An introduction to survival analysis

Georg Wölflein

School of Computer Science, University of St Andrews

April 11, 2022

Contents

What is time-to-event (TTE) data?

We can measure **time** in:

- years
- months
- seconds

What is time-to-event (TTE) data?

We can measure **time** in:

- years
- months
- seconds

The **event** could be:

- death from disease
- product failure
- losing a customer

What is time-to-event (TTE) data?

We can measure **time** in:

- years
- months
- seconds

The **event** could be:

- death from disease
- product failure
- losing a customer

must be a binary variable

What is time-to-event (TTE) data?

We can measure **time** in:

- years
- months
- seconds

The **event** could be:

- death from disease
- product failure
- losing a customer

must be a binary variable

TTE data consists of $(\textit{time}, \overbrace{\textit{event}}^{\text{yes/no}})$ tuples.

Time-to-event (TTE) data

TTE analysis is also known as:

- survival analysis
- failure time analysis
- reliability theory (engineering)
- duration modelling (economics)
- event history analysis (sociology)

Time-to-event (TTE) data

TTE analysis is also known as:

- survival analysis
- failure time analysis
- reliability theory (engineering)
- duration modelling (economics)
- event history analysis (sociology)

Use cases for TTE analysis:

- TODO

Example: Covid-19 treatment trial

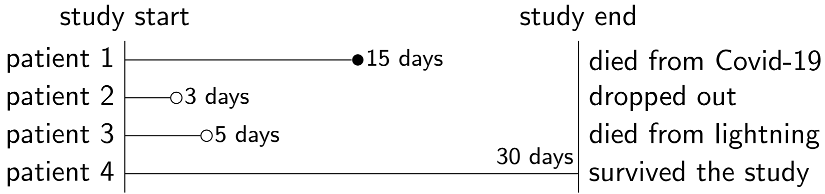
A randomised controlled trial ($n = 4$) was conducted to assess the efficacy of drug ABC in treating Covid-19. This is what happened to the patients:

Example: Covid-19 treatment trial

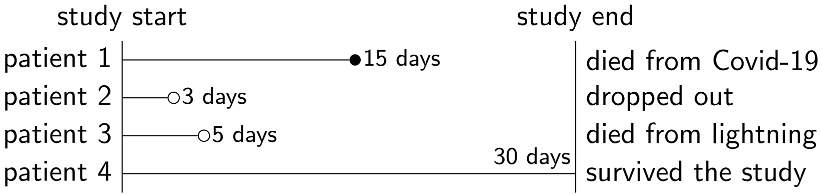
A randomised controlled trial ($n = 4$) was conducted to assess the efficacy of drug ABC in treating Covid-19. This is what happened to the patients:

patient	received ABC?	outcome
1	yes	died from Covid-19 on day 15
2	no	dropped out of the study after day 3
3	yes	died by a lightning stroke on day 5
4	no	survived the study (30 days)

Example: Covid-19 treatment trial

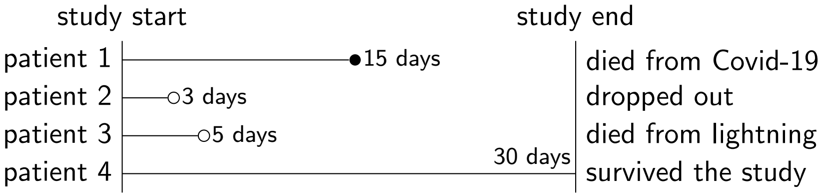


Example: Covid-19 treatment trial



The **time** is the number of days since testing positive for Covid-19.
The **event** is whether the patient died due to Covid-19.

Example: Covid-19 treatment trial

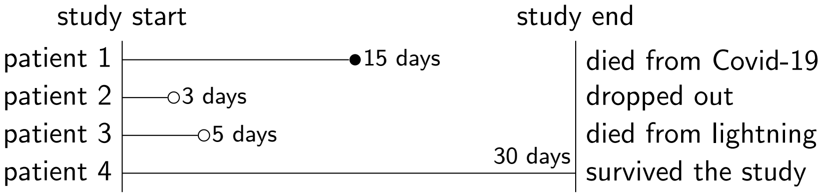


The **time** is the number of days since testing positive for Covid-19.
The **event** is whether the patient died due to Covid-19.

Time-to-event data

patient	time	event
1	15	yes
2	?	?
3	?	?
4	?	no

Example: Covid-19 treatment trial



The **time** is the number of days since testing positive for Covid-19.
The **event** is whether the patient died due to Covid-19.

Time-to-event data

patient	time	event
1	15	yes
2	$[0, 3]$	no
3	$[0, 5)$	no
4	$[0, 30]$	no

Censoring

We just saw examples of **right-censored** data.
TODO

Survival function

Let T be a continuous random variable representing survival time. The **survival function** $S(t)$ is the probability that an individual will survive past time t.

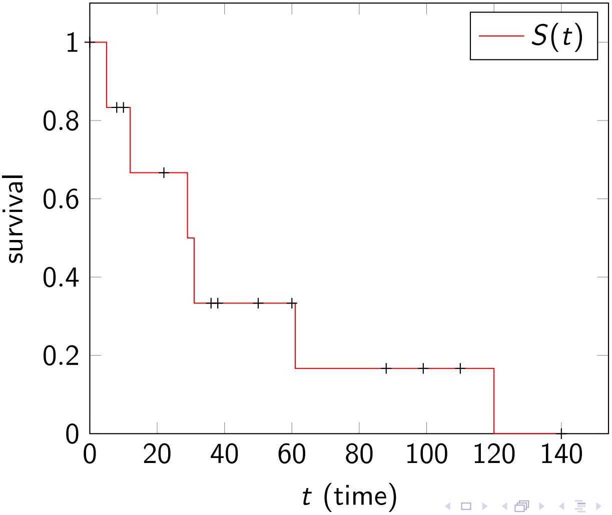
Survival function

Let T be a continuous random variable representing survival time. The **survival function** $S(t)$ is the probability that an individual will survive past time t.

