# Lecture 10

Monday, November 20, 2023 10:01

BIOS522_Slides10

EMORY ROLLINS SCHOOL OF PUBLIC HEALTH | Department of Biostatistics and Bioinformatics

*BIOS 522: Survival Analysis Methods*

## Lecture 10:

## Left truncation and interval censoring

## Previously


- Observation starts at the time origin
- The time of the event is known precisely (unless right-censored)
- We are interested in only one type of event

2

# Other data structures

- Observation starts at the time origin → this assumption is relaxed
  - **Delayed/staggered entry**: observation can start after the time origin
- The time of the event is known precisely (unless right-censored) → this assumption is relaxed
  - **Interval censoring**: time of the event is known to lie within an interval



# Delayed entry

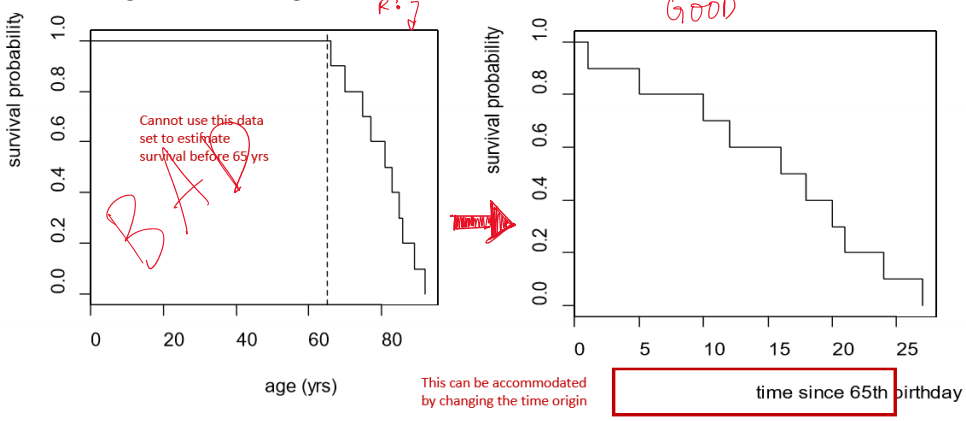
- *Example:* Study human survival
- Enroll participants aged 65 yrs and follow them over time
- Event of interest is death
- Participants die at the following ages:

$$\begin{pmatrix} 66, 70, 75, 77, 81, \\ 83, 85, 86, 89, 92 \end{pmatrix}$$



# Delayed entry

if we just plug this data into R :)



# Time origin examples

| Setting | Time origin | Event | Time scale |
|---|---|---|---|
| Human mortality | Birth, 65th birthday | Death | Age |
| Clinical trial | Randomization | Stroke | Follow-up time |
| Pregnancy cohort | 12 weeks gestation | Fetal death | Gestational age |
| Hospital study | Admission | Discharge | Time in hospital |
| Cancer cohort | Diagnosis | Tumor recurrence | Time since diagnosis |
| Ebola survival study | Date of symptom onset | Death | Time since symptom onset |
| Influenza study | Start of flu season (October 1, 2018) | Influenza symptom onset | Calendar time |

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

6

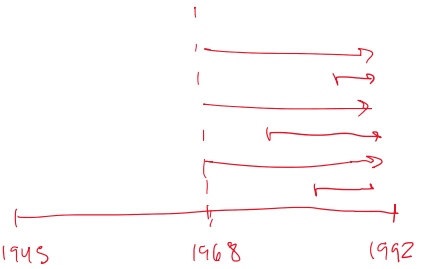
# Atomic bomb long-term survivor study

- Study of long-term outcomes in survivors of the atomic bomb dropped in Hiroshima in 1945
- Study started in 1968 – 23 years later
- *What can we say about outcomes within the first 23 years?*

- In addition, not everyone enrolled in 1968
- Rolling enrolment all the way to study end in 1992
- *What are our analytical options*

