Intro to Survival Analysis + my project

Peem L. May 27th

What is survival analysis?

- A regression problem on **time until an event occurs.**
- **Examples:**
  - Predict how long an admitted patient will stay in a hospital.
  - Predict when a subscriber unsubscribes from a service.

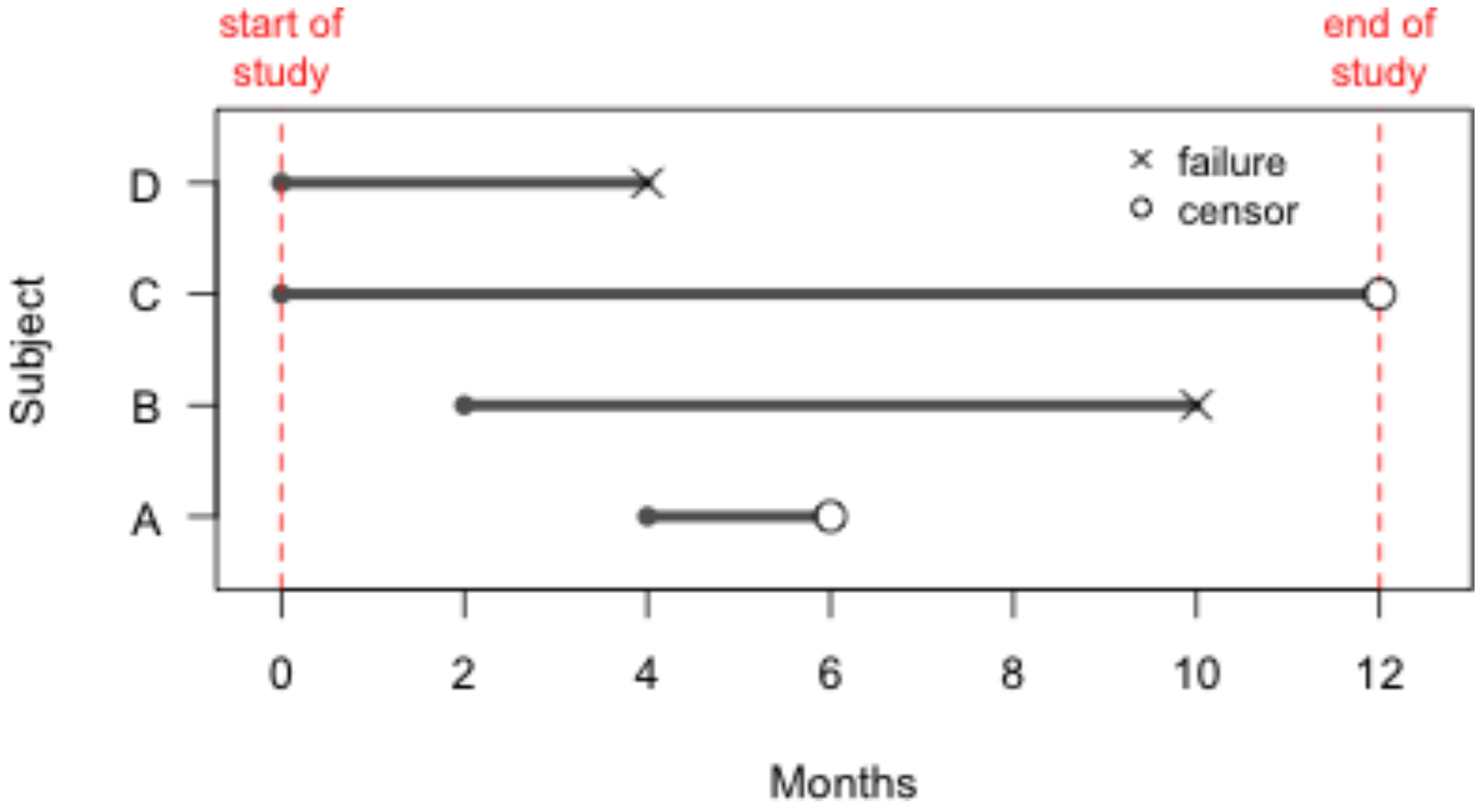
Censored data makes survival analysis challenging.

