2024 Summer Institute In Statistics for Clinical & Epidemiological Research

Module 3:

Design, Conduct, and Analysis of Randomized Clinical Trials with Time to Event Primary Endpoints

Lecture 9:

Estimation of Survival Curves

Scott S. Emerson, M.D., Ph.D.
Professor Emeritus of Biostatistics
University of Washington

1

Noninformative Censoring

- When estimating survivor functions using censored data:
  - Censoring must not be informative
    - Censored subjects neither more nor less likely to have an event in the immediate future
  - Censored individuals must be a random sample of those at risk at time of censoring: MAR
    - Missingness depends on time last observed
    - But random among all subjects at that time
  - Later: a random sample from all subjects at risk having similar modeled covariates: MAR
    - Missingness depends on time last observed and some other measured and modeled covariates

2

Informative Censoring Examples

- Subjects in a RCT are withdrawn due to treatment failure
  - (likely they would die sooner than those remaining)
- Subjects in a RCT in a fatal condition are lost to follow up when they go on vacation
  - (likely they are healthier than those remaining)
- Leukemia patients in a RCT of bone marrow transplantation are censored if they die of infections rather than dying of cancer
  - (they might have had a more effective regimen to wipe out existing cancer)

3

Detecting Informative Censoring

- As a general rule it is impossible to use the data to detect informative censoring
- The necessary data is almost certainly missing in the data set
- In some cases, it is impossible to ever observe the missing data: "Competing Risks"
  - Nonfelines can only die once
  - We cannot observe whether subjects dying of one cause are more or less likely to die of another if we cure them of the first cause

4

Life Table Methods

- In the actuarial (e.g., insurance) setting
  - The time intervals are often chosen by years, decades, etc.
  - The data are presented for each year as
    - N_j: Number of subjects at risk at start of interval
    - C_j: Number censored during interval (these will contribute half a person)
    - D_j: Number of events in interval

5

Life Table Methods: Notation

- Number at risk, censored, failed in each interval

$$\begin{aligned} &\text{Time interval:} && (t_{j-1}, t_j] \\ &\text{Number at risk:} && N_j \\ &\text{Number censored:} && C_j \\ &\text{Number of events:} && D_j \end{aligned}$$

6

Life Table Methods: Formula

- Computation of probability of survival

Conditional probability of survival in interval:

$$\Pr\left(T^0 \geq t_j \mid T^0 \geq t_{j-1}\right) = 1 - \frac{D_j}{N_j - 0.5 \times C_j}$$

Cumulative probability of survival:

$$\Pr\left(T^0 \geq t_j\right) = \Pr\left(T^0 \geq t_j \mid T^0 \geq t_{j-1}\right)\Pr\left(T^0 \geq t_{j-1}\right)$$

7

Kaplan-Meier Estimates

- Kaplan-Meier (Product Limit) Estimates

- With more precisely measured individual data
  - The time intervals are defined by unique observation times
  - The data are presented for each year as
    - N_j: Number of subjects at risk at start of interval
    - D_j: Number of events at end of interval
    - (Note no censoring or events during interval by definition)
    - (Note also that for ties, censoring occurs after deaths)

8

Kaplan-Meier Notation

- Definition of intervals, number at risk, failures

Ordered distinct observation times:

$$t_1 \le t_2 \le \cdots \le t_k$$

Time interval: $(t_{j-1}, t_j]$

Number at risk at t_j: N_j

Number of events at t_j: D_j

9

Kaplan-Meier Hazard Estimates

- Computation of hazard and conditional probability of survival in interval

Hazard for event in interval: $\frac{D_j}{N_j}$

Conditional probability of survival in interval:

$$\Pr\left(T^0 \ge t_j \mid T^0 \ge t_{j-1}\right) = 1 - \frac{D_j}{N_j}$$

10

Kaplan-Meier Survival Estimate

- Estimating survival probability

$S(t) = Pr(T^0 > t)$

Cumulative probability of survival:

