
Department of Statistical Sciences

# Part I. Introduction

Sarah Lotspeich

STA779: *Applied Survival Analysis (Spring 2023)*

# Fathers of survival

CAPTAIN JOHN GRAUNT

EDMOND HALLEY

*From an engraving by Vertue of the original portrait by R. Phillips*

# It all starts with the end

- More than 360 years ago, "survival analysis" was born out of the **seventeenth century mortality studies**.
- Unclear who *really* did it first: **John Graunt** with his Weekly Bill of Mortality in London or **Edmond Healey** with his lifetables.

| *Age.* *Curt.* | *Persons* | *Age.* *Curt.* | *Persons* | *Age.* *Curt.* | *Persons* | *Age.* *Curt.* | *Persons* | *Age.* *Curt.* | *Persons* | *Age.* *Curt.* | *Persons* | *Age* | *Persons* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1000 | 8 | 680 | 15 | 628 | 22 | 586 | 29 | 539 | 36 | 481 | 7 | 5547 |
| 2 | 855 | 9 | 670 | 16 | 622 | 23 | 579 | 30 | 531 | 37 | 472 | 14 | 4584 |
| 3 | 798 | 10 | 661 | 17 | 616 | 24 | 573 | 31 | 523 | 38 | 463 | 21 | 4270 |
| 4 | 760 | 11 | 653 | 18 | 610 | 25 | 567 | 32 | 515 | 39 | 454 | 28 | 3964 |
| 5 | 732 | 12 | 646 | 19 | 604 | 26 | 560 | 33 | 507 | 40 | 445 | 35 | 3604 |
| 6 | 710 | 13 | 640 | 20 | 598 | 27 | 553 | 34 | 499 | 41 | 436 | 42 | 3178 |
| 7 | 692 | 14 | 634 | 21 | 592 | 28 | 546 | 35 | 490 | 42 | 427 | 49 | 2709 |
| *Age.* *Curt.* | *Persons* | *Age.* *Curt.* | *Persons* | *Age.* *Curt.* | *Persons* | *Age.* *Curt.* | *Persons* | *Age.* *Curt.* | *Persons* | *Age.* *Curt.* | *Persons* | 56 | 2194 |
| | | | | | | | | | | | | 63 | 1694 |
| 43 | 417 | 50 | 346 | 57 | 272 | 64 | 202 | 71 | 131 | 78 | 58 | 70 | 1204 |
| 44 | 407 | 51 | 335 | 58 | 262 | 65 | 192 | 72 | 120 | 79 | 49 | 77 | 692 |
| 45 | 397 | 52 | 324 | 59 | 252 | 66 | 182 | 73 | 109 | 80 | 41 | 84 | 253 |
| 46 | 387 | 53 | 313 | 60 | 242 | 67 | 172 | 74 | 98 | 81 | 34 | 100 | 107 |
| 47 | 377 | 54 | 302 | 61 | 232 | 68 | 162 | 75 | 88 | 82 | 28 | | 34000 |
| 48 | 367 | 55 | 292 | 62 | 222 | 69 | 152 | 76 | 78 | 83 | 23 | Sum | Total |
| 49 | 357 | 56 | 282 | 63 | 212 | 70 | 142 | 77 | 68 | 84 | 20 | | |

Figure: Edmond Healy's lifetable of the city of Breslau (1687–1691)

Read more: **(1)** Lee, Elisa T. and Go, Oscar T. "Survival Analysis in Public Health Research." *Annual Review of Public Health* vol. 18, no. 1, 1997, pp. 105–134 and **(2)** Bellhouse, David R. "A New Look at Halley's Life Table." *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, vol. 174, no. 3, 2011, pp. 823–32.

# From time to death to time to... anything

Graunt and Healey started out by measuring **time to death**, but in the centuries since survival analysis methods have **become more widely applicable**.

- During World War II (1939–1945), reliability of military equipment became a critical issue. This led to the study of the durability or the **"lifetime" of industrial devices** (rather than people).
- After the War, what engineers called "lifetime analysis" was adopted by **cancer researchers and rebranded as "survival analysis."**
- Become **one of the most frequently used methods** in many disciplines: medicine, epidemiology, environmental health, criminology, marketing, astronomy.