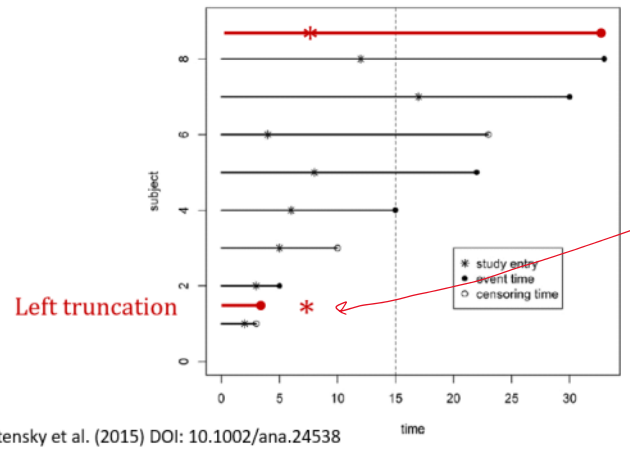
7

People entering at different times




# Staggered entry

Challenge: Same study entry time, but one who survives for a long time and one who does before entry




Betensky et al. (2015) DOI: 10.1002/ana.24538

8

compared to censoring:
At least in censoring you know this person exists.
W/ left trunc. you don't even know they exist

# Notation

- Study entry time $S_i$
- Failure/censoring time $T_i^*$
- Event indicator $\delta_i$

- If $T_i^* < S_i$.... this person is never observed! (is left truncated)



# Staggered entry

You need to have some people at time 0

| Entry time $S_i$ | Follow-up time $T_i^*$ | Event indicator $\delta_i$ |
|---|---|---|
| 0 | 2 | 1 |
| 2 | 4 | 1 |
| 4 | 6 | 1 |
| 0 | 6 | 1 |
| 2 | 8 | 1 |
| 4 | 8 | 1 |
| 0 | 28 | 0 |
| 2 | 28 | 0 |
| 4 | 28 | 0 |
| 0 | 28 | 0 |
| 2 | 28 | 0 |
| 4 | 28 | 0 |

| Unique failure/ censoring time $t_j$ | Number at risk $n_j$ during $(t_{j-1}, t_j]$ | Number of deaths $d_j$ at $t_j$ | Conditional survival probability $\hat{q}_j$ | Kaplan-Meier estimate in $[t_j, t_{j+1})$ |
|---|---|---|---|---|
| $t_0 = 0$ | | | | $t = [0,2)$ $\hat{S}(t) = 1$ |
| $t_1 = 2$ | $t = (0,2]$ $n_1 = 4$ | $d_1 = 1$ | $\hat{q}_1 = \left(1 - \frac{1}{4}\right)$ | $t = [2,4)$ $\hat{S}(t) = 0.750$ |
| $t_2 = 4$ | $t = (2,4]$ $n_2 = 7$ | $d_2 = 1$ | $\hat{q}_2 = \left(1 - \frac{1}{7}\right)$ | $t = [4,6)$ $\hat{S}(t) = 0.643$ |
| $t_3 = 6$ | $t = (4,6]$ $n_3 = 10$ | $d_3 = 2$ | $\hat{q}_3 = \left(1 - \frac{2}{10}\right)$ | $t = [6,8)$ $\hat{S}(t) = 0.514$ |
| $t_4 = 8$ | $t = (6,8]$ $n_4 = 8$ | $d_4 = 2$ | $\hat{q}_4 = \left(1 - \frac{2}{8}\right)$ | $t = [8,28)$ $\hat{S}(t) = 0.386$ |
| $t_5 = 28$ | $t = (8,28]$ $n_5 = 6$ | $d_5 = 0$ | $\hat{q}_5 = 1$ | $t = [28, \infty)$ $\hat{S}(t) = 0.386$ |

**1st interval:**

- 4 enter at time 0, 1 dies at time 2

**2nd interval:**

- 3 remain from the 1st interval, 4 more enter at time 2, 1 dies at time 4



# Partial likelihood

One covariate X

$$L(\beta) = \prod_{i=1}^{m} \frac{\exp(\beta X_i)}{\sum_{j \in R(T_i)} \exp(\beta X_j)}$$

