INTRODUCTION TO CLINICAL RESEARCH

Survival Analysis – Getting Started

Gayane Yenokyan, MD, MPH, PhD
Associate Director, Biostatistics Center
Department of Biostatistics, JHSPH

Topics to be covered today

• Motivation: why do we need to learn about this method?
• Unique nature of “survival” data
• Non-parametric approach: Kaplan-Meier Estimates
• Comparing 2 survival curves: log-rank test
• Semi-parametric method: Cox proportional hazards model
• Parametric approach: WILL NOT BE COVERED

Terminology

• First applications in engineering
  – Time until failure of certain machine or equipment
• Term “survival analysis” because most of earlier applications were time to death
• Alternative names: “time-to-event”, “event history analysis” (sociology), “reliability analysis” or “failure time analysis” (engineering), transition analysis (economics)
Motivation I

• Interested is not just in probability of the event, but the time to certain event
• Event – qualitative change that can be localized in time (death, recurrence, marriage, promotion, war etc.)
• There are 2 important considerations:
  – Times to event are not normal
  – Times to event for some patients are unobserved

Motivation II

• If patients do not develop the event (did not fail), they are said to censored

• Not good approaches:
  – Discarding censored cases
  – Assigning all censored cases the end of the follow-up time

Examples of Censoring

• Uncensored data: The event has occurred
• Censored data: The event has yet to occur
  – Event-free at the current follow-up time
  – A competing event that is not an endpoint stops follow-up
  – Death (if not part of the endpoint)
  – Clinical event that requires treatment, etc.
Types of Censoring

- Left vs. Right censoring
  If $T$ = time to event,
  - Right censored observations: $T > c$
  - Left censored observations: $T < c$
  - Interval censored observations: $c_1 < T < c_2$

- Type I censoring
- Type II censoring
- Random censoring

Survival Data (Type I censoring)

Survival Data (Random censoring)

From: Allison PD. Event History Analysis: Regression for Longitudinal Event Data (Quantitative Applications in the Social Sciences). SAGE University Press
Survival Analysis Methods

Survival Analysis methods need to accommodate:
- Censoring
- Time-dependent predictors (more later)
The Problem with Standard Analyses of Times to Events

- Mean time to event: \( (1 + 3 + 5 + 6 + 8)/5 = 4.6 \)
  Is this right?
- Median: 5
- Histogram

Censoring

> 3 is not 3, it may be 33

Mean is not 4.6, it may be \( (1 + 33 + 5 + 6 + 8)/5 = 10.6 \)

Or any value greater than 4.6

> 3 is a right "censored value" – we only know the value exceeds 3

> x is often written "x+"

Overall Event Rate

Event Rate = \( \frac{\text{# of events}}{\text{Total time at risk}} \)

- Gives an average event rate over the follow-up period; actual event rate may vary over time
- For a finer time resolution, do the above for smaller intervals
**Interpretation of the Overall Event Rate**

- The numeric value depends on the time frame
- Example: 2 events in 23 person-months
  - $1$ event per 11.5 months =
  - $0.09$ event per 1 month =
  - $1.04$ events per year =
  - $104$ events per 100 person-years

**Second Option: Natural history**

“One day at a time”
Thinking about Times to Events

<table>
<thead>
<tr>
<th>Event Times</th>
<th>0-1</th>
<th>1-2</th>
<th>2-3</th>
<th>3-4</th>
<th>4-5</th>
<th>5-6</th>
<th>6-7</th>
<th>7-8</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3+</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>6+</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8+</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>No. at risk</th>
<th>5</th>
<th>4</th>
<th>4</th>
<th>3</th>
<th>3</th>
<th>2</th>
<th>1</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fraction of events=&quot;hazard&quot;</td>
<td>0.2</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>33</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
</tbody>
</table>

Survival Function

“Survival function”, \( S(t) \), is defined to be the probability a person survives beyond time \( t \)

\[ S(0) = 1.0 \]

\[ S(t+1) \leq S(t) \]

Hazard Function

- Hazard at time \( t \), \( h(t) \), is the probability per unit time of having the event in a small interval around time \( t \)
- Hazard = Force of mortality
- \( \sim \frac{\Pr\{\text{event in } (t,t+dt)\}}{dt} \)
- Need not be between 0 and 1 because it is per unit time
- \( h(t) \sim \frac{(S(t)-S(t+dt))/(S(t) \ dt)}{dt} \)
Hazard Function

- Basic idea: Live your life one interval (day, month, or year) at a time
- Example: \( S(3) = \Pr(\text{survive for 3 months}) = \Pr(\text{survive 1st month}) \times \Pr(\text{survive 2nd month} | \text{survive 1st month}) \times \Pr(\text{survive 3rd month} | \text{survive 2nd month}) \)
- Thus, \( S(3) = \frac{S(1)}{S(2)} \frac{S(2)}{S(3)} = S(1) S(2) \)

