Introduction; The nature of time-to-event data

Patrick Breheny

August 25

Survival analysis

- A very common outcome in medical studies is the time until an event occurs:
  - The time until a patient dies
  - The time until a patient suffers a heart attack
  - The time until a liver transplant patient needs a new liver
  - The time until the recurrence of cancer following treatment
- Data involving such an outcome is often called "time-to-event" data or "failure-time data", and the branch of statistics that deals with analyzing these data is called *survival analysis*

A new type of data

As we will see, time-to-event data is a fundamentally different type of data – neither continuous nor categorical – and requires entirely new approaches at each level of statistical analysis:

- New summary statistics
- New methods for plotting/visualizing the data
- New methods for inference
- New methods for modeling

What's wrong with a t-test?

- At first, it might seem that the time until an event occurs is continuous, and that we could use methods for continuous data to analyze time-to-event data
- However, there is a fundamental feature of time-to-event data that prevents any attempt to use such methods: the event doesn't always occur!
- For example, in our hypothetical heart attack study, some patients will never experience a heart attack

It takes time to measure time

- Even if we're studying an event that is certain to occur *eventually*, such as death, it is typically inefficient to have to wait indefinitely until the final event occurs before we can analyze the data
- For example, in our hypothetical liver transplant study, a patient may live for decades following transplantation
- Not only is it impractical to ask researchers to delay publishing their findings so long, it is also unethical in the sense of keeping important medical research hidden for an unnecessarily long time

Couldn't we just throw out the missing data?

- It might seem as though you could fix this problem by throwing out the subjects with missing data
- Nothing could be further from the truth!
- Suppose 20 years have gone by and there's still one individual from our transplant study who is alive; does it make sense to throw that person out and pretend that we know nothing about their survival?
- Of course not; we know a great deal about their survival – they survived for at least 20 years following transplantation

Censoring and partial information

- Thus, what we see isn't "missing" data; it's just incompletely observed; the statistical term for this is that the survival time is *censored*
- Observing a patient to survive for at least 20 years contains quite a lot of information about the distribution of survival time
- On the other hand, if the survival time was censored after a week (e.g., the patient dropped out of the study), this would provide very little information about the distribution of survival time
- Any meaningful analysis of time-to-event data has to take this kind of partial information into account; doing so is what survival analysis is all about

Parametric distribution for post-surgery survival?

- A second reason that survival data analysis tends to differ substantially from analyzing other types of data is that parametric approaches are typically inadequate
- Consider, for example, survival time following some surgical procedure
- To begin, survival times are always positive
- Furthermore, typically the risk of death is high in the days and weeks following surgery, as complications from the surgery itself can lead to death

Parametric distribution for post-surgery survival? (cont'd)

- Risk then declines, but at some point starts rising again due to the simple fact that older individuals are at greater risk of death
- Finally, there are frequently large outliers when it comes to survival
- Coming up with a parametric distribution to describe all this, while not impossible, is certainly not straightforward
- Furthermore, such a parametric approach would necessarily involve a large number of assumptions that would be open to debate

Moments for survival data?

- On a somewhat related note, moment-based statistics such as means and variances are also typically ill-suited to survival data
- Not only are they inadequate for describing unusual distributions like the one we just mentioned, they are often impossible to estimate with incompletely observed data
- For example, consider again our transplant patient who has survived for decades
- Whether she survives until 80, or 90, or 110, certainly impacts the mean and variance

Moments for survival data? (cont'd)

- It does not, however, impact the median
- Indeed, in order to estimate a median, we only need to wait until half of the events occur
- For all of these reasons, nonparametric (or *semiparametric*) approaches tend to more widely used in survival analysis

Survival at 70 vs. survival at 100

- Finally, another way in which survival analysis is unique is that it is typically most natural to think about survival times conditionally
- For example, presumably we would all agree that a 100-year old person is at greater risk of death (in, say, the next year) than a 70-year old person
- However, 70 is a much more common age of death than 100
- To put it another way, the probability density at 70 is higher than the probability density at 100

Survival at 70 vs. survival at 100 (cont'd)

- For many other types of data, we're used to thinking in terms of distributions and densities
- The probability distribution may lead us to remark that a person is much more likely to die at 70 than at 100
- While technically true, this is potentially quite misleading

The hazard function

- Instead, it is typically most natural to work with survival data on a conditional level: what is the risk of death, given that an individual has survived up until a certain age?
- This concept is known as the hazard function – we will define it more formally in the next lecture
- Thus, it is common in survival analysis to estimate, analyze, and model things on the level of the hazard function, rather than the distribution or density function directly

Life expectancy: Expectation

- A good example of this is the widely misunderstood concept of life expectancy
- For example, in England in 1841 (the oldest date for which systematically collected population survival data was collected) the life expectancy at birth was 40 years old
- A common misconception is that this implies that most people lived until about 40, then died (as you might expect if, say, survival times were normally distributed),
  as in the following ad from the British Pharmaceutical Industry:

  > *They say life begins at 40. Not so long ago, that's about when it ended.*

Life expectancy: Conditional expectation

- However, it was never true that 40 was a common age of death in England
- The average age at death may have been 40, but this is only because death in infancy was very common in the 1800s
- Indeed, if you survived until your third birthday in 1841, you could then expect to live another 50 years, until the age of 53
- If you survived until the age of 40, you could expect to survive another *27 years*, until the age of 67

