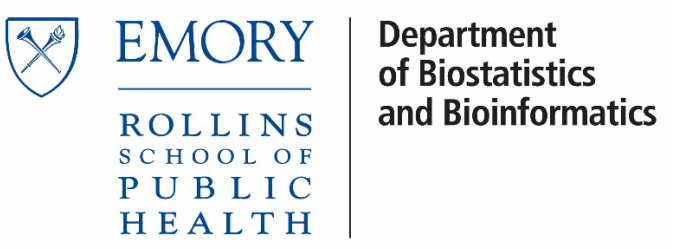


BIOS 522: Survival Analysis Methods

Reading 1:

Introduction to time-to-event data

This week, we will provide an overview of time-to-event data. We will study examples of failure time events, time origins, and time scales. We will define and characterize right censoring. We will review the analysis of continuous, binary, and person-time data.

Part 1. Welcome to survival analysis

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.

The time until an event occurs is called the **survival time** or, equivalently, the **failure time**. It is also referred to as the **occurrence time**, **event time**, or **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

Examples from other disciplines:

- Education: time to graduation
- Sociology: time to divorce
- Economics: length of unemployment
- Engineering: the lifetime of mechanical or electrical devices

Often, we are interested in summarizing the distribution of survival times for a particular population, such as 5-year survival from the time of diagnosis, or the median length of stay in the hospital.

Example: 5-year survival for women with stage II breast cancer is 93%, meaning 93% of women will survive at least 5 years after diagnosis.

We may also examine whether survival times are different between groups, such as shorter times until suppression of symptoms with an experimental treatment as compared to standard of care.

Example: Among women with stage II breast cancer, women receiving tamoxifen survive significantly longer than women who do not receive tamoxifen.

Time origin, event, and scale

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

(1) **Event**: the outcome of interest
(2) **Time origin**: the beginning of the survival time ("time zero")

Some examples are provided in the table below:

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

Note that the event of interest can be a single event or a **composite** (e.g., death or complication, whichever happens first). The event can also be desirable, undesirable, or neutral. For example, early discharge from the hospital is desirable while early death is not. The analytical methods we use are the same regardless.

When we analyze survival data, we align all observations with respect to their time origin. This determines the time scale used in the analysis. A good goal is to pick a time scale that best captures the underlying biological, physical or other mechanism we seek to study.

Example: For pregnancy-related outcomes, it makes sense to align participants with respect to their gestational age. Twenty weeks gestation will have a similar meaning across all participants. This captures the biological process of fetal development.

Example: For a surgical study, risk of death or complications may be highest in the weeks immediately after surgery. Time since surgery is thus a natural time scale for post-surgical outcomes. For personal-decision making, patients will want an estimate of the level of risk they face post-surgery, and they will want to know when they have passed the period of highest risk.

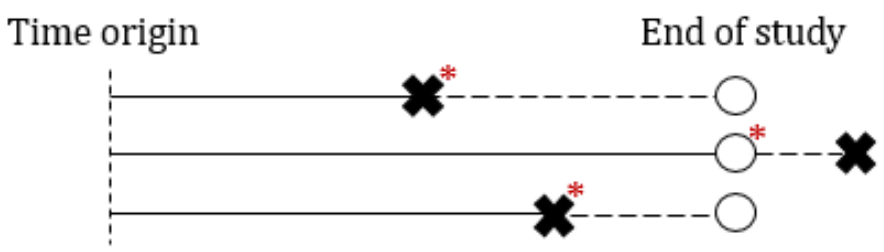
Note that, except for the influenza study example, the calendar date of the time origin will vary across study participants.

Example: For a surgical study, patients may undergo surgery on different calendar dates. Yet in our analysis, they will be aligned so that their day zero is the date of surgery. They will subsequently be combined and compared one week after surgery, one month after surgery, and so on.

Part 2. Censoring

Survival times are typically not known for all subjects. We may be studying events that are not inevitable or that may not occur during the period of observation.

Example: Consider the following hypothetical data, where a black X is time of death, and a white O is the end of the study. Data are shown for three participants. We will only observe the deaths that occur before the end of the study. The observed times are marked with red stars.