Estimating the Survival Function: Kaplan-Meier Method

<table>
<thead>
<tr>
<th>Interval of Follow-Up</th>
<th>0-1</th>
<th>1-2</th>
<th>2-3</th>
<th>3-4</th>
<th>4-5</th>
<th>5-6</th>
<th>6-7</th>
<th>7-8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Times</td>
<td>9</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>No. at risk</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>No. of events</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Fraction of events=&quot;hazard&quot;</td>
<td>0.2</td>
<td>0.0</td>
<td>0.0</td>
<td>0.33</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Fraction without event in interval</td>
<td>0.8</td>
<td>1.0</td>
<td>1.0</td>
<td>0.67</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>Fraction without event since start</td>
<td>1.0</td>
<td>0.8</td>
<td>0.8</td>
<td>0.8</td>
<td>0.53</td>
<td>0.53</td>
<td>0.53</td>
<td>0.53</td>
</tr>
</tbody>
</table>

\[ \Pr(\text{survive past 5 and survive past 4}) = \Pr(\text{survive past 5}) \times \Pr(\text{survive past 4}) \]

Displaying the Survival Function

![Graph of survival function](image)
Notes on Estimating Survival Function

Estimates of Survival Function (Kaplan-Meier estimates)
- This is a non-parametric method
- Estimate only changes in intervals where an event occurs
- Censored observations contribute to denominators, but never to numerators
- Intervals are arbitrary; want narrow ones

Assumptions of Kaplan-Meier Method

- Observations are independent (patients represent a random sample from a population)
- Everybody has the same survival curve (homogenous group)
- Non-informative censoring (independent censoring): probability of censoring is not related to the probability of event
  - In other words: those who are censored are not more or less likely to develop the event than those who “stay” in the study

Kaplan-Meier Example

- DATA (5 patients): 25, 18+, 17, 22, 27
- These data can be represented as pair of variables: y and c, where y = duration, and c = censoring indicator: c = 1 – censored and 0 – failed (developed the event of interest)

<table>
<thead>
<tr>
<th>Patient</th>
<th>Y</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>25</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>18</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>17</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>22</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>27</td>
<td>0</td>
</tr>
</tbody>
</table>
Kaplan-Meier survival estimate

Y (duration) 25 18 17 22 27
C (censored) 0 1 0 0 0

Kaplan-Meier Survival Estimates

<table>
<thead>
<tr>
<th>Time</th>
<th>Total</th>
<th>Fail</th>
<th>Lost</th>
<th>Survivor</th>
<th>Std. Error</th>
<th>Std. Conf. Int.</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>5</td>
<td>-</td>
<td>-</td>
<td>1.0000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>17</td>
<td>5</td>
<td>1</td>
<td>0</td>
<td>0.8000</td>
<td>0.1789</td>
<td>0.2038</td>
</tr>
<tr>
<td>22</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>0.5333</td>
<td>0.2483</td>
<td>0.0683</td>
</tr>
<tr>
<td>25</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0.2667</td>
<td>0.2257</td>
<td>0.0097</td>
</tr>
<tr>
<td>27</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0.0000</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

\[
S(17) = 1 - \frac{d}{n} = 1 - \frac{1}{5} = 0.8
\]

K-M Estimates for A Clinical Trial
Comparing Survival Functions

• Suppose we want to test the hypothesis that two survival curves, \( S_1(t) \) and \( S_2(t) \) are the same
  – Null hypothesis is that two survival curves are the same
• Common approach is the “log-rank” (chi-square) test
• It is effective when we can assume the hazard rates in the two groups are roughly proportional over time

Logrank test: “Drug trial” data

![Graph showing Kaplan-Meier survival estimates by drug with Logrank chi-square statistic: 1.72, p-value: < 0.19.]

Conclusion - ?

More Advanced Approach

• Log-rank test allows to incorporate one categorical predictor
• What if we want to include more predictors?
• Regression analysis: Cox proportional hazards model
• Also allows to include time-dependent predictors
**Cox Proportional Hazards Model**

- Proposed in 1972 by Sir David Cox
- Not just accommodates censoring (like KM method)
- Is a Regression Method:
  - Adjusts for multiple covariates
  - Including covariates that change with time (time-dependent covariates)

**Time-dependent predictors**

- Predictors that change the value over time
- Examples:
  - Marital status
  - Serum albumin value
  - Blood pressure (compare to, baseline blood pressure)
- Why these predictors are important?
  - They change the risk of event over time

**Example: Heart Transplant Data**

- Famous Stanford Heart transplant data (Crowley and Hu, 1977)
- 103 cardiac patients enrolled in heart transplantation program between 1967 and 1974
- Patients were followed from the date of acceptance into the program to either death or end of the follow-up (April 1974)
- Data on first 10 patients on the next slide...
Heart Transplant DATA

Variables:
- DOB – date of birth
- DOA – date of admission
- DOT – date of transplant
- DLS – date last seen (dead or censored)
- Dead – vital status: 1 – died, 0 – alive
- Surg – prior open heart surgery
- M1, M2, and M3 – mismatch scores

Research Question(s)
- Does cardiac transplant improve survival?
- What other factors (age at transplant, mismatch score) predict survival?

