Introduction and Background

Dave Harrington

May 14 - 18, 2018

Administrative Information

Instructor coordinates

Dave Harrington

- Department of Biostatistics, Harvard T.H. Chan School of Public Health
- Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute
- Email: dharrington846@gmail.com

Course format

- Lectures, but please ask questions
- Labs (practical exercises)
- Readings (typically, papers from literature)
- Computing (in R)

Acknowledgements

Many individuals have contributed to these notes in courses at Harvard, Ghent University, Hasselt University, and other institutions.

Special acknowledgements to

- Rui Wang, Paige Williams, Rebecca Betensky, and Paul Catalano (Harvard)
- Louise Ryan (University of Technology, Sydney)
- Els Goetghebeur (U Ghent)
- Julie Vu (University College London)

Important Definitions

Main ideas

Survival analysis typically focuses on *time-to-event* data.

Examples:

- Time to death from a chronic disease
- Time to progression of a disease
- Time to onset (or relapse) of a disease
- Length of stay in a hospital or a nursing home

The terms *survival time* and *event time* will be used to mean the time to an event.

Most useful references

- Klein and Moeschberger: *Survival Analysis: Techniques for censored and truncated data*
- Therneau and Grambsch: *Modeling Survival Data: extending the Cox Model (R)*
- Collett: *Modelling Survival Data in Medical Research*
- Hosmer, Lemeshow, and May: *Applied Survival Analysis, 2nd ed.*
- Kleinbaum: *Survival Analysis: A self-learning text*
- Cox and Oakes: *Analysis of Survival Data*

I recommend the K&M text, supplemented by T&G.

A few more references

- Fleming and Harrington: *Counting Processes and Survival Analysis*
- Kalbfleisch and Prentice: *The Statistical Analysis of Failure Time Data*
- Allison: *Survival Analysis Using the SAS System*
- Miller: *Survival Analysis*

Packages used

```
install.packages("devtools")
library(devtools)
install.packages(c("survival", "KMSurv",
                   "gsDesign", "Hmisc"))
install_github("keaven/nphsim",
               "dave-harrington/eventtimedata")
```