The survival time of a study subject is said to be **censored** when the endpoint of interest has not been observed for that individual. Instead of capturing the failure time, we capture the **censoring time**, which is the last time the participant was observed. We refer to this type of censoring as **right censoring** because we won't observe a failure time to the *right* of a censored variable (see figure above).

When a study ends, participants are no longer under observation. In this setting, we refer to any individual who did not experience the event by the time of study end as **administratively censored**. Their censoring time is the end of the study.

If an individual is **lost-to-follow-up** during the study, that person is no longer under observation. If that person went on to have the event, we would not observe it. This person is censored at the time of loss-to-follow-up.

Example: For a study of cancer recurrence, some participants may move out of town while the study is ongoing. They will have new doctors managing their care, and so follow-up records will not be available after they move. Study investigators conduct an exit interview with the participant, and this person is censored at the time of the move.

In the above examples, individuals were censored because they were no longer under observation (e.g., the study ended or they moved away). In some cases, we censor participants because they are no longer **at-risk** for the event, even if they are still observable.

Example: In a study for women at high risk for the development of endometrial cancer, a woman who has a hysterectomy is no longer at risk for the event of endometrial cancer. She is censored at the time of the hysterectomy.

Sometimes a person may die for reasons unrelated to the event being studied. They may be censored at the time of death.

Example: In a study where the event is cardiovascular death or stroke, a person who dies in a traffic accident is no longer at risk for cardiovascular death or stroke. The study protocol specifies that participants who die due to unrelated causes are censored at the time of death.

Censored data formatting

In time-to-event data, the time of the event (measured as time elapsed since the time origin) is recorded for each participant. For any participant whose event time is not observed, the censoring time is recorded instead. To distinguish

which times are event times and which times are censoring times, an additional event indicator variable is created.

Imagine that each individual has a hypothetical failure time T ($T \geq 0$). Each individual also has a hypothetical censoring time C ($C \geq 0$). In practice, we will only observe whichever time comes first.

We introduce notation for our observed data. We observe the random variable $T^* = \min(T, C)$. Thus, we only directly observe the failure time T if $T \leq C$; otherwise we observe C. We refer to T^* as a (right) censored failure time random variable. The * indicates that some of the times are censored.

In our data, we must keep track of whether a time is a failure time or a censoring time. To capture this, we use an **event indicator** $\delta = I[T \leq C]$. The indicator $\delta = 1$ if we observe a failure time, and $\delta = 0$ if we observe a censoring time.

Thus, we represent each observation in our data set by two variables, T^* and δ, which tell us respectively the follow-up time and whether this follow-up ended in a failure or censoring.

Example: Consider hypothetical survival data for three individuals, described below. The censoring time is the end of the study, which occurs at 6 months. The theoretical underlying data are on the left. The observable data are on the right.

Underlying data		**Observed data**	
T_i	C_i	T_i^*	δ_i
3 (failed first)	6	3	1 (failure time)
5 (failed first)	6	5	1 (failure time)
8	6 (censored first)	6	0 (censoring time)

Example: A surgical study is conducted where the event is death by any cause. Participants are administratively censored one month after the date of surgery. The underlying data may look like the following:

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

To convert these data to an analyzable form, the elapsed times are calculated, and an event indicator is added.

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

We could also display these data using common shorthand, where "+" indicates censoring times. 31+ denotes that patients survived at least 31 days but were subsequently censored.

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

Or we could sort the data by time to make it more readable:

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

Part 3. Analysis of continuous data, proportions, and incidence rates

Why do we have specialized analysis methods for time-to-event data? To motivate the methods we will cover in this course, we will review other common types of analysis methods. These are methods for continuous, binary, and rate data.

We will use the motivating example of estimating survival for women newly diagnosed with stage II breast cancer. We can imagine forming a cohort of newly diagnosed women and following them prospectively for five years to monitor for death by any cause. The $i = 1, \dots, n$ women in the cohort have true survival times T_i, but we only direct observe the right-censored times T_i^*.

Analysis of continuous data

Time from cancer diagnosis to death is a continuous variable. Could we analyze time from cancer diagnosis to death as a continuous variable?