Life expectancy woes

Failing to appreciate these subtle distinctions leads to some hilariously wrong statements, like

> *Middle age is a modern phenomenon – a hundred years ago, life expectancy was 47.*

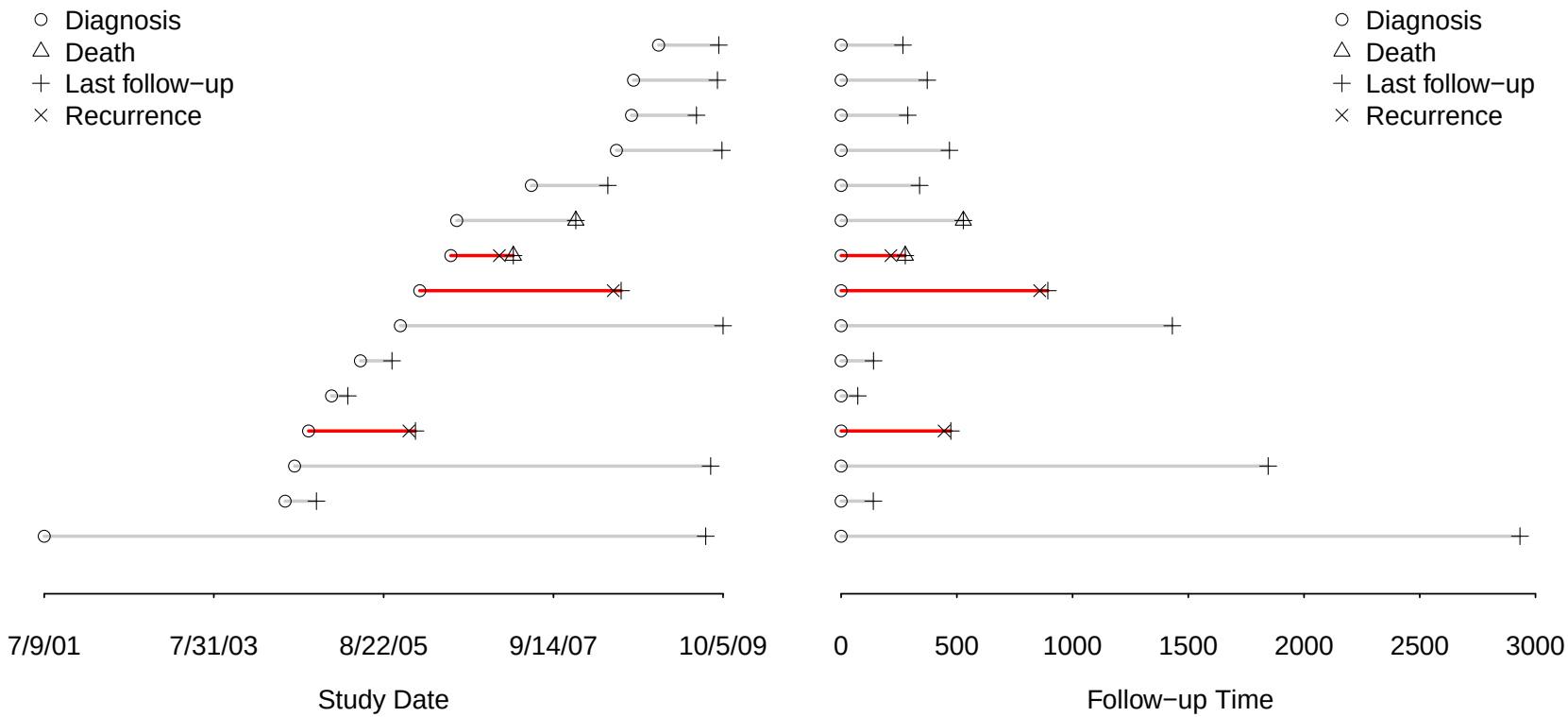
and

> *Mothers have always provoked rage and resentment in their adult daughters. . . In past centuries, daughters could bury their rage. . . while they cared for their mothers who, turning 40, rapidly aged, grew frail and died. Now mothers turning 40 are strong and healthy, and only half way through their lives.*

both of which appeared in *The Guardian* (different authors, different dates)

Raw and transformed survival data

The following is a chart of the "raw" data for 15 subjects from a study of survival in patients with endometrial carcinoma



Examples of censoring

We can see from the chart that censoring is pretty common; it can be caused by many things:

- The end of the study
- Patient moves and the investigators lose contact with them
- Patient drops out of the study
- Death: for example, if we're studying cancer recurrence, the patient may die before we get to see when their cancer would have come back

General notation

We will adopt the following general notation in this course:

- Let T denote the time from some specified origin (e.g., birth, time of surgery, date of diagnosis) until the event we are studying
- We do not always observe T, however; instead, we observe $\{t_i, d_i, x_i\}_{i=1}^n$, where
  - t_i: The follow-up time
  - d_i: An indicator for whether the event was observed (i.e., if the follow-up time is equal to the failure time; $d_i = 1$ if we observe the event and $d_i = 0$ if the event was censored)
  - x_i: Explanatory variables (e.g., an indicator for which treatment group the subject was in, or a long vector of covariates for a regression model)

Conclusion

- Previously, we remarked that hazard functions are important for understanding and modeling survival distributions
- Tomorrow, we will define and begin to work with the hazard function for T, and see how it relates to the distribution function, density function, and various other properties of the survival time distribution