- Suppose we study how long patients with stage-2 cancer will survive. We might collect data for 5 years, but at the end of study, some patients might still survive.
- **Reasons for censoring**: Study ends, patients withdraw from studies, etc.
- **Challenge:** We haven't fully observed their outcomes, but we don't want to discard the data either.

Survival time and censoring time

- For each individual i, suppose there exists a true failure time T_i and a true censoring time C_i. However, we can only observe $y_i = min(T_i, C_i)$.
  - If an event appears before the study ends, $y_i = T_i$.
  - Else, if the study ends and nothing happens to that individual (i.e., is censored), $y_i = C_i$.
- We can denote our dataset as (x_i, y_i, δ_i), where x_i denotes feature vectors, y_i as above, and δ_i = 0 implies censored δ_i = 1 implies fully observed.

Examples


start of study
end of study
failure
censor
D
C
B
A
Subject
0
2
4
6
8
10
12
Months

What are we estimating in survival analysis?

- **Survival function:** $S(t|x) = Pr(T > t|x) = 1 - F(t|x)$, where $F(t|x)$ denotes CDF of conditional survival time.

- **Cumulative hazard function:** $H(t|X) = -logS(t|X)$.

- **Hazard function:** $h(t|x) = \frac{\partial}{\partial t} H(t|x)$

Q: Why do we care about censoring?

- **Answer:** Yes, especially to examine whether there are systematic reasons why censoring occurs.

- **Example:**

  - **Data:** SUPPORT dataset on survival time of seriously ill patients ($n = 8{,}873$)

  - **Task:** We want to estimate how many patients survive beyond 100 days.

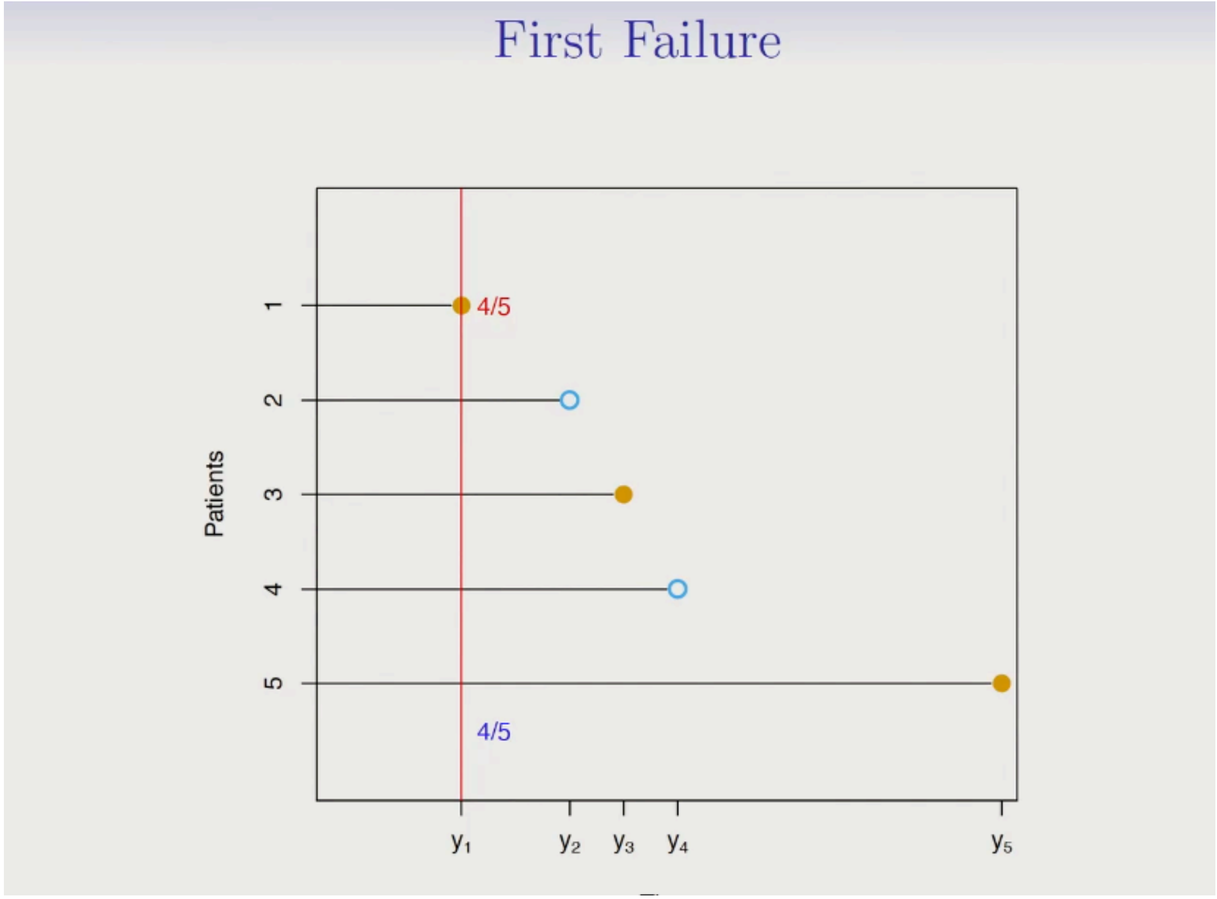
Examples on why censoring matters.