$$\Pr\left(T^0 > t_j\right) = \Pr\left(T^0 > t_j \mid T^0 > t_{j-1}\right)\Pr\left(T^0 > t_{j-1}\right)$$

$$\hat{S}\left(t_j\right) = \left(1 - \frac{D_j}{N_j}\right) \times \left(1 - \frac{D_{j-1}}{N_{j-1}}\right) \times \cdots \times \left(1 - \frac{D_1}{N_1}\right)$$

$$= \prod_{i=1}^{j}\left(1 - \frac{D_i}{N_i}\right)$$

11

If Last Observation Censored

- For an interval which ends in a censored observation with no observed events, the conditional probability of surviving within the interval is 1.
- Note also that if the largest observation time is censored, the KM (PLE) survivor function never goes to zero
  - We generally regard the KM (PLE) survivor function to be undefined for times beyond the largest observation time in this situation

12

Kaplan-Meier Properties

- The KM (PLE) survivor functions can be shown to be
  - Consistent: As sample sizes go to infinity, they estimate the true value
  - Nonparametric maximum likelihood estimates
    - But usual asymptotic (large sample) theory for regular, parametric MLE's does not apply
    - Asymptotic (large sample) normal distribution for estimates was established differently

13

Other Derivations of KM

- The KM (PLE) survivor functions can also be derived as the
  - Self-consistent estimator
    - (see Miller, Survival Analysis)
  - "Redistribute to the right" estimator

14

Redistribute to the Right

- Basic idea
  - Recall the empirical cdf assigns probability 1/n to each observation
  - A censored observation should be equally likely to have event time like any of the remaining uncensored observations
    - Recursively redistribute the mass of each censored observation among the subjects remaining at risk

15

Ex: Redistribute to the Right

- Data: 1, 3, 4+, 5, 7+, 9, 10
  - (plus sign means censored)
- Initially: each point has mass 1/7
- Determine probability of events at earliest observed (uncensored) event times
  - Pr ($T^0 = 1$) = 1/7
  - Pr ($T^0 = 3$) = 1/7

16

Ex: Redistribute to the Right

- Censored observation at 4
  - Divide the mass at 4 equally among the remaining subjects at risk
    - Now mass of 1/7 + 1/28 = 5/28 for each of 5, 7, 9, 10
- Determine probability of events at next observed (uncensored) event times
  - Pr (T^0 = 5) = 5/28

17

Ex: Redistribute to the Right

- Censored observation at 7
  - Divide the mass at 7 equally among the remaining subjects at risk
    - Now mass of 5/28 + 5/56 = 15/56 for each of 9, 10
- Determine probability of events at next observed (uncensored) event times
  - Pr (T^0 = 9) = 15/56
  - Pr (T^0 = 10) = 15/56

18

Ex: Redistribute to the Right

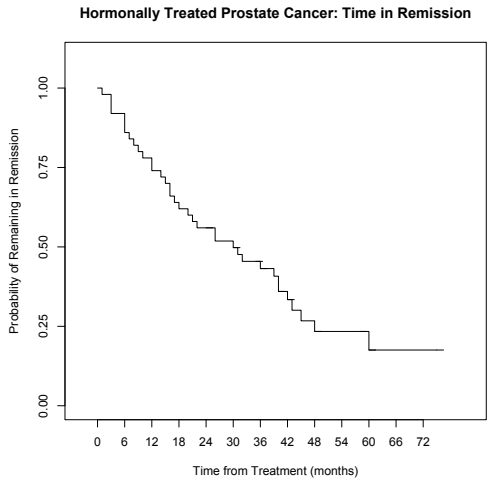
Kaplan-Meier estimate of Survival

t	Pr (T^0 = t)	Pr (T^0 > t)
0		1.000
1	1/ 7 = 0.143	.857
3	1/ 7 = 0.143	.714
4	0.000	.714
5	5/28 = 0.179	.536
7	0.000	.536
9	15/56 = 0.268	.268
10	15/56 = 0.268	.000