For $j$ in risk set $R(T_i)$, $S_j < T_i$ and $T_j^* \geq T_i$

In order to be in the risk set, you need to be observed

*Alcohol & Alcoholism* Vol. 37, No. 1, pp. 87–92, 2002

MODERATE ALCOHOL INTAKE IN PREGNANCY AND THE RISK OF SPONTANEOUS ABORTION

ULRIK KESMODEL[1,3]*, KIRSTEN WISBORG[1,2], SJÚRÐUR FRÓÐI OLSEN[4], TINE BRINK HENRIKSEN[1,2] and NIELS JØRGEN SECHER[1,5]

[1]Perinatal Epidemiological Research Unit, Department of Obstetrics and Gynaecology, [2]Department of Paediatrics, Aarhus University Hospital, Skejby Sygehus, 8200 Aarhus N, [3]Department of Epidemiology and Social Medicine, University of Aarhus, Vennelyst Boulevard 6, 8000 Aarhus C, [4]Maternal Nutrition Group, Danish Epidemiology Science Centre, Statens Serum Institut, Artillerivej 5, 2300 Copenhagen S, Denmark and [5]Department of Obstetrics and Gynaecology, King Faisal Specialist Hospital and Research Centre, Riyadh, Kingdom of Saudi Arabia

(*Received* 4 *May* 2001; *in revised form* 9 *July* 2001; *accepted* 7 *August* 2001)

**Goal**: Researchers sought to examine the relationship between **alcohol intake and the risk of spontaneous abortion** in pregnant people.

**Population**: A cohort was formed of all pregnant people receiving antenatal care at a university hospital in Denmark from 1989 to 1996 who had completed a routine questionnaire on alcohol intake. They included **24,679 singleton pregnancies** in the study.

↳ no twins

Kesmodel et al. (2002) *Alcohol & Alcoholism* https://doi.org/10.1093/alcalc/37.1.87

12

**Outcome variable**: Spontaneous abortion was defined as **spontaneous fetal death before 28 weeks of pregnancy**. Gestational age was calculated from the last menstrual date or using ultrasonographic measurements. Induced abortions were treated as censored observations.

Pregnancies entered into the cohort at the time of questionnaire completion (staggered entry). **No data was available** on spontaneous abortions **before 7 completed weeks of gestation**, so the authors used 7 weeks as the time origin for the study.

↳ staggered entry

**Predictor variables**: The key predictor variable was **alcohol intake as measured by self-report questionnaire**, categorized into <1, 1-2, 3-4, ≥5 drinks/week. Other covariates include maternal smoking habits, caffeine intake, age, pre-pregnant body mass index, marital status, occupational status, education, and parity (number of prior births).

**Statistical analysis:** Researchers fit Cox proportional hazards models with **gestational age as the underlying time scale**.

13

**Results:** Table 2 summarizes the unadjusted $HR_u$ and adjusted (for the predictor variables listed above) $HR_a$ hazard ratios of first trimester spontaneous abortions, Aarhaus, Denmark, 1989-1996. N is the total number of pregnancies, and n is the number of spontaneous abortions.

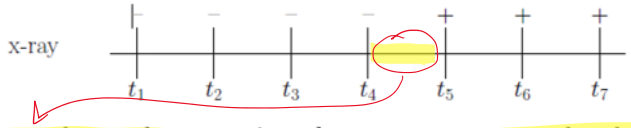
| Alcohol (drinks/week) | Weeks 7–11 (1st trimester)[b] | | | | | | |
|---|---|---|---|---|---|---|---|
| | *N* | Risk weeks | *n* | $HR_u$ | 95% CI | $HR_a$ | 95% CI |
| <1 | 3813 | 7378 | 53 | 1.0 | — | 1.0 | — |
| 1–2 | 1229 | 2472 | 28 | 1.5 | 1.0–2.4 | 1.3 | 0.8–2.0 |
| 3–4 | 403 | 895 | 8 | 1.2 | 0.6–2.5 | 0.8 | 0.4–1.7 |
| ≥5 | 168 | 377 | 15 | 5.2 | 2.9–9.2 | 3.7 | 2.0–6.8 |

Briefly, the hazard of spontaneous abortion is significantly higher for people **reporting drinking ≥5 drinks/week** than people drinking <1 drink/week (adjusted HR 3.7 (95% CI: 2.0 to 6.8)).