- **Idea:** Number of patients surviving beyond 1,000 days / total patients = 1,522/8,873 = 17.15%.
  - However, **1,584** patients are censored with observed duration < 1,000. So, 17.15% is an underestimated.
  - We essentially assume these censored patients are all dead.

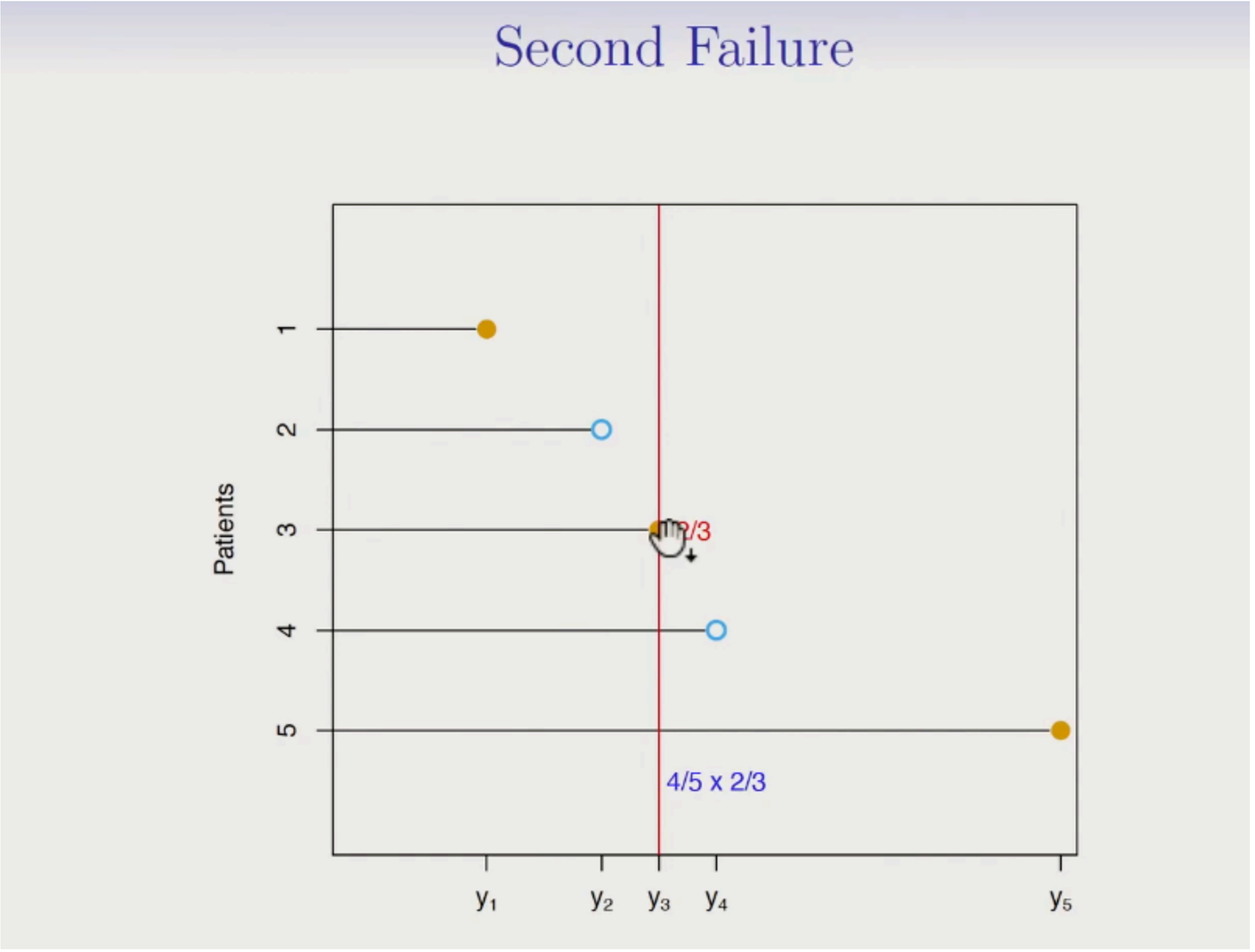
Kaplan-Meier estimator (1959)

- **What it does:**
  - 1. List of periods when an event occurs.
  - 2. Compute products of conditional probability of surviving until each period.
- **Assume:** Independent censoring

K-M estimator (1959) (cont.)