Proportional Hazards Assumption

The model is:
\[ \log[h(t)] = \log[h_0(t)] + (\beta_1 X_1 + ... + \beta_k X_k) \]

BASELINE HAZARD – is left unspecified

MODEL: \( \beta_k \) – change in log(hazard) for 1 unit difference in \( X \)
\( \exp(\beta_k) \) – hazard ratio (HR) of outcome per 1 unit difference in \( X \)

Proportional hazards assumption means that the ratio of hazards for each binary predictor or for 1 unit difference in non-binary predictor stays the same over time.
Proportional Hazards Assumption

From: Allison PD. Event History Analysis - Regression for Longitudinal Event Data (Quantitative Applications in the Social Sciences). SAGE University Press

(Naïve) Analysis

Analysis of Maximum Likelihood Estimates

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Std. Error</th>
<th>Chi-Square</th>
<th>Pr &gt; Chi-Square</th>
<th>Std. Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time</td>
<td>-1.769240</td>
<td>0.2860</td>
<td>37.5904</td>
<td>0.0001</td>
</tr>
<tr>
<td>Odds</td>
<td>-0.421480</td>
<td>0.2720</td>
<td>1.2903</td>
<td>0.2560</td>
</tr>
<tr>
<td>ABC/CD</td>
<td>0.058697</td>
<td>0.1515</td>
<td>19.1631</td>
<td>0.0001</td>
</tr>
</tbody>
</table>

Hazard Ratio of death comparing those who received transplant vs. those who did not after adjusting for surgery prior to the program and age at transplant. The estimated benefit of 82% in those who received transplant vs. those who did not.

Is this effect real?

Time-Dependent Covariates

Complication: some people die before receiving transplant

If these people are included in the non-transplant group, this group will have more deaths than the transplant group, and the transplant group will look too good (Hazard Ratio will overestimate the truth)
Possible Solutions

• Better way: restrict the analysis to those who received transplant. Ask a question what determines survival among patients who received it?
• Downside: Cannot look at the effect of transplant.
• Solution: Treat transplant status as time-dependent covariate

New Approach

Re-write the model as:

\[ \log[h(t)] = \log[h_0(t)] + (\beta_1 X_1 + \ldots + \beta_k X_k(t)) \]

Divide each person’s time into prior to transplant and after transplant
  – “Prior to transplant” time is counted as “not on transplant”
  – “After transplant” time is counted as “on transplant”

New Results

<table>
<thead>
<tr>
<th>Variable</th>
<th>Parametric Estimate</th>
<th>Standard Error</th>
<th>Wald Chi-Square</th>
<th>Pr &gt; Ch</th>
<th>Risk Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>PLANT</td>
<td>-0.046152</td>
<td>0.20276</td>
<td>0.03224</td>
<td>0.8595</td>
<td>0.965</td>
</tr>
<tr>
<td>SURG</td>
<td>-0.775456</td>
<td>0.35961</td>
<td>4.60216</td>
<td>0.0339</td>
<td>0.462</td>
</tr>
<tr>
<td>AGENDOCY</td>
<td>0.032086</td>
<td>0.01391</td>
<td>6.99524</td>
<td>0.0091</td>
<td>1.052</td>
</tr>
</tbody>
</table>

Hazard Ratio of death comparing those who received transplant vs. those who did not after adjusting for surgery prior to the program and age at transplant. The estimated benefit of 5% in those who received transplant vs. those who did not (and it is not significant).
Main Points Once Again I

• Time to event data can be censored because every person does not necessarily have the event during the study

• Think of time to event as a natural history, that is 0 before the event and then switches to 1 when the event occurs; analysis counts the events

• Survival function, S(t), is the probability a person’s event occurs after each time t

Main Points Once Again II

• Kaplan-Meier (non-parametric) estimator of the survival function is a product of interval-specific survival probabilities

• Hazard function, h(t), is the risk per unit time of having the event for a person who is at risk (not previously had event)

• Logrank tests evaluate differences among survival in population subgroups

• Cox model used for regression for survival data

References


3. Allison PD. Event History Analysis: Regression for Longitudinal Event Data (Quantitative Applications in the Social Sciences). SAGE University Press