We have many statistical tools for the analysis of continuous variables. We could report mean and median time-to-death. We could conduct two-sample t-tests or Wilcoxon rank sum tests to compare independent groups (e.g. stage 2A vs. stage 2B cancer). We could use linear regression to model the effects of continuous covariates or many covariates at once (e.g. age at diagnosis, race/ethnicity).

Importantly, we do not use these continuous data methods to analyze time-to-event data *survival times may not be observed for everyone in the study*. The

survival time T_i is not known for any individual who is censored. The continuous data methods listed are not designed to handle censoring.

Returning to our breast cancer cohort example, unless the women were very old and the study lasted a very long time, many women will still be alive at the end of the study period. Taking the average of follow-up times $\sum_{i=1}^{n} T_i^* /n$, where some times are time-to-death and some are censoring times, is not meaningful for understanding breast cancer survival.

Analysis of binary data

Another way to approach the analysis is to define a set time-period (e.g. five years) and treat the outcome of whether or not the event occurred during the time period as a binary (yes/no variable). For example, we are interested in calculating the **five-year survival** (the proportion of women who survive to five years) or **five-year mortality** (the proportion of women who die by five years).

$$\text{Five-year mortality} = \frac{1}{n}\sum_{i=1}^{n} I[T_i \leq 5]$$

Five-year mortality is the same as the five-year **cumulative incidence** of death. Recall that the cumulative incidence is the probability that a person at risk at the start of follow-up experiences the event during a pre-specified time interval, and the cumulative incidence can only be meaningfully interpreted with knowledge of the time period to which it applies. (A cumulative incidence of 3% is very different for a five-year period than a five-day period.) Cumulative incidence is sometimes called **incidence proportion**, **risk**, or **attack rate** (though the term rate is best reserved for person-time incidence rates, which this is not). For example, the five-year survival probability is often called the five-year survival rate, though it is not technically a "rate."

We have many statistical tools for the analysis of proportions. We can estimate sample proportions. We can construct confidence intervals using the normal approximation, Wilson, or exact methods. We can use 2x2 tables to compare two independent groups, and we can apply chi-squared tests or Fisher's exact tests to assess differences in proportions. We can use logistic regression to model the effects of continuous covariates or many covariates at once.

Why do we not simply use these binary data methods for time-to-event data?

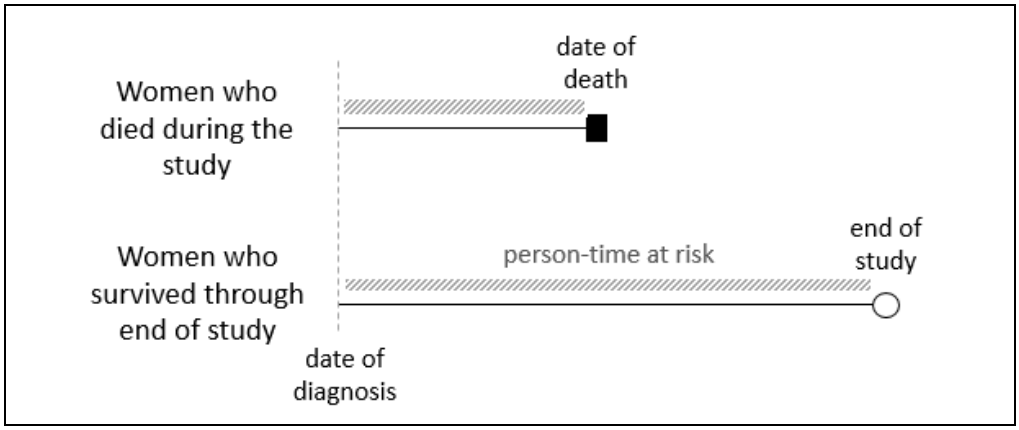
(1) **There may be censoring before the end of the study**. If some women are censored at four years (e.g. move out of the study area), their five-year outcome is not known. We could calculate $\frac{1}{n}\sum_{i=1}^{n} I[T_i^* \leq 5]$, but this is the cumulative incidence of *death or censoring*, which is not of direct scientific interest. Alternatively, we could analyze four-year survival, when our data are complete. But if women are censored even earlier, this quickly becomes problematic.