Read more: Lee, Elisa T. and Go, Oscar T. "Survival Analysis in Public Health Research." *Annual Review of Public Health* vol. 18, no. 1, 1997, pp. 105–134.

# Examples of time-to-event data

Survival analysis methods are most often employed with **"time-to-event"** or **"failure-time"** data. Now with applications across many disciplines, such as medicine, biology, public health, epidemiology, examples might include

- Time to death (in days),
- Time to onset of a disease (in days),
- Time to relapse of a disease (in days),or
- Length of stay in a hospital, i.e., time to discharge (in days).

These methods are also applied in fields like engineering, economics, and demography, but we will focus on biomedical applications in this class.

# And really, it doesn't have to be "time" at all

More generally, survival analysis methods can be used on **any positive real-valued random variables**, like

- Money paid by health insurance (in dollars),
- Viral load measurements for patients with HIV (in copies/milliliter),
- Patients' weight after starting a new treatment (in pounds or kilograms), or
- Distance run in a 1-hour period (in miles or kilometers).

However, the "time" variables are probably still the most common.

# Your turn: Examples of survival data

- time to cheese maturation
- time to accident (for insurance)
- time to COVID recovery
- time to seasonal depression
- time to protein dissociation
- time to "recession"
- time to graduation

# Course objectives

In this course, we will discuss how to do the following.

- Diagnose different types of survival data and appropriate methods to analyze them.
- Identify basic quantities and models, as well as know how to compare them between samples or subgroups.
- Conduct basic inference for the associations between time-to-event outcomes and covariates. Also, learn the relative merits of different approaches.
- Know and test for all necessary assumptions to ensure valid statistical inference, as well as methods to correct for violations of these assumptions.
- Apply all of these concepts in R to describe and analyze various survival datasets.

# Connection to discrete outcomes

As the name "time-to-event" suggests, we are often interested in the **absence or presence of an event** (e.g., death or disease onset) as our outcome.

- Such an outcome is captured as a discrete (i.e., binary) variable like

$$\Delta_i = \mathrm{I}(\text{Event happened}),$$

  where $\mathrm{I}(\cdot)$ is the **indicator function**, i.e., $\mathrm{I}(\cdot) = 1$ if the condition is true and $0$ otherwise.

- If data were collected over a fixed time period (like 2 years) and you were interested in comparing 2-year mortality between subgroups, you might fit a **logistic regression**.

- But we often don't collect data this way and/or aren't interested in these questions.

- Plus, this model is throwing away potentially valuable information about **how fast** people in different subgroups died. You'd need survival methods to capture that.

# Data collection

Survival data can be collected many ways, but most commonly...

1. **Clinical trials**:
2. **Prospective cohort studies**:
3. **Retrospective cohort studies**:

There is one major way in which survival data tend to be unique: they are often **not fully observed, but rather censored or truncated**.

# Data collection

Survival data can be collected many ways, but most commonly...

1. **Clinical trials**: "A research study in which one or more human subjects are prospectively assigned to one or more interventions (which may include placebo or other control) to evaluate the effects of those interventions on health-related biomedical or behavioral outcomes." (from the National Institutes of Health [NIH])
2. **Prospective cohort studies**:
3. **Retrospective cohort studies**:

There is one major way in which survival data tend to be unique: they are often **not fully observed, but rather censored or truncated**.

# Data collection

Survival data can be collected many ways, but most commonly...

1. Clinical trials:
2. **Prospective cohort studies**: "A research study that follows over time groups of individuals who are alike in many ways but differ by a certain characteristic (for example, female nurses who smoke and those who do not smoke) and compares them for a particular outcome (such as lung cancer)." (from the National Cancer Institute [NCI])
3. **Retrospective cohort studies**:

There is one major way in which survival data tend to be unique: they are often **not fully observed, but rather censored or truncated**.

# Data collection

Survival data can be collected many ways, but most commonly...