19

Example: Prostate Ca Time in Remission

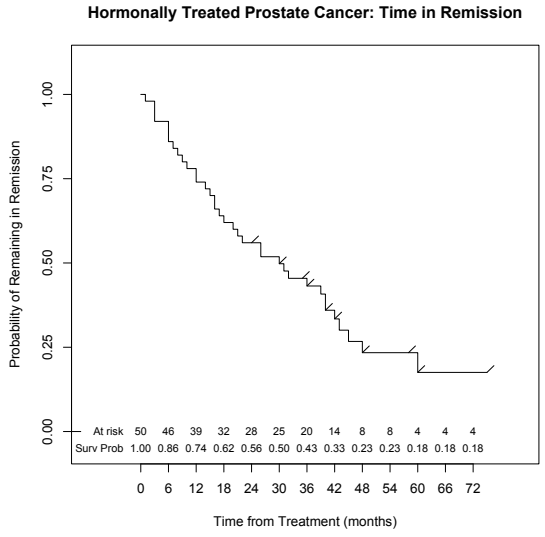
- Time in remission among observational cohort of hormonally treated prostate cancer



20

Better Example: Prostate Ca Remission

- Include number at risk and display censoring times



21

Risk Sets

- Most often, we recognize that the probability of an event depends in some way upon time
- In many cases, that time dependence is something we merely want to adjust for as we compare different groups
  - It is not as important to contrast the event probability over time
- We can sometimes think of our analysis as stratifying on time and analyzing the instantaneous probability of an event

22

Hazard Functions

- We are often interested in the rate (over time) at which individuals convert from being "event-free" to having had the event
  - Time can be calendar time, age, study time …
  - (Differ in what we call time zero and how data is pooled)
- At each point in time, we essentially compute a proportion
  - Denominator: Individuals currently "event-free" (random sample)
  - Numerator: Among those in the denominator, who converts in the next instant
- Referred to as
  - Epidemiology: incidence / mortality rates, force of mortality
  - Statistics and probability: hazard function

23

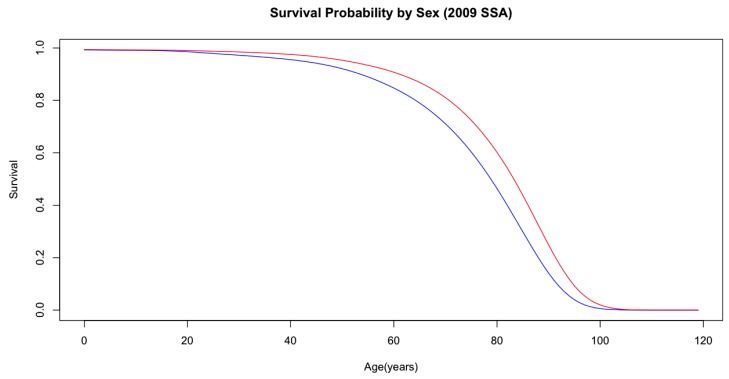
Left Entry

- We can also handle data in which subjects enter the risk set after some unobserved period
  - Potentially three variables may be used
    - Start of interval
      - Usually assumed to be at time 0 if nothing supplied
    - End of interval
    - Status at end of interval
      - 0 = censored
      - Nonzero = event occurred at end of interval
- We just need a risk set that is a random sample of subjects who would have still been at risk at each time point
  - We can follow a population for a year, and estimate the lifetime experience within individual age strata
    - (Presuming no calendar year or birth cohort effects)

