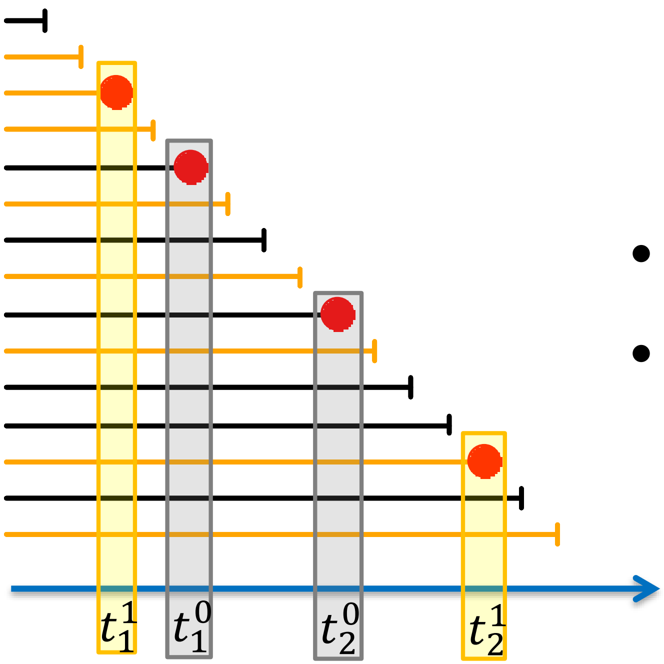
(available from R package "simPH")

Cox regression model

	exp(coef)	exp(-coef)	lower .95	upper .95
DRUG[T.Drug use]	2.764	0.3618	1.673	4.567
AGE	1.102	0.9074	1.062	1.143

- HIV+ individuals with drug use have significantly higher hazard when compared with those with no drug use after adjusting for age (HR = 2.8, 95%CI: 1.7 to 4.6).
- When age increases by 1 unit, the hazard increases by a factor of 1.10 (95% CI: 1.06-1.14; P-value) after adjusting for drug use.

Stratified Cox regression model



- If different baseline hazards for each level of a binary confounder (0: black vs 1: yellow),
- PH assumption not satisfied.
- Can perform a ***stratified*** Cox model (assumes $h_0(t)$ constant within strata):

$$L(\beta) = \prod_{s} \prod_{t_i^s} \frac{\exp^{\beta X_i^s}}{\sum_{k \epsilon R_i^s} \exp^{\beta X_k^s}}$$

Note the parallel to stratified logistic regression, with stratum effect

Survival analysis –final comments

- Kaplan-Meier curves are often used to present data, and a log-rank test used to compare groups
- Most common model in survival analysis is **Cox regression** which estimates the hazard ratio for the exposed compared to the unexposed.