Example: Time to death or hospitalization

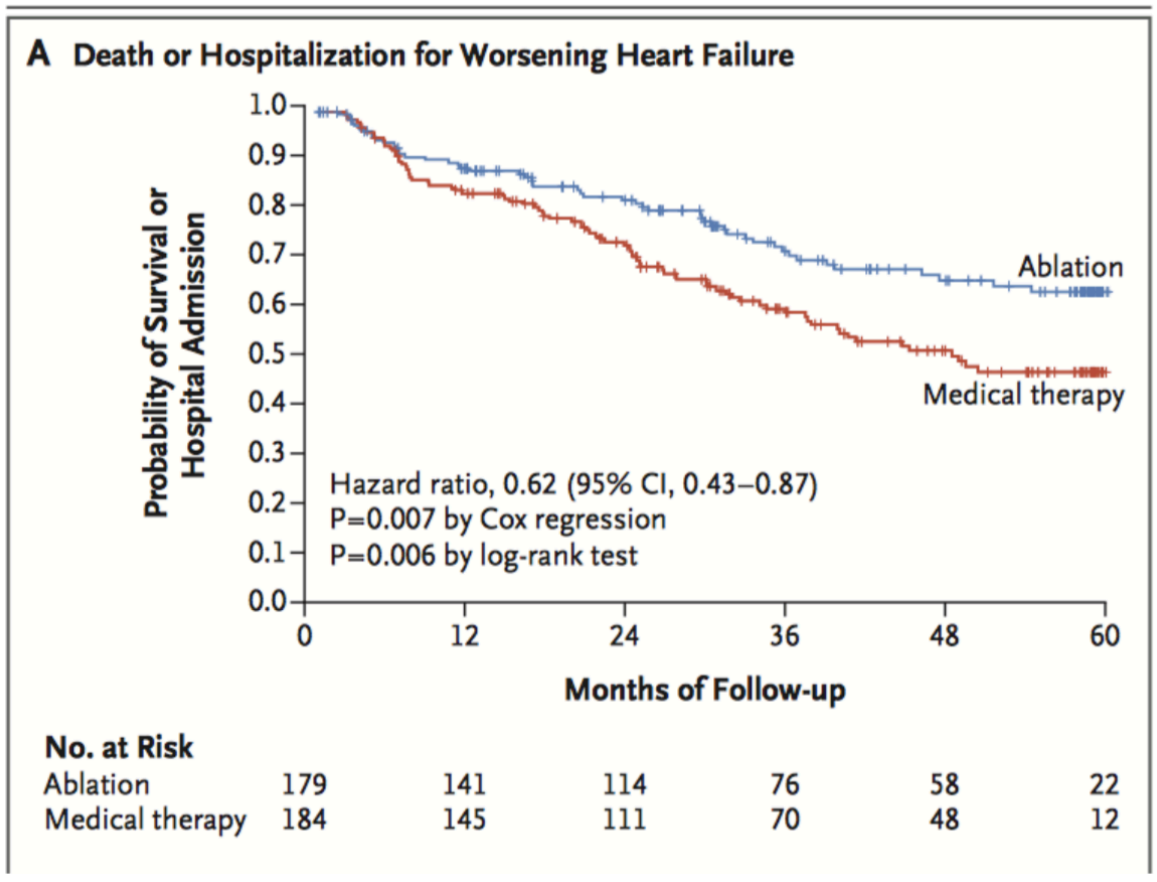

Figure 1: Figure from Marrouche, et al., *NEJM* 2018

See Marrouche, et al.

Structure of event time data

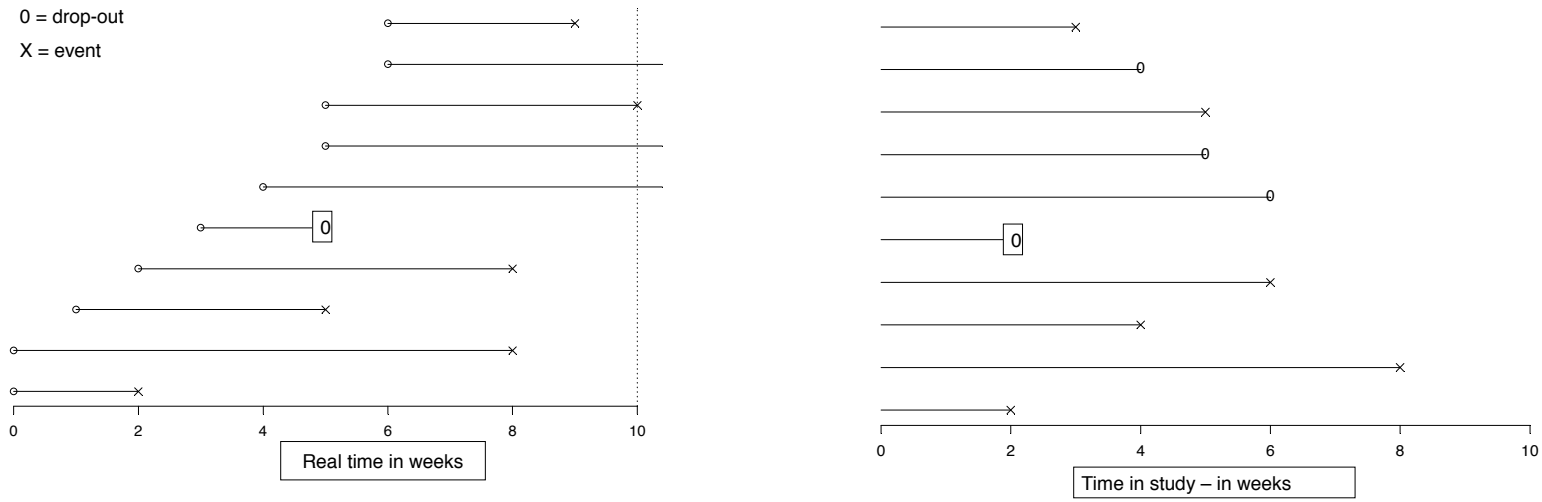

Figure 2: Event time data as observed (L) versus to a data analyst (R)