Survival function

$$S(t) = \Pr(T > t)$$

Survival curve



Modelling the survival function

The **Kaplan-Meier estimator** provides a non-parametric estimate of the survival function $S(t)$ using the survival curve.

Modelling the survival function

The **Kaplan-Meier estimator** provides a non-parametric estimate of the survival function $S(t)$ using the survival curve.

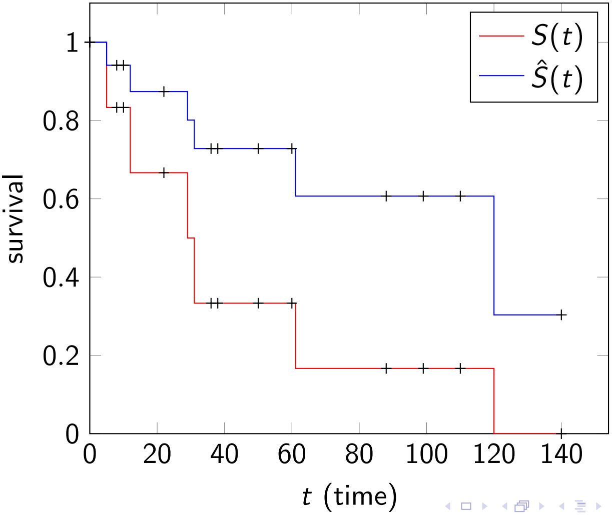
Kaplan-Meier estimator

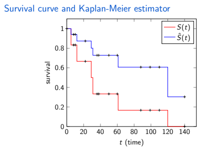
$$\hat{S}(t) = \prod_{i:t_i \leq t} \left(1 - \frac{d_i}{n_i}\right)$$

where

- t_i is an event time
- d_i is the number of deaths at time t_i
- n_i is the number of individuals *known to have survived* until t_i

Survival curve and Kaplan-Meier estimator





- When there is no censoring, $S(t) = \hat{S}(t)$.
- Commonly used to compare two study populations.
- Does not control for covariates.

Hazard function

The **hazard function** expresses the *instantaneous rate of occurence* of the event.

Hazard function

The **hazard function** expresses the *instantaneous rate of occurence* of the event.
Supposing an individual survived until time t, it expresses the probability of dying within a short additional time dt, per unit time.

Hazard function

The **hazard function** expresses the *instantaneous rate of occurence* of the event.
Supposing an individual survived until time t, it expresses the probability of dying within a short additional time dt, per unit time.

Hazard function

$$\lambda(t) = \frac{\Pr\left(\qquad\qquad \middle| \qquad \right)}{\qquad}$$

Hazard function

The **hazard function** expresses the *instantaneous rate of occurence* of the event.
Supposing an individual survived until time t, it expresses the probability of dying within a short additional time dt, per unit time.

Hazard function

$$\lambda(t) = \frac{\Pr(t \leq T \leq t + dt | \qquad)}{}$$

Hazard function

The **hazard function** expresses the *instantaneous rate of occurence* of the event.
Supposing an individual survived until time t, it expresses the probability of dying within a short additional time dt, per unit time.

Hazard function

$$\lambda(t) = \frac{\Pr\left(t \leq T \leq t + dt \middle| T \geq t\right)}{}$$

Hazard function

The **hazard function** expresses the *instantaneous rate of occurence* of the event.
Supposing an individual survived until time t, it expresses the probability of dying within a short additional time dt, per unit time.

Hazard function

$$\lambda(t) = \frac{\Pr\left(t \leq T \leq t + dt | T \geq t\right)}{dt}$$

Hazard function

The **hazard function** expresses the *instantaneous rate of occurence* of the event.
Supposing an individual survived until time t, it expresses the probability of dying within a short additional time dt, per unit time.

Hazard function

$$\lambda(t) = \lim_{dt \to 0} \frac{\Pr(t \leq T \leq t + dt | T \geq t)}{dt}$$

Hazard function

The **hazard function** expresses the *instantaneous rate of occurence* of the event.
Supposing an individual survived until time t, it expresses the probability of dying within a short additional time dt, per unit time.

Hazard function

$$\lambda(t) = \lim_{dt \to 0} \frac{\Pr\left(t \le T \le t + dt \middle| T \ge t\right)}{dt} = \lim_{dt \to 0} \frac{\Pr\left(t \le T \le t + dt\right)}{dt \cdot \Pr\left(T \ge t\right)}$$

Hazard function

The **hazard function** expresses the *instantaneous rate of occurence* of the event.
Supposing an individual survived until time t, it expresses the probability of dying within a short additional time dt, per unit time.

Hazard function

$$\lambda(t) = \lim_{dt \to 0} \frac{\Pr\left(t \leq T \leq t + dt \middle| T \geq t\right)}{dt}$$
$$= \lim_{dt \to 0} \frac{\Pr\left(t \leq T \leq t + dt\right)}{dt \cdot S(t)}$$

Cox proportional hazards model