1. Clinical trials:
2. Prospective cohort studies:
3. **Retrospective cohort studies**: "A research study in which the medical records of groups of individuals who are alike in many ways but differ by a certain characteristic (for example, female nurses who smoke and those who do not smoke) are compared for a particular outcome (such as lung cancer). Also called historic cohort study." (from the National Cancer Institute [NCI])

There is one major way in which survival data tend to be unique: they are often **not fully observed, but rather censored or truncated**.

# Data collection

Survival data can be collected many ways, but most commonly...

1. Clinical trials:
2. Prospective cohort studies:
3. Retrospective cohort studies:

There is one major way in which survival data tend to be unique: they are often **not fully observed, but rather censored or truncated**.

# Key factors in defining a survival outcome

Defining a failure-time variable $X$ requires the following:

1. A clear time **origin** (i.e., when you started counting $X$ or what $T = 0$ represents),
2. A set time **scale** (i.e., how you're going to count or increment $X$),
3. A definition of the failure **event** (i.e., what you're interested in or where $X$ is intended to "stop").

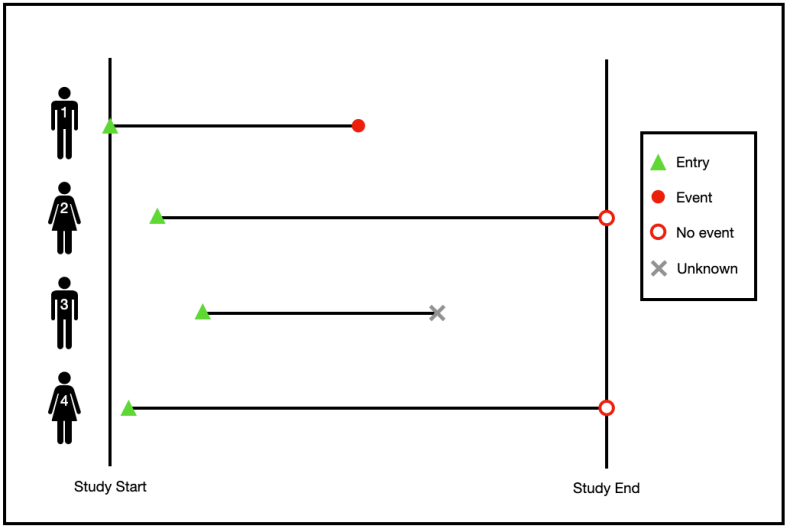


Figure: Example of (right) censored data.

# Typical features of survival data

The Figure on the previous slide illustrates several typical features of survival data:

- **Staggered entry:** Individuals all enter the study at different times.
  - Ex: The study may start when the funding comes through, but individuals are recruited (and therefore, enter the study) on a rolling basis thereafter.
- **Censoring:** Not everyone had the event during the study, but they (probably) aren't immune. We just know that it hasn't happened to them yet.
  - Ex: If someone is still alive at the end of a 10-year study of mortality, then they (probably) aren't immortal. We think that they will one day die, but that day must be after the study ends.
  - Note that Subjects 2–4 are all censored, but Subjects 2 and 4 were censored at a different value than Subject 3 was.

# Your turn: Identifying key factors

**AIDS Diagnosis:** *Consider an existing cohort of people living with HIV (PLWH). Researchers are interested in studying the time between antiretroviral treatment (ART) initiation and AIDS diagnosis. (They are specifically interested in comparing this variable between subgroups, but we'll get to that later.)*

Define a failure-time outcome $X$ that would help them answer this question. Specifically, note the (1) origin, (2) scale, and (3) event definition for $X$.

(1) origin: ART initiation

(2) scale: days

(3) event: AIDS diagnosis

# A primer on censoring

- **Censoring** is a key characteristic (and challenge) in analyzing survival data.
- There are many **reasons for censoring**, including loss to follow-up, drop-out, study termination, or machine limits of detection (LOD).
- Different reasons lead to different **types of censoring**, most common ones being
  1. Right censoring (RC),
  2. Left censoring (LC),
  3. Interval censoring (IC), or
  4. Double censoring (DC).
- Also, any of these types can be **random or non-random censoring**, depending on whether $C$ is fixed (e.g., by the machine's LOD) or a random itself (e.g., subject-specific time to drop-out).