Lecture 1

Monday, August 28, 2023 09:56

BIOS522_Slides1

EMORY ROLLINS SCHOOL OF PUBLIC HEALTH | Department of Biostatistics and Bioinformatics

BIOS 522: Survival Analysis Methods

Lecture 1:

Introduction to time-to-event data

Welcome to BIOS 522!

- Survival analysis is the branch of statistics that deals with times to events. It has many important applications in clinical research and epidemiology. The goal of this course is to give you a solid understanding of survival analysis and its applications.

2

Your instructor

- Dr. Natalie Dean
  - Assistant Professor in the BIOS Department
  - Research on emerging infectious diseases, vaccine study design
- Office: GCR 336
- Contact: nataliedean@emory.edu or Canvas

3

Your teaching assistant

- Emily Wu
- Contact: emily.wu2@emory.edu or Canvas

4

Course structure

- Before class
  - Read the lecture notes
  - Take a short Canvas quiz
  - One other pre-class responsibility (e.g. homework, discussion, review computing handout)
- In-class
  - Non-exhaustive review of concepts… building content!
  - Examples from the literature
  - Small group activities
- *Active learning!*

5

Review syllabus

Final is take-home project w/ short answer

Course philosophy

- Emphasis on literacy, understanding, and context
- Building a map of concepts >>>> Covering every detail
- Participation >>>> perfection

- *I am glad you are joining me this semester to learn about this important topic. Your feedback throughout the semester is valued. Please do not hesitate to contact me with questions or concerns.*

Today's learning objectives

- *Identify examples of time-to-event analyses in practice*
- *Define the time-to-event data format*
- *Define right censoring and censoring notation*
- *Identify the time origin, event, or time scale from an example*
- *Diagnose the limitations of analyzing data as continuous, binary, or as incidence rates*

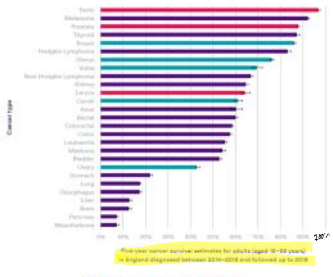
Survival analysis

- Survival analysis is the branch of statistics that deals with times to events.
  - Survival time, failure time, occurrence time, event time, time-to-event

- Examples of survival times in clinical research:
  - Time from diagnosis until death
  - Time from infection until diagnosis
  - Time from treatment until suppression of symptoms
  - Length of stay in a hospital

Survival analysis examples

- Survival analysis in action



Source: https://www.gov.uk/government/statistics/cancer-survival-in-england-for-patients-diagnosed-between-2014-and-2018-and-followed-up-until-2019

Estimated by survival analysis!

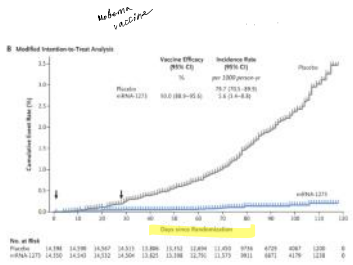


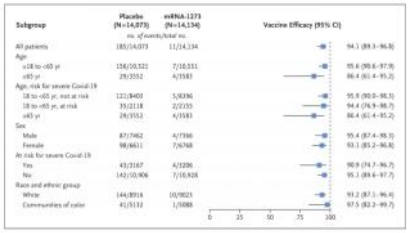
Figure 3. Vaccine Efficacy of mRNA-1273 to Prevent Covid-19.

Shown is the cumulative incidence of Covid-19 events in the primary analysis based on adjudicated assessment starting 14 days after the second vaccination in the per-protocol population (Panel A) and after randomization in the modified intention-to-treat population (Panel B) (see the Supplementary Appendix). The dotted line in Panel A indicates day 42 (14 days after vaccination 2), when the per-protocol follow-up began, and arrows in both panels indicate days 1 and 29, when injections were administered. Tick marks indicate censored data. Vaccine efficacy was defined as 1 minus the hazard ratio (mRNA vs. placebo), and the 95% confidence interval was estimated with the use of a stratified Cox proportional hazards model, with Efron's method of tie handling and with treatment group as a covariate, with adjustment for stratification factor. Incidence was defined as the number of events divided by number of participants at risk and was adjusted by person-years. Symptomatic Covid-19 case accrual for placebo and vaccine in the modified intention-to-treat population is displayed (does not include asymptomatic cases of SARS-CoV-2 detected at the day 29 bi-nasopharyngeal swab).