Characteristics of event time data

- 'Individuals' do not all enter the study at the same time.
  - This is referred to as *staggered entry*.
- When the study ends, some individuals still haven't had the event.
- Other individuals drop out or are lost during the study.
  - The last time they were still "free" of the event is all that is known.

The last two features relate to *censoring* of the failure times.

The first of the times until the study ends or the subject drops out is called a *censoring time*.

A hypothetical example

Assume 10 subjects assigned to each of four treatments after cancer remission, followed until death or end of study at 36 months.

The times to death (months):

Trt 1: Deaths at 2, 3, 7, 9, 15, 16 (4 alive at end)

Trt 2: Deaths at 1, 1, 2, 4, 4, 6, 7, 9, 11 (1 alive at end)

Trt 3: Deaths at 1, 1, 2, 4, 4, 5 (4 alive at end)

Trt 4: Deaths at 2, 3, 7, 9, 15, 22, 27, 28, 29 (1 alive at end)

- For Treatment 1, what is the average time to death?
- Comparing Treatments 1 and 2, which appears better?
- Comparing Treatments 1 and 3, which appears better?
- Comparing Treatments 3 and 4, which appears better?

Descriptive comparisons of "average" death time

Treatment Group	Among Deaths: Mean	Among Deaths: Median	Median adjusting for Censoring (KM)
1	8.67	8.00	15.5
2	5.00	4.00	5.0
3	2.83	3.00	4.5
4	15.78	15.00	18.5

Comparisons of treatments (p-values)

Methods to do these calculations coming in this course.

Comparison	Log-Rank Test	Wilcoxon Test	Exponential Model
1 vs 2	0.045	0.048	0.014
3 vs 4	0.62	0.67	0.56
1 vs 3	0.63	0.37	0.73
2 vs 4	0.15	0.06	0.09
1 vs 4	0.37	0.68	0.34

More key features of survival data

- Survival times are often *right-skewed*, so the median is usually a better measure of center than the mean.
- The median can be estimated from data that include censored observations (not always possible with the mean).
- Calculating summary statistics and comparing survival distributions must account for cases without events.
- Comparisons between survival distributions may yield different conclusions depending on assumptions.

Types of censoring

Right-censoring

Let T_i be the time to event and U_i be the time to censoring for an individual i. Only the r.v. $X_i = \min(T_i, U_i)$ is observed, due to

- loss to follow-up
- drop-out
- study termination ('administrative censoring')

Called *right*-censoring because the true unobserved event is to the right of the censoring time (i.e., after the censoring time).

In addition to X_i, the *failure indicator* δ_i is observable:

$$\delta_i = \begin{cases} 1 & \text{if} \quad T_i \leq U_i \\ 0 & \text{if} \quad T_i > U_i \end{cases}$$

Types of censoring ...

Left-censoring

The r.v.'s observed are $Y_i = \max(T_i, U_i)$ and the failure indicator ϵ_i:

$$\epsilon_i = \begin{cases} 1 & \text{if} \quad U_i \leq T_i \\ 0 & \text{if} \quad U_i > T_i \end{cases}$$

Ex.: study of age at which African children learn a task (Miller).

- Some already knew the task (left-censoring)
- Some learned during study (exact)
- Some had not yet learned by end of study (right-censoring)

Types of censoring...