14

# Interval censoring

- Imagine an event that is "silent" and requires a **specialized test** to detect (e.g. X-ray, blood test)
- We repeatedly test study participants over time until the event is detected




- We do not know the exact time that an event occurred, only that it lies within some **interval**
- A naïve strategy is to set the event time as the interval midpoint, but this can induce bias, especially when the intervals are long

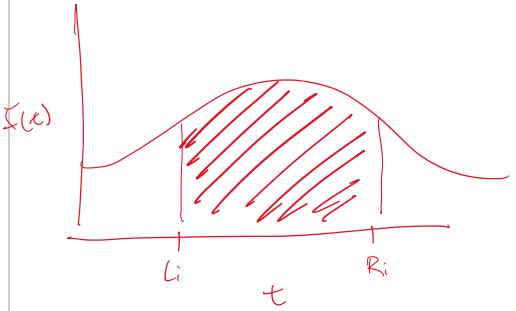
15

# Maximum likelihood for interval censoring

- Instead of $T_i$, we have bounds $(L_i, R_i]$ — true time lies somewhere within this
- Likelihood contribution is the probability that the event occurs between left bound $L_i$ and right bound $R_i$, $\Pr(L_i < T_i \leq R_i)$
- Likelihood for parameter(s) $\theta$:

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} \big(F(R_i|\theta) - F(L_i|\theta)\big)$$

Difference in these densities




17

# Maximum likelihood for interval censoring

- Parametric (Weibull) or non-parametric (Turnbull estimator) approaches exist
- *Example:* exponential model

$$\mathcal{L}(\lambda) = \prod_{i=1}^{n} \Big(\big(1 - e^{-\lambda R_i}\big) - \big(1 - e^{-\lambda L_i}\big)\Big) = \prod_{i=1}^{n} \big(e^{-\lambda L_i} - e^{-\lambda R_i}\big)$$

cdf right hand interval — cdf left hand interval

- For censored observations, the interval is $(L_i, \infty)$ and the likelihood contribution is $\exp(-\lambda L_i)$

$\prod_{i=1}^{n} \left(e^{-\lambda L_i} - e^{-\lambda \infty}\right)$ — $e^{-\lambda\infty}$ goes to 0

The NEW ENGLAND JOURNAL of MEDICINE

ORIGINAL ARTICLE

# Persistence of Zika Virus in Body Fluids — Preliminary Report

Gabriela Paz-Bailey, M.D., Ph.D., Eli S. Rosenberg, Ph.D., Kate Doyle, M.P.H., Jorge Munoz-Jordan, Ph.D., Gilberto A. Santiago, Ph.D., Liore Klein, M.S.P.H., Janice Perez-Padilla, M.P.H., Freddy A. Medina, Ph.D., Stephen H. Waterman, M.D., M.P.H., Carlos Garcia Gubern, M.D., Luisa I. Alvarado, M.D., and Tyler M. Sharp, Ph.D.

**Goal:** Researchers sought to estimate the **frequency and duration of detectable Zika virus (ZIKV) in human body fluids**. These estimates can guide public health policy regarding blood donation and reducing risk of sexual transmission.

**Population:** A prospective cohort was formed from **newly ZIKV infected participants** in Puerto Rico. Patients with acute febrile illness who were laboratory confirmed for ZIKV infection were invited to participate along with **any infected household contacts**. The study included 127 index participants and 23 household contacts (n=150).

Paz-Bailey et al. (2018) *NEJM* https://www.nejm.org/doi/full/10.1056/nejmoa1613108

18

**Outcome variable:** **Serum, urine, saliva, semen** (adults only), and **vaginal secretions** (adults only) were collected weekly for the first month and at 2, 4, and 6 months thereafter. Among the participants in whom ZIKV was detected in any specimen at week 4, biweekly collection continued until all the specimens tested negative.

The outcome variable was **time from the onset of ZIKV symptoms to the first negative RT-PCR result** (no viral RNA detected). For participants with intermittent viral shedding, the outcome was time to the first negative result after the final positive result. Data were censored for participants who still had positive results at the time of data analysis.