24

2009 SSA: Age Effects on Survival

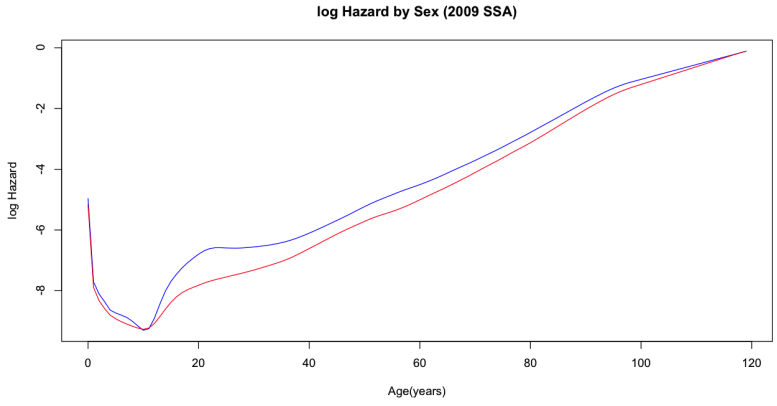
- Estimated survival curves for US population
  - But did not have to follow a single birth cohort



25

2009 SSA: Age Effects on Hazards

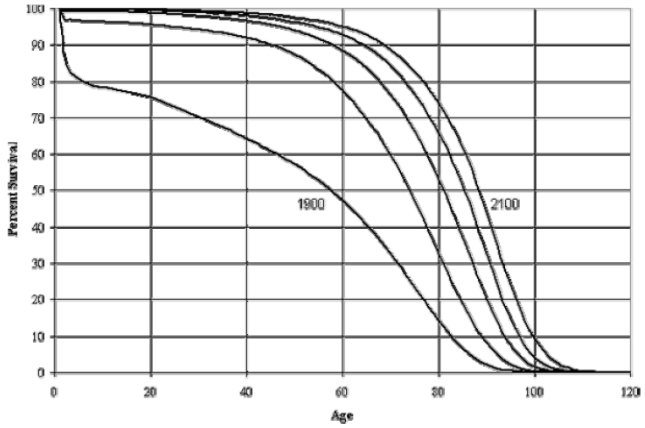
- Could consider hazard in age strata, with some adjustment for calendar (birth cohort) effects
  - Calendar or birth year cohort effects can be major



26

Birth Cohort Effects on Mortality

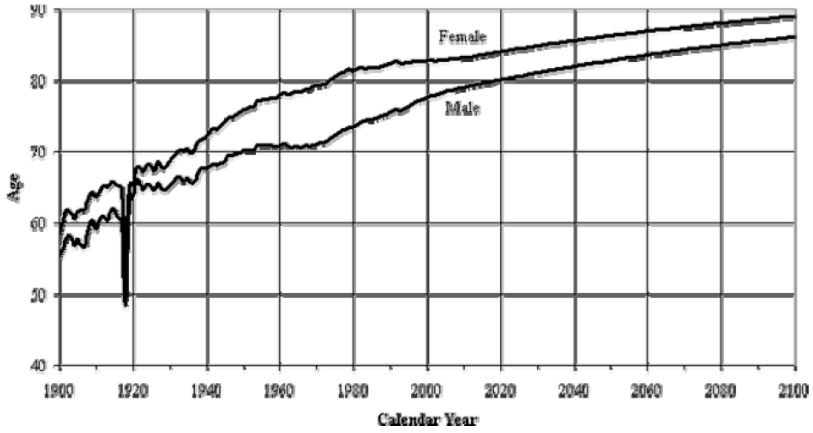
- Survival curves 1900 to 2100 by 50 year increments



27

Calendar Year Effects on Mortality

Figure 4a—Median Age at Death (S(x) = .5)
by Sex and Calendar Year
(Based on Period Tables)