(2) **We need to specify a single time point for analysis**. By only studying five-year survival, we lose information about how survival compares at other time points. In patients diagnosed with aggressive pancreatic cancer, five-year survival may be similarly poor regardless of treatment. But for two treatments with similar safety profiles, we would prefer a treatment that extends short-term survival (e.g. superior one-year survival). Should we analyze the data using five-year survival, or one-year survival? Ideally, we could capture differences between the two treatments without having to pre-specify a time point. This allows us to summarize the effects of treatments that extend life, even if they are not cures.

Analysis of person-time data

The last analysis approach is of person-time data. A **rate-based analysis** can be used when there are differing lengths of follow-up. We can calculate **incidence rates** using methods that account for person-time at risk. 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**. People do not contribute person-time after the event has occurred because they are no longer at risk for the event.

In the figure below, the gray bars denote the person-time at risk for each woman.



The **incidence rate** is defined as the number of events divided by the total person-time at risk. Person-time could be measured in years, months, days, etc.

Recall that δ_i is our event indicator (=1 if $T_i^* = T_i$). The incidence rate (yearly mortality rate) in our study is defined as follows:

$$\text{incidence rate} = \frac{\sum_{i=1}^{n} \delta_i}{\sum_{i=1}^{n} T_i^*}$$

Unlike a probability, the incidence rate has a lower bound of zero and no upper bound, meaning that the incidence rate can exceed the value of one. This does not mean that more than 100% of persons in a population can experience the event. We must consider the units of time.

An incidence rate of 100 cases per 1 person-year might be expressed as 100 cases/person-year. It might also be expressed as 8.33 cases/person-month, 1.92 cases/person-week, or 0.27 cases/person-day. The value of the incidence rate depends on the time unit selected. *Thus, it is critical to report incidence rates with their time units*.

We have many statistical tools for the analysis of incidence rates. We can estimate sample incidence rates. We can construct normal approximation, log-transformed or exact confidence intervals. We can use a Poisson test to compare two independent groups. We can use Poisson regression to model the effects of continuous covariates or many covariates at once.

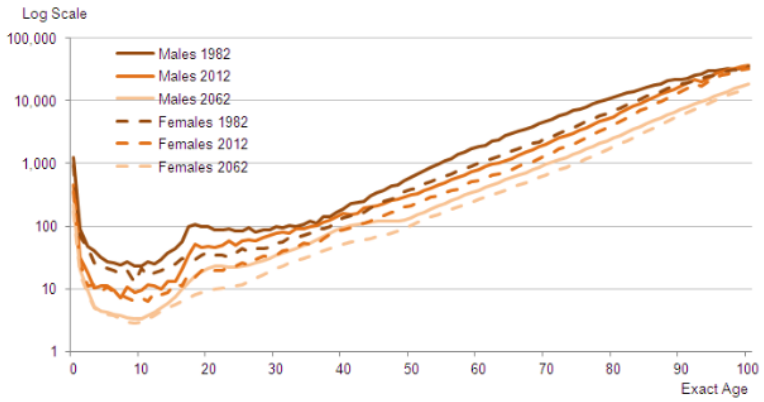
We will see that rate-based analyses are closely related to certain types of simple survival analysis methods, but survival analysis methods in general are much more flexible. *Rate-based analyses generally assume that the event rate is constant over the study period*. The majority of survival analysis methods that we will learn about in this class are flexible in that they allow the event rate to vary over time.

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. Thus, the mortality rate varies over time since diagnosis.

Example: Human mortality rate is known to vary across the age span. Mortality rates are particularly high in the first year of life and then again in old age. Assuming that mortality is the same across the age span would be overly restrictive.

Figure 3: 2012-based Period Mortality Rates (qx), United Kingdom, 1982, 2012, 2062

Principal Projection



Source: Office for National Statistics

Survival methods

Survival analysis methods are special because they are designed to accommodate censored observations, varying lengths of follow-up, and time-varying event rates.

Part 4. Looking ahead

This week we introduced time-to-event data and right censoring. We motivated the need for survival analysis methods by reviewing the limitations of methods for continuous, binary, and person-time data. Next week we will learn about the **survival function** – a key estimand in survival analysis – and methods to estimate it non-parametrically.

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.