Interval-censoring

Observe (L_i, R_i) where $T_i \in (L_i, R_i)$

Examples:

- Time to prostate cancer, observe longitudinal PSA measurements
- Time to undetectable viral load in AIDS studies, based on measurements of viral load taken at each clinic visit
- Detect recurrence of colon cancer after surgery. Follow patients every 3 months after resection of primary tumor.

This short course is restricted to right-censoring.

Independent versus informative censoring

Censoring is **independent** if U_i is independent of T_i.

Examples:

- If U_i is the planned end of the study (say, 2 years after the study opens), then it is usually independent of the event times.
  - What if there is a trend over calendar time in the survival times?
- If U_i is the time that a patient drops out of the study because they have become much sicker and/or had to discontinue taking the study treatment, then U_i and T_i are probably not independent.

An individual censored at U should be *representative* of all subjects who survive to U.

Censoring is considered **informative** if the distribution of U_i contains information about the parameters characterizing the distribution of T_i.

Survival Distributions

Some mathematical definitions

There are several equivalent ways to characterize the probability distribution of a survival random variable.

- The density function $f(t)$
- The survivor function $S(t)$
- The hazard function $\lambda(t)$
- The cumulative hazard function $\Lambda(t)$

Some are special to survival analysis.

Density function

For a *discrete* random variable

Suppose that T takes values in $a_1, a_2, \ldots, a_J$.

$$
\begin{aligned}
f(t) &= P(T = t) \\
&= \begin{cases} f_j & \text{if} \quad t = a_j, j = 1, 2, \ldots, J \\ 0 & \text{if} \quad t \neq a_j, j = 1, 2, \ldots, J \end{cases}
\end{aligned}
$$

For a *continuous* random variable

$$
f(t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} P(t \leq T < t + \Delta t)
$$

Survivorship function $S(t)$

In other settings, the cumulative distribution function

$$F(t) = P(T \leq t) = 1 - S(t)$$

is of interest.

In survival analysis, interest tends to focus on $S(t) = P(T > t)$, the *survival* or *survivorship* function.[1]

The survival function measures the probability an individual will experience the event beyond time t.

[1] Be careful: some books use the definition $S(t) = P(T \geq t)$.

Survivorship function $S(t)$ ...

For a *continuous* random variable:

$$S(t) = \int_t^\infty f(u)du$$

For a *discrete* random variable:

$$S(t) = \sum_{u>t} f(u) = \sum_{a_j>t} f(a_j) = \sum_{a_j>t} f_j$$

Hazard function $\lambda(t)$

The hazard function measures the probability of death at time t, conditional on having survived until that time.

- This is sometimes called the *instantaneous failure rate*.

For a *continuous* random variable T:

$$\begin{aligned}
\lambda(t) &= \lim_{\Delta t \to 0} \frac{1}{\Delta t} P(t \leq T < t + \Delta t | T \geq t) \\
&= \lim_{\Delta t \to 0} \frac{1}{\Delta t} \frac{P([t \leq T < t + \Delta t] \cap [T \geq t])}{P(T \geq t)} \\
&= \lim_{\Delta t \to 0} \frac{1}{\Delta t} \frac{P(t \leq T < t + \Delta t)}{P(T \geq t)} \\
&= \frac{f(t)}{S(t)}
\end{aligned}$$

Hazard function $\lambda(t)\ldots$

For a *discrete* random variable T:

$$\lambda(a_j) \equiv \lambda_j = P(T = a_j | T \geq a_j) = \frac{P(T = a_j)}{P(T \geq a_j)}$$

$$= \frac{f(a_j)}{S(a_j-)} = \frac{f(t)}{\sum_{k:a_k \geq a_j} f(a_k)}$$

The form of the denominator for both continuous and discrete variables is the reason some books use $P(T \geq t)$ as the definition of the survivor function.