Source: Baden et al. (2020) NEJM https://www.nejm.org/doi/full/10.1056/nejmoa2035389

Cox model
hazard ratio

Figure 4. Vaccine Efficacy of mRNA-1273 to Prevent Covid-19 in Subgroups.



Source: Baden et al. (2020) NEJM https://www.nejm.org/doi/full/10.1056/nejmoa2035389

Survival analysis overview

In order to analyze time-to-event data, it is necessary to define:

1. **The time origin**: the beginning of the survival time
2. **The failure time**: the end of the survival time

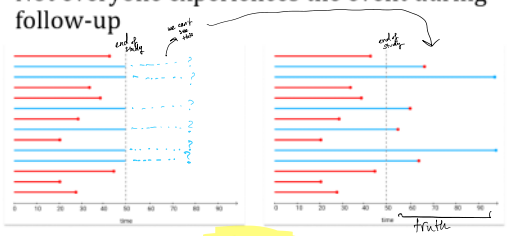
Survival time T measures the time elapsed from the origin ("time zero") until the event of interest

Setting	Time origin	Event	Time scale
Human mortality	Birth	Death	Age
Clinical trial of treatment	Randomization	Stroke or cardiovascular death	Time since start of treatment
Pregnancy cohort	12 weeks gestation	Fetal death	Gestational age
Hospital study	Admission	Discharge	Time in hospital
Surgical study	Surgery	Death or complication	Time since surgery
Cancer cohort	Diagnosis	Tumor recurrence	Time since diagnosis
Ebola survival study	Date of symptom onset	Death due to Ebola	Time since symptom onset
Influenza study	Start of flu season (October 1, 2018)	Influenza symptom onset	Calendar time

? Calendar date

eg. used in pregnancy studies

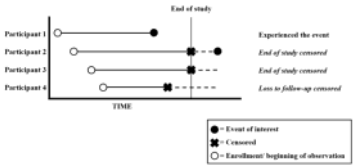
Not everyone experiences the event during follow-up



Source: https://flexy.cc/censoring-data-in-survival-analysis/

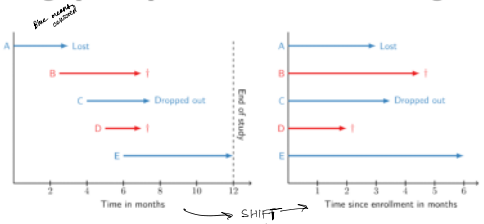
10

Some may be censored earlier
Enrollment may be rolling over time



17

Align participants at common time origin



Source: https://k-d-w.org/blog/2019/07/survival-analysis-for-deep-learning/

18

Censoring notation

- We imagine that everyone in our study has two random variables:
  - T is their **failure time**
  - C is their **censoring time** (end of study, date when they will move)
- In practice, we only observe whichever comes first
- We introduce new notation for the observed data
  - $T^* = \min(T, C)$ is the observed time (failure or censoring)
  - $\delta = I[T \leq C]$ is an indicator (0 or 1) for observing a failure time

T^* for observed time of what-comes-first

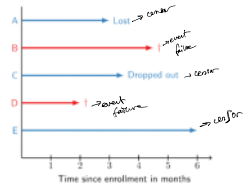
δ → if this event is FAIL or CENSOR

Underlying data		Observed data	
T_i	C_i	T_i^*	δ_i
3	6	3	1
5	6	5	1
8	6	6	0

Truth — What we see

15

Censoring notation



	Observed data	
ID	T_i^*	δ_i
A	3	0
B	4.5	1
C	3.5	0
D	2	1
E	6	0

Using shorthand:
3+, 4.5, 3.5+, 2, 6+
2, 3+, 3.5+, 4.5, 6+

+ means censoring time

20

So what do survival data look like?

Patient ID	Date of surgery	Date of death	Date of censoring
1	January 1, 2020		February 1, 2020
2	January 5, 2020	January 13, 2020	
3	January 7, 2020		February 7, 2020
4	January 18, 2020		February 18, 2020
5	January 20, 2020	January 31, 2020	
6	January 22, 2020		February 22, 2020

Patient ID	Time-to-event	Event indicator (1=death/0=censoring)
1	31 days	0
2	8 days	1
3	31 days	0
4	31 days	0
5	11 days	1
6	31 days	0

Using shorthand:

31+, 8, 31+, 31+, 11, 31+

Using shorthand, sorted:

