Survival analysis for healthcare data

M2 Données massives en santé

Juliette Murris

January 2025

Outline

Introduction

Time-to-event and right censoring

Non-parametric estimations

Semi-parametric estimations

Parametric estimations

Overall

Introduction

Introduction

You've just learned that you have a serious, potentially fatal illness. What questions come to mind?

- How can I reduce my risk and improve my chances of survival?
- What are my chances of being alive in 10 years?
- Among others with this same condition, what are their survival rates at 3 months, 1 year, and 5 years?
- How much time do I have left?

These fundamental questions drive the importance of survival analysis in healthcare.

Objectives of survival analysis

Survival analysis studies the time until specific events occur

- Medical Applications: Death, Relapse, Hospitalization, Remission
- Other Applications: Gaming-level progression, machine failure, PhD completion

Key goals

- Estimate survival time distributions
- Compare survival functions between groups
- Analyze effects of explanatory variables

Survival analysis in healthcare research

- **Critical tool in evaluating treatment effect**
  - Primary endpoints in oncology clinical trials
  - Measuring patient outcomes over time

- **Usual survival endpoints in clinical studies**

  Overall Survival (OS) Time from randomization to death

  Progression-Free Survival (PFS) Time from treatment start to disease progression

eunethta
EUROPEAN NETWORK FOR HEALTH TECHNOLOGY ASSESSMENT

FDA

NICE National Institute for Health and Care Excellence

The dataset we'll use throughout this course

```r
require(survival)
data(bladder1)
head(bladder1)
```

- **id**: Patient identifier
- **treatment**: Placebo vs. thiotepa vs. pyridoxine
- **number**: Initial tumor count (8=8+)
- **size**: Largest initial tumor (cm)
- **recur**: Number of recurrences
- **start, stop**: Interval times
- **status**: 0=censored, 1=recurrence, 2=cancer death, 3=other death
- **rtumor**: Tumors at recurrence
- **rsize**: Largest tumor at recurrence
- **enum**: Recurrence number (max 4)

Evaluating Thiotepa in bladder cancer recurrence

Why not compare recurrence percentages?

Patients lost to follow-up:

× Excluding them → Reduced statistical power

× Including them → Invalid unless follow-up periods are equal

Why not compare time to recurrence?

Patients without recurrence:

× Excluding them → Loss of power and information

× Including with arbitrary values → Artificially inflated means

✓ Solution: Survival analysis combining time data and recurrence status

Definitions – Key Dates

Origin Date (OD): Study entry date for the patient

- Randomization date
- Study inclusion date
- Diagnosis date

Last Follow-up Date (LFD): Most recent patient contact

- Death date
- Last completed visit date

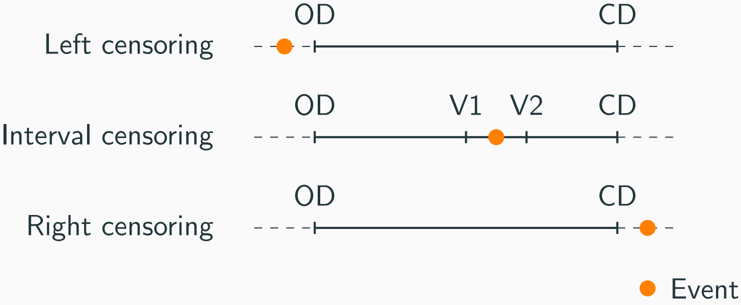
Cut-off Date (CD)
Pre-specified date (in protocol) marking the study analysis point. Any information collected after this date is not considered in the analysis.

Definitions – Censoring

Censoring occurs when the exact date of an event is unknown.

There are three types of censoring:

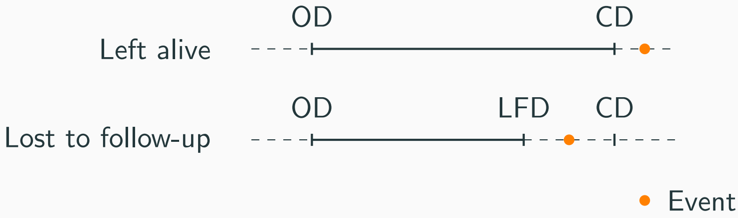
- Left censoring: Event occurs before the OD
- Interval censoring: Event occurs between two observations (e.g., visits)
- Right censoring: Event occurs after the end of subject observation



Right censoring

Right censoring occurs in two scenarios:

- Alive without event: Subject hasn't experienced the event by CD
- Lost to follow-up: Unknown if subject experienced the event