**Predictor variables:** Data were collected on age, sex, pregnancy, symptoms at enrollment, time from symptom onset to enrollment, and other laboratory findings (white-cell count, platelet count, and hematocrit).

19

**Statistical analysis:** To estimate the population distribution of viral clearance times, they used two methods:

(1) the non-parametric maximum-likelihood Turnbull estimator, and

(2) parametric Weibull regression models.

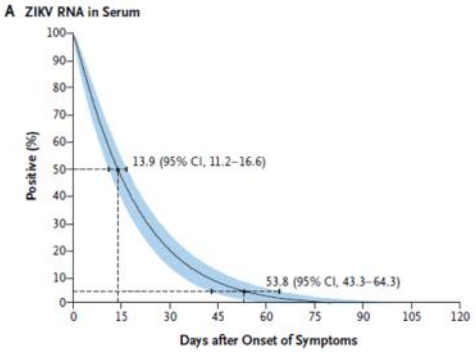
Both methods account for interval censoring since the loss of RNA detection occurred within an interval between visits.

20

**Statistical analysis:** Figure 1A summarizes the estimated time until loss of ZIKV RNA detection after the onset of symptoms in serum as estimated with the use **Weibull regression**. Median and 95th percentiles are noted. Blue shading denotes 95% confidence intervals.

The median time was 14 days (95% CI: 11 to 17), and the 95th percentile time was 54 days (95% CI: 43 to 64).

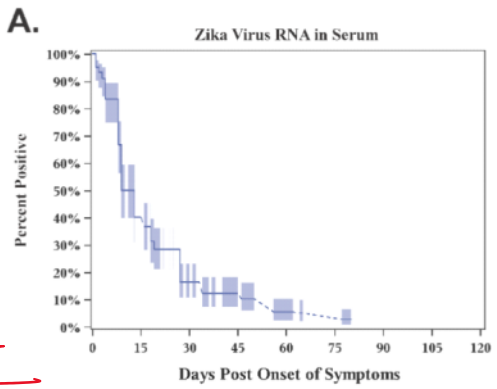
Note the smoothness of the curve as is consistent with the smoothness of the Weibull distribution.




21

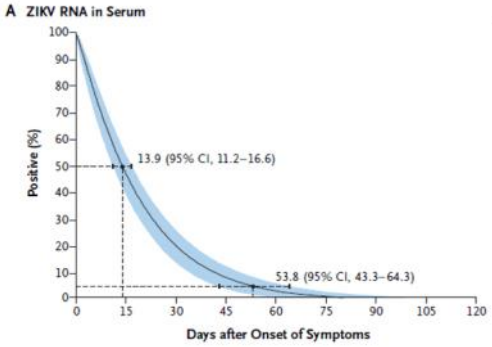
Supplement Figure S2A summarizes the estimated clearance time distribution using the **non-parametric Turnbull method**. Blue shading denotes 95% CI. Median time to loss of ZIKV RNA detection was 13 days (95% CI: 9 to 13).

Note the stepped shape of the curve, as is consistent with a nonparametric estimator. Both return similar estimates of the distribution of clearance times.

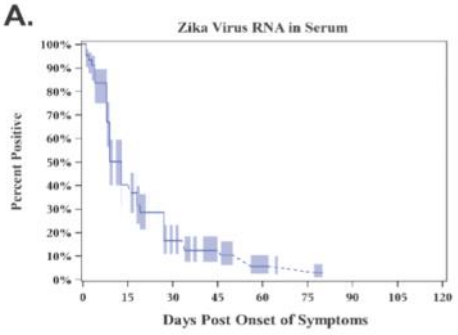



22

**Parametric estimator**
(e.g. Weibull distribution)




**Nonparametric estimator**
(e.g. Turnbull's method)




23

# Preventing Microalbuminuria in Type 2 Diabetes

Piero Ruggenenti, M.D., Anna Fassi, M.D., Anelja Parvanova Ilieva, M.D., Simona Bruno, M.D., Ilian Petrov Iliev, M.D., Varusca Brusegan, M.D., Nadia Rubis, R.N., Giulia Gherardi, R.N., Federica Arnoldi, R.N., Maria Ganeva, Stat.Sci.D., Bogdan Ene-Iordache, Eng.D., Flavio Gaspari, Ph.D., Annalisa Perna, Stat.Sci.D., Antonio Bossi, M.D., Roberto Trevisan, M.D., Alessandro R. Dodesini, M.D., and Giuseppe Remuzzi, M.D., for the Bergamo Nephrologic Diabetes Complications Trial (BENEDICT) Investigators