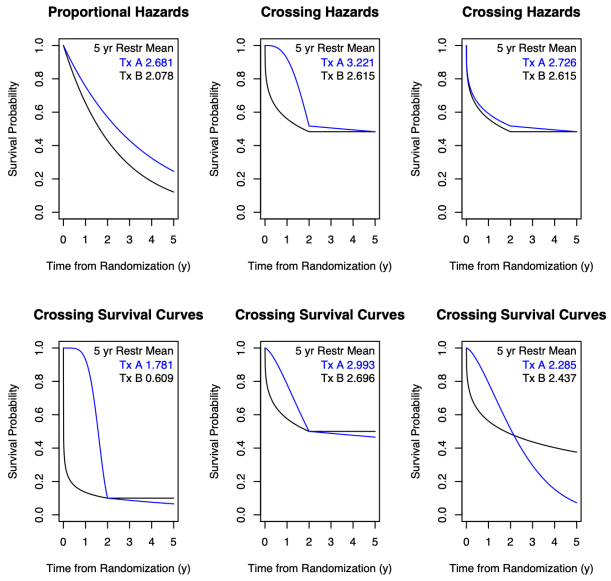
28

Restricted Means

- The area under the KM curve computed for a positive random variable will be related to mean survival time
  - If there is no censoring, this is exactly the sample mean
- If there is not enough follow-up to observe a KM curve decrease to 0, we can only estimate a "restricted" mean
  - E.g., Average years alive during first 5 years
- Best: Pre-specify time restriction prior to analysis of data
  - Some authors have found that using the maximum observation time behaves well

29

Restricted Means (Hypothetical Data)



30

30

Mean Residual Life Expectancy

- We sometimes talk about the "residual" life expectancy within a particular risk set
  - This should not be confused with restricted mean
- For instance, using the 2009 SSA survival estimates

	Mean Residual Life Expectancy	
If survive to age	Males	Females
0	75.9	80.1
20	56.8	61.5
40	38.2	42.2
60	21.3	24.3
80	8.1	9.7

31

Estimating Cumulative Incidence

- Suppose we are estimating the distribution of time to death, with no intercurrent events to worry about
  - We are thus estimating the proportion of subjects who are in each of two possible states: alive or dead
- The Kaplan-Meier estimate of the survival curve can be used to estimate the cumulative incidence

$$F(t) = \Pr(T^0 \leq t) = 1 - Pr(T^0 > t) = S(t)$$

$$\hat{F}(t) = 1 - \hat{S}(t) = 1 - \prod_{j:t_j \leq t} \left(1 - \frac{D_j}{N_j}\right)$$

32

32

Cause Specific Hazards

- Now suppose we are estimating the distribution of time to death from two possible causes, say, CVD or nonCVD
  - We have to estimate the proportion of subjects who are in each of three possible states: alive ($X = 0$), dead from CVD ($X = 1$), or dead from nonCVD ($X = 2$)
- Cause specific hazard $\hat{h}_{jk}$ at time t_j can be based on number of deaths D_{jk}, from cause k where number still at risk (alive) just prior to time t_j is N_j

$$\hat{h}_{jk} = \frac{D_{jk}}{N_j}$$

33

33

Cause Specific Cumulative Incidence

- We cannot just use Kaplan-Meier estimates to compute cumulative incidence in the presence of competing risks
  - Such would pretend that the censored subjects were potentially at risk for the other cause of death
- We must use the Aalen-Johansen estimator for cumulative incidence

$$\widehat{Pr}(T^0 \leq t, X = k) = \sum_{j:t_j \leq t} \widehat{Pr}(T^0 \geq t_j) \frac{D_{jk}}{N_j} = \sum_{j:t_j \leq t} \hat{S}(t_j - 0) \frac{D_{jk}}{N_j}$$

34

34

Informative Competing Risks

- Note that the Aalen-Johansen estimator "accounted" for the competing risk in the sense that it ensured that subjects who failed due to the competing risk would not be presumed to still be at risk for the primary event
- However, in an extreme setting in which a treatment causes nonCVD death just prior to when a CVD death would have been observed
  - the cause specific incidence of CVD death would decrease, and
  - the cause specific incidence of nonCVD death would increase

35

35