Cumulative hazard function $\Lambda(t)$

For a *continuous* random variable T:

$$\Lambda(t) = \int_0^t \lambda(u)du$$

For a *discrete* random variable T:

$$\Lambda(t) = \sum_{k:a_k \leq t} \lambda_k$$

Relationship between $S(t)$ and $\lambda(t)$

For a continuous random variable,

$$\lambda(t) = \frac{f(t)}{S(t)}.$$

For a right-continuous survivor function $S(t)$,

$$f(t) = -S'(t) \text{ or } S'(t) = -f(t).$$

These relationships can be used to show another way to write $\lambda(t)$:

$$\lambda(t) = \frac{f(t)}{S(t)} = -\frac{-f(t)}{S(t)} = -\left(\frac{1}{S(t)}\right) S'(t) = -\frac{d}{dt}[\log S(t)]$$

$$\lambda(t) = -\frac{d}{dt}[\log S(t)]$$

Relationship between $S(t)$ and $\Lambda(t)$

For a *continuous* random variable:

$$\begin{aligned}\Lambda(t) &= \int_0^t \lambda(u)du \\ &= \int_0^t -\frac{d}{du}\log S(u)du \\ &= -\log S(t) + \log S(0)\end{aligned}$$

Thus, $S(t) = e^{-\Lambda(t)}$.

Relationship between $S(t)$ and $\Lambda(t)\ldots$

For a *discrete* random variable:

Suppose that $a_j \leq t < a_{j+1}$. Then

$$\begin{aligned} S(t) = P(T > t) &= P(T > a_1, T > a_2, \ldots, T > a_j) \\ &= P(T > a_1) \times P(T > a_2 | T > a_1) \times \cdots \\ &\quad \cdots \times P(T > a_j | T > a_{j-1}) \\ &= (1 - \lambda_1) \times \cdots \times (1 - \lambda_j) \\ &= \prod_{k: a_k \leq t} (1 - \lambda_k) \end{aligned}$$

Relationships: an overview

$$f(t)\Delta t \approx P(t \leq T < t + \Delta t)$$

$$\lambda(t)\Delta t \approx P(t \leq T < t + \Delta t | T \geq t)$$

$$S(t) = P(T > t) = \int_t^\infty f(u)du$$

$$f(t) = -\frac{d}{dt}S(t)$$

$$\lambda(t) = \frac{f(t)}{S(t)}$$

$$\lambda(t) = -\frac{d}{dt}[\log S(t)]$$

$$S(t) = e^{-\Lambda(t)}$$

$S(t) \approx 1 - \Lambda(t)$ while the cumulative hazard is small.

Some parametric survival distributions

The *exponential distribution*

- Simplest distribution, only one unknown parameter
- Plays a role similar to that of the normal distribution in linear regression

$$f(t) = \lambda e^{-\lambda t} \text{ for } t \geq 0$$

$$S(t) = \int_t^\infty f(u)du = e^{-\lambda t}$$

$$\lambda(t) = \frac{f(t)}{S(t)} = \lambda \qquad \text{constant hazard}$$

$$\Lambda(t) = \int_0^t \lambda(u)\, du = \int_0^t \lambda\, du = \lambda t$$

Some parametric survival distributions...

The *Weibull distribution* generalizes the exponential, and has two parameters

- λ: the *scale* parameter
- γ: the *shape* parameter

$$S(t) = e^{-\lambda t^\gamma}$$
$$f(t) = \frac{-d}{dt} S(t) = \gamma \lambda t^{\gamma-1} e^{-\lambda t^\gamma}$$
$$\lambda(t) = \gamma \lambda t^{\gamma-1}$$
$$\Lambda(t) = \int_0^t \lambda(u) du = \lambda t^\gamma$$

The Weibull distribution...

The Weibull distribution is convenient because of simple forms. It includes several hazard shapes:

- $\gamma = 1 \rightarrow$ constant hazard
- $0 < \gamma < 1 \rightarrow$ decreasing hazard
- $\gamma > 1 \rightarrow$ increasing hazard