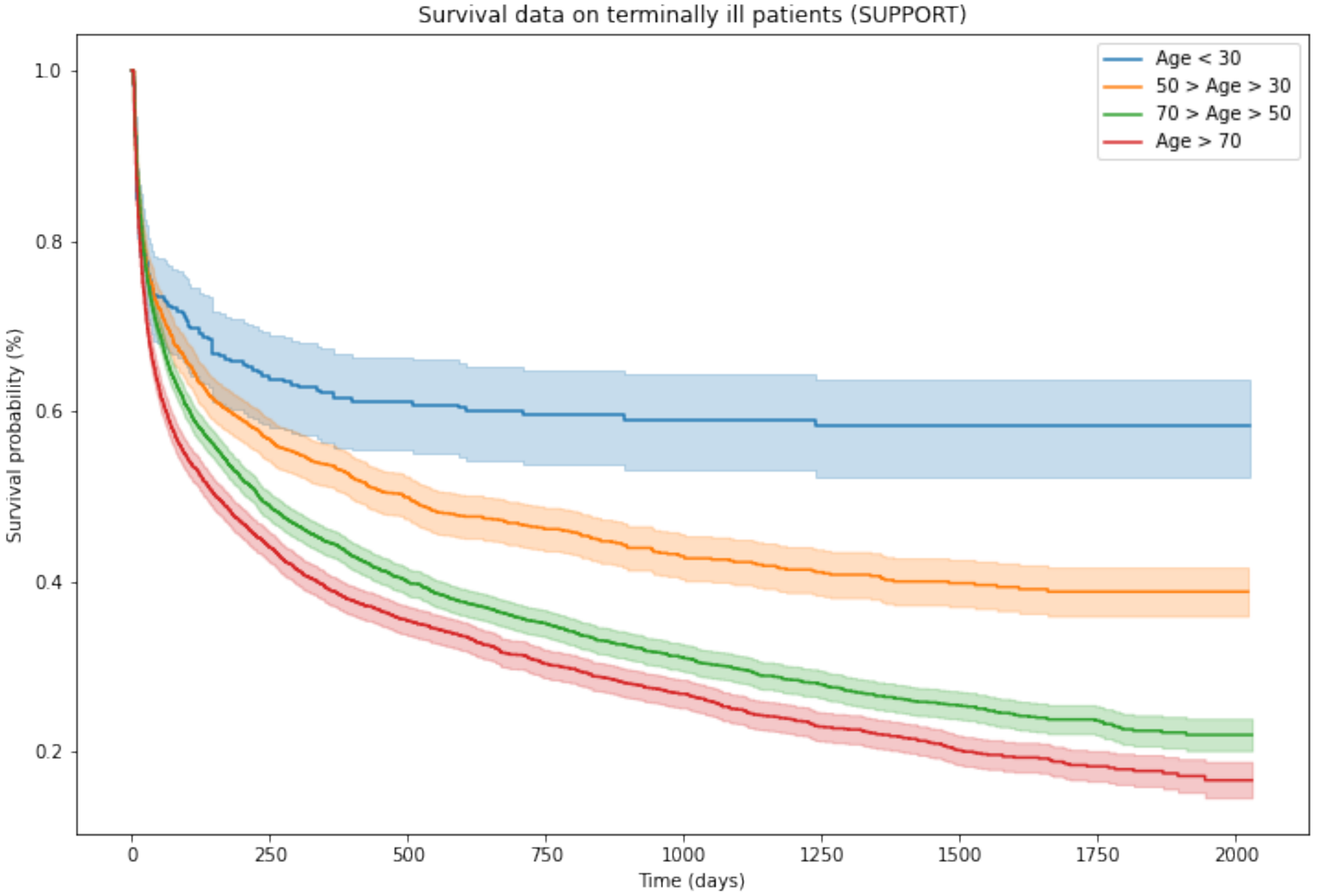


K-M estimator (1959) (cont.)



K-M estimator (1959) (cont.)

K-M estimator on SUPPORT.



Age < 30: 304

50 > Age > 30: 1,550

70 > Age > 50: 3,854

Age > 70: 3,165

Cox-proportional hazard model

- Note that K-M doesn’t use any covariate. However, we might want to study how each term influences the survival probability. Recall the Hazard function:

$$h(t|x) = \frac{\partial}{\partial t} H(t|x)$$

- $$= \lim_{\Delta t \to 0} \frac{Pr(t \leq T \leq T + \Delta t)}{\Delta t}$$

Cox-proportional Hazard model

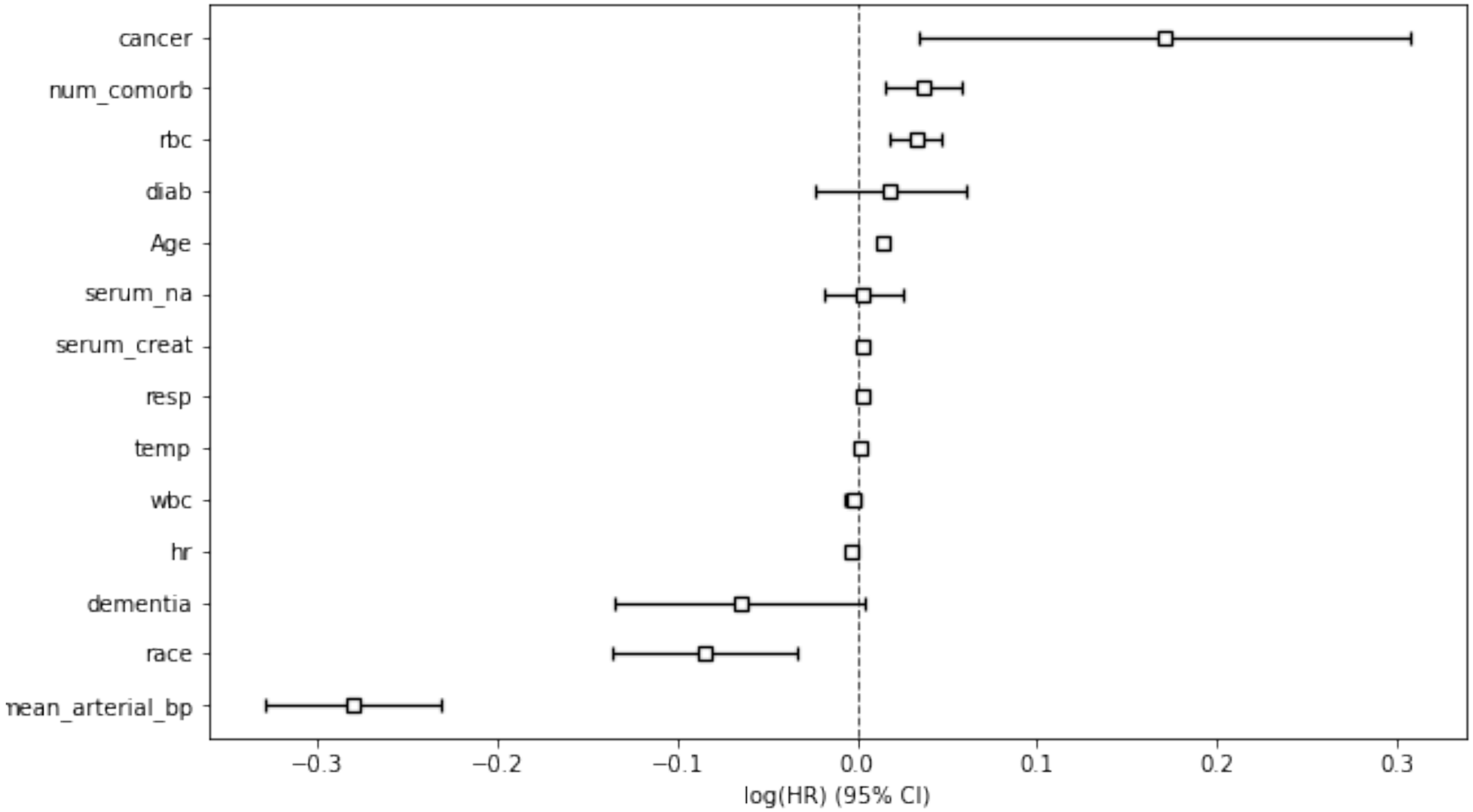
- Use hazard function as the basis for regression:

$$h(t|x_i) = h_0(t)\exp\left(\sum_{j\in[p]} x_{ij}\beta_j\right)$$

- Here, β_j refers to the coefficient for covariates of interests.