8, 11, 31+, 31+, 31+, 31+

Example: Worcester Heart Attack Study

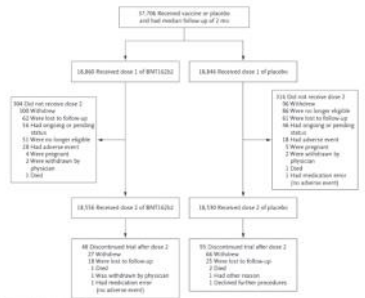
id	age	gender	admitdate	fdate	S fstat
3	70	1	01/01/1997	12/31/2002	0
154	72	1	01/02/1999	06/17/2003	1
163	84	1	01/03/1999	12/27/1999	1
230	63	0	01/03/1999	12/31/2002	0
23	73	0	01/05/1997	03/07/1997	1
325	61	0	01/05/1999	12/31/2002	0
466	57	1	01/05/2001	12/31/2002	0
96	68	0	01/07/1997	08/11/1998	1
219	47	0	01/07/1999	12/31/2002	0
311	84	1	01/07/1999	12/31/2002	0
166	72	1	01/08/1999	01/19/1999	1
142	58	1	01/10/1999	12/31/2002	0
6	66	0	01/11/1997	02/19/2001	1
177	90	1	01/11/1999	01/12/1999	1
184	88	1	01/12/1999	04/21/2000	1
171	68	0	01/12/1999	12/31/2002	0
1	83	0	01/13/1997	12/31/2002	0
485	77	0	01/15/2001	04/27/2002	1

→ need another column for difference in days

Source: Hosmer, Lemeshow, May (2008) John Wiley and Sons Inc.

Other reasons for censoring

- Lost to follow-up
- Move out of the study region
- No longer meet study eligibility criteria
- No longer "at-risk" (e.g., death from unrelated cause)



Source: Polack et al. (2020) NEJM doi: 10.1056/NEJMoa2034577

Why survival analysis?

- Why do we need specialized methods for time-to-event data when we have methods for analyzing:
  - Continuous data
  - Binary data
  - Incidence rate data (person-time)

Hypothetical study

- We design a study to estimate survival for women diagnosed with stage II breast cancer.
- We can imagine forming a cohort of newly diagnosed women and following them prospectively in time to track survival outcomes.

Continuous data

- Imagine that our cohort was comprised of **elderly women** (≥75 years), and we continued our study for 20+ years until all women had died. For each woman we can calculate the event time, which is the **time elapsed from diagnosis to death**.
- We can report the **mean or median survival time**.

- What if our cohort was comprised of **middle-aged women** tracked for 5 years only?
- *Can we calculate the mean or median survival time?*

Why not continuous data?

- $n = 5$ women
  - 1 woman dies 3 years after diagnosis
  - 1 woman dies 4 years after diagnosis
  - 3 survive to the end of the 5-year study (survival time is "**censored**")

- Calculate mean time:

$$\frac{1}{n}\sum_{i=1}^{n} T^* = \frac{3+4+5+5+5}{5} = 4.4 \text{ years}$$

? this is **MEAN FOLLOW-UP TIME**

not mean diagnosis until death

Why not continuous data?

- $n = 5$ women
  - 1 woman dies 3 years after diagnosis
  - 1 woman dies 4 years after diagnosis
  - 3 survive to the end of the 5-year study (survival time is "**censored**")

- Calculate mean SURVIVAL time:

$$\frac{1}{n}\sum_{i=1}^{n} T = \frac{3+4+?+?+?}{5} = ?? \text{ years}$$

Why not continuous data?

- $n = 5$ women
  - 1 woman dies 3 years after diagnosis
  - 1 woman dies 4 years after diagnosis
  - 3 survive to the end of the 5-year study (survival time is "**censored**")

- Median survival time: (3, 4, 5, ?, ?) = ??

Why not continuous data?

- $n = 5$ women
  - 1 woman dies 3 years after diagnosis
  - 1 woman dies 4 years after diagnosis
  - *1 woman dies 4.5 years after diagnosis*
  - 2 survive to the end of the 5-year study (survival time is "**censored**")

- Median survival time: (3, 4, 4.5, ?, ?) = 4.5 years

these times are > 5

Administrative censoring = censoring at end of study

Why not continuous data?

- $n = 5$ women
  - 1 woman dies 3 years after diagnosis
  - 1 woman dies 4 years after diagnosis
  - *1 woman moves away 2 years after diagnosis* ***(censored)***
  - 3 survive to the end of the 5-year study (survival time is "**censored**")

- Median survival time: (?, 3, 4, ?, ?) = ??

Why not binary data?