Goal: Assess whether **ACE inhibitors** and **calcium channel blockers**, alone or in combination, **prevent microalbuminuria in type 2 diabetics**

Population: **1204 subjects** with type 2 diabetes mellitus and normal urinary albumin excretion were **randomized to** receive at least three years of treatment with **one of the four following regimens**:

(1) trandolapril + verapamil, (2) trandolapril only,

(3) verapamil only, or (4) placebo.

24

Outcome variable: Development of **persistent microalbuminuria** (overnight albumin excretion, ≥20 µg per minute at two consecutive visits). Measured from date of administration of the first study drug.

Morning **urine samples were collected** at the time of randomization, at one week, one month, and three months after randomization, and **every three months thereafter**.

Statistical analysis: Time to persistent microalbuminuria was **interval-censored**. The method accounts for the fact that the true event time lies between the last time the subject had normal excretion and the first time the subject had microalbuminuria validated.

25

(1) Statistical analysis (continued): The primary analysis used an **accelerated failure-time model with a log-normal baseline hazard** which directly incorporated the interval-censored-data.

we haven't talked about log normal for AFT but it's just the baseline hazard function

The **model was adjusted** for study site, patient age, sex, smoking status (never vs. current/former smokers), diastolic blood pressure, and baseline albumin excretion.

(2) **Kaplan-Meier curves** were plotted for each treatment group, with the use of the **midpoint of the intervals** as event times. A **Cox** (3) **regression model** was also applied, to ensure the robustness of the results.

she doesn't recommend mid point but it is a way to get a crude plot (could have produced log normal curves)

26

looking at the midpoint ones

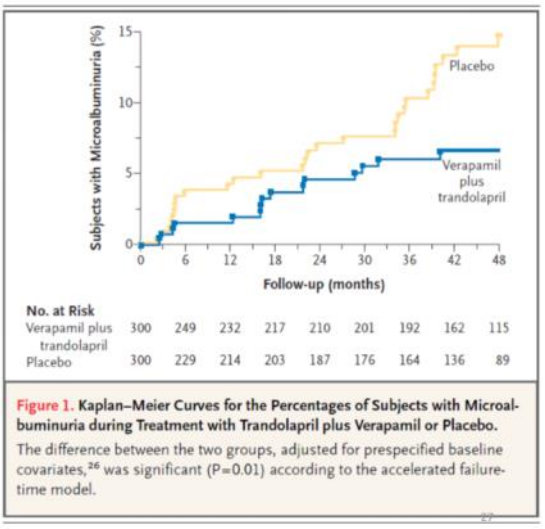
Results: KM curves for verapamil + trandolapril versus placebo are shown. These groups clearly separate at 3 months.

from the AFT

The estimated **acceleration factor** after controlling for predefined baseline variables was **0.39** (95% CI: 0.19 to 0.80; p=0.01) for the combined regimen vs. placebo.*

"Thus, the combined regimen significantly **delayed microalbuminuria onset by a factor of 2.6**" (=1/0.39).

* They have reported exp(−α). The quoted sentence interpreting the results is very important for assessing the direction of the effect.




| No. at Risk | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Verapamil plus trandolapril | 300 | 249 | 232 | 217 | 210 | 201 | 192 | 162 | 115 |
| Placebo | 300 | 229 | 214 | 203 | 187 | 176 | 164 | 136 | 89 |

**Figure 1. Kaplan–Meier Curves for the Percentages of Subjects with Microalbuminuria during Treatment with Trandolapril plus Verapamil or Placebo.**

The difference between the two groups, adjusted for prespecified baseline covariates,[26] was significant (P=0.01) according to the accelerated failure-time model.

diff models described in one figure