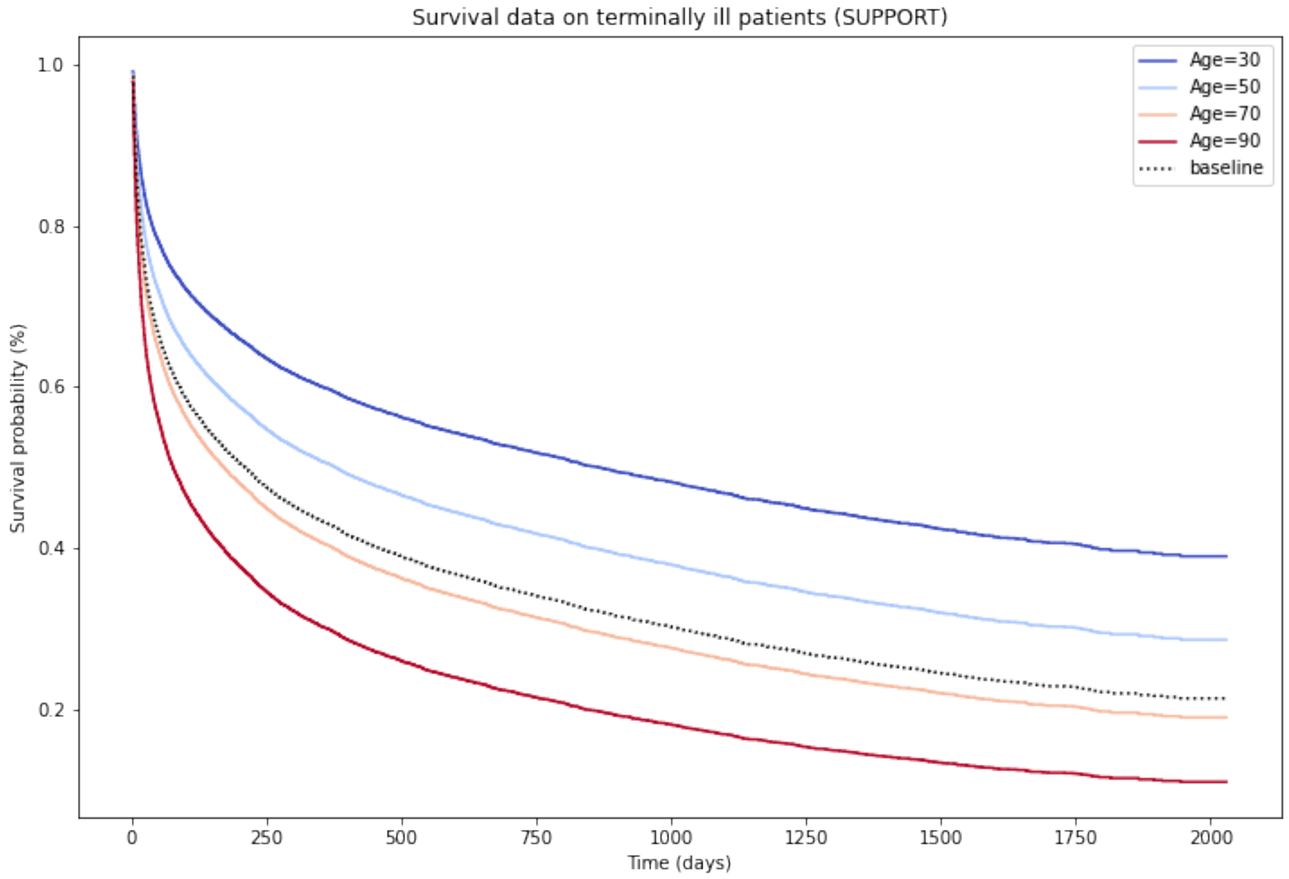
- Then, $h(t|x_i) \rightarrow H(t|x_i) \rightarrow S(t|x_i)$

Cox-proportional hazard on SUPPORT.



Cox-proportional hazard model on SUPPORT.

C-PH on SUPPORT (cont.)



My project with Nynke, Larry, George Chen.

- Currently, there exist many survival models beyond Kaplan-Meier and Cox proportional hazard models, specifically one that uses machine learning (or deep + sth.).

- Besides performance on test set, we want to **quantify uncertainty** in the predicted survival curve, which is more difficult than a point. Methods such as conformal prediction provide too large of an interval.

- **Question:** Can we provide conditional guarantees on a specific interval of interest?

- **Data:** Medical Information Mart for Intensive Care III (MIMIC-III) (EHR on 2,183 patients).

Credit

- https://sakai.unc.edu/access/content/group/2842013b-58f5-4453-aa8d-3e01bacbfc3d/public/Ecol562_Spring2012/docs/lectures/lecture27.htm

- ISLR Version 2 by Hastie et. al.