- Another data summary is **5-year mortality** (yes/no, binary)
- $n = 5$ women
  - 1 woman dies 3 years after diagnosis
  - 1 woman dies 4 years after diagnosis
  - 3 survive to the end of the 5-year study (survival time is **censored**)

- 5-year mortality

$$\frac{1}{n}\sum_{i=1}^{n} I[T_i \leq 5] = \frac{2}{5}$$

probability that they will die 5 years before

Why not binary data?

- $n = 5$ women
  - 1 woman dies 3 years after diagnosis $T_i = 3 \rightarrow T_i^* = 3, \delta_i = 1$
  - 1 woman dies 4 years after diagnosis $T_i = 4 \rightarrow T_i^* = \delta_i = 1$
  - 3 survive to the end of the 5-year study $C_i = 5 \rightarrow T_i^* = 5, \delta_i = 0$

- For the 3 women censored at 5 years, we know $T_i > 5$

Why not binary data?

- $n = 5$ women
  - 1 woman dies 3 years after diagnosis
  - 1 woman dies 4 years after diagnosis
  - *1 woman moves away at 2 years after diagnosis*
  - 2 survive to the end of the 5-year study (survival time is **censored**)

Why not binary data?

- $n = 5$ women
  - 1 woman dies 3 years after diagnosis $T_i = 3 \rightarrow T_i^* = 3, \delta_i = 1$
  - 1 woman dies 4 years after diagnosis $T_i = 4 \rightarrow T_i^* = 3, \delta_i = 1$
  - *1 woman moves away at 2 years after diagnosis $C_i = 2 \rightarrow T_i^* = 2, \delta_i = 0$*
  - 2 survive to the end of the 5-year study $C_i = 5 \rightarrow T_i^* = 5, \delta_i = 0$

- 5-year mortality

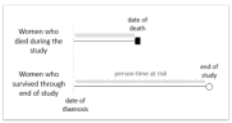
$$\frac{1}{n}\sum_{i=1}^{n} I[T_i \leq 5] = \frac{?}{5}$$

Why not binary data?

- Necessary to pick a single time point for analysis
- Is 10 years more important than 5 years?

Why not incidence rate data?

- A **rate-based analysis** can be used when there are differing lengths of follow-up.
- For each woman, we can calculate the length of time during which the event could have occurred and would have been counted in the population, known as the **person-time**.



Why not incidence rate data?

- $n = 5$ women
  - 1 woman dies 3 years after diagnosis
  - 1 woman dies 4 years after diagnosis
  - 3 survive to the end of the 5-year study (survival time is **censored**)

Yearly mortality rate

$$\text{incidence rate} = \frac{\#\ of\ women\ who\ died\ in\ our\ study}{total\ person\text{-}years\ at\ risk\ after\ diagnosis}$$

Why not incidence rate data?

- $n = 5$ women
  - 1 woman dies 3 years after diagnosis $T_i = 3 \rightarrow T_i^* = 3, \delta_i = 1$
  - 1 woman dies 4 years after diagnosis $T_i = 4 \rightarrow T_i^* = 3, \delta_i = 1$
  - 3 survive to the end of the 5-year study $C_i = 5 \rightarrow T_i^* = 5, \delta_i = 0$

Yearly mortality rate

$$\text{incidence rate} = \frac{\sum_{i=1}^{n} \delta_i}{\sum_{i=1}^{n} T_i^*}$$
$$= \frac{2}{(3+4+5+5+5)}$$
$$= 0.09\ \text{deaths/person-year}$$

but what if you wanted it to vary over time?

Why not incidence rate data?

- **Rate-based analyses** are related to simple survival analysis methods.
- Rate-based analyses assume that the **event rate is constant**.
- The majority of survival analysis methods that we will learn about in this class **allow the event rate to vary over time**.

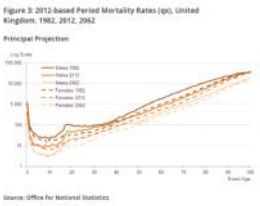
- For example, mortality rates for women newly diagnosed with breast cancer may be **initially high**, as some women may have aggressive or difficult-to-treat forms. Women who survive more than 5 years after diagnosis, though, may have mortality rates closer to the general population.

Why not incidence rate data?

- **Rate-based analyses** are related to simple survival analysis methods.
- Rate-based analyses assume that the **event rate is constant**.
- The majority of survival analysis methods that we will learn about in this class **allow the event rate to vary over time**.

- For example, mortality rates for women newly diagnosed with breast cancer may be **initially high**, as some women may have aggressive or difficult-to-treat forms. Women who survive more than 5 years after diagnosis, though, may have mortality rates closer to the general population.

Human mortality



Source: Office for National Statistics UK, 2013

Today's activity

- Word problems
- Work in assigned small groups