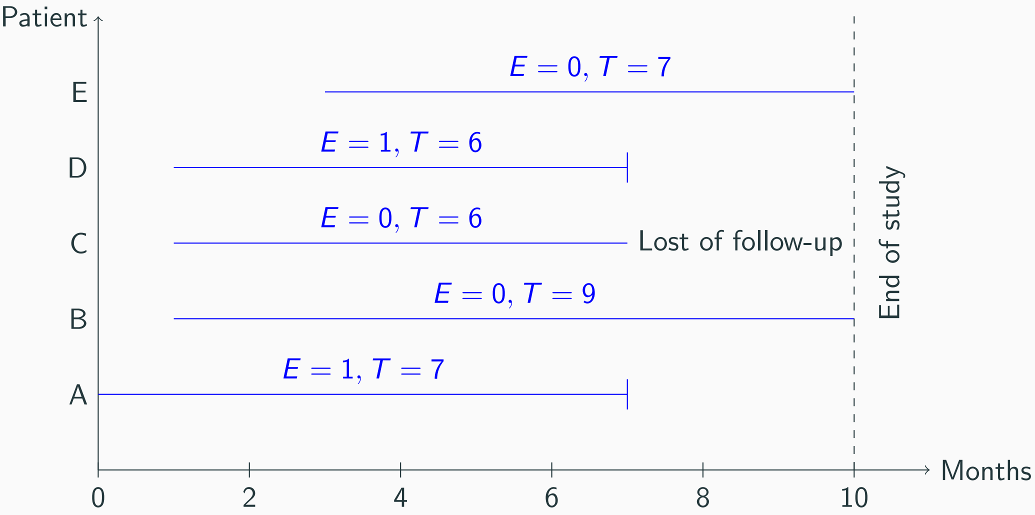
Survival data – Requirements

For survival analysis, two key elements are needed:

1. A **binary variable**, E:
   - **0**: Event of interest not observed during follow-up
   - **1**: Event occurred during study period

2. A **duration**, T:
   - If $E = 1$: Time from study start to event occurrence
   - If $E = 0$: Time from study start to last follow-up

Survival data – Example



Time-to-event and right censoring

Survival time and right censoring

Survival time is the duration from the origin date until a specific event occurs. Let:

- $C \in [0, \infty)$: time to right censoring
- T^*: time to event of interest
- $\delta \in \{0, 1\}$: event status indicator

Two possible scenarios for individual i

- If $T_i^* \leq C_i$: Event is observed before censoring
  - Event time is known
  - $\delta_i = 1$
- If $T_i^* > C_i$: Event is not observed during study period
  - Event time unknown or event did not occur
  - $\delta_i = 0$

With $T \perp\!\!\!\perp C$, survival time $T = T^* \wedge C$ where $a \wedge b = \min(a, b)$ and T is non-negative with continuous distribution

Independent censoring

Key Assumption
Initially, we assume independence of the censoring process: subjects censored at time t should not constitute a biased sample of those at risk at the same time t.

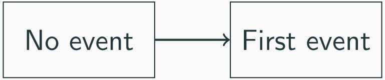
Note
It is generally impossible to verify the independent censoring assumption from available data. However:

- Censoring due to being alive at study end can usually be considered "independent"
- It is recommended to follow up on lost subjects
- Document reasons for loss when possible (e.g., discontinued follow-up, emigration)

Time to first event

For *classical* survival analysis, we focus on the occurrence of the first event

The individual starts in state 0, meaning they have not experienced any event and may remain in this state
As soon as an event occurs, the individual transitions to state 1



Basic functions

The **distribution function** $F(t)$ and density function $f(t)$ are related by:

$$F(t) = P(T \leq t) = \int_0^t f(u)du \tag{1}$$

The **survival function** $S(t)$ is the probability of not experiencing the event before time t:

$$S(t) = 1 - F(t) = P(T > t) \tag{2}$$

where $S(0) = 1$ and $\lim_{t \to \infty} S(t) = 0$

Hazard functions

The **instantaneous hazard function** $\lambda(t)$ is the instantaneous risk of event occurrence at time t, given survival until t:

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} \tag{3}$$

The **cumulative hazard function** $\Lambda(t)$ represents the accumulated risk up to time t:

$$\Lambda(t) = \int_0^t \lambda(u) du \tag{4}$$

These functions are related by:

$$\begin{cases} \lambda(t) = \frac{f(t)}{S(t)} \\ S(t) = \exp(-\Lambda(t)) = \exp(-\int_0^t \lambda(u) du) \end{cases} \tag{5}$$

Non-parametric estimations

Survival function estimation with Kaplan-Meier (1958)

For ordered event times t_k, the KM estimate is defined as:

$$\hat{S}_{KM}(t) = \prod_{k=1}^{K}(1 - \frac{d_k}{n_k}) \tag{6}$$

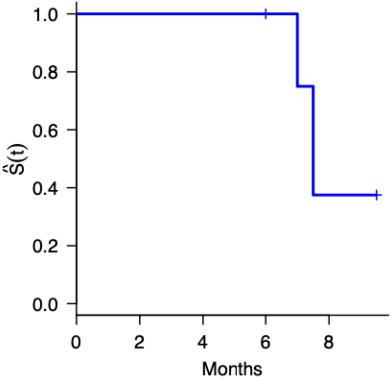
where:

- $k : t_k \leq t$ are times with at least one event
- d_k is the number of events between t_k and t_{k-1}
- n_k is the number of subjects at risk just before t_k

Note
In the absence of censoring, the Kaplan-Meier estimator is equivalent to the empirical survival function.

Example of Kaplan-Meier estimation

t_k	N_k	D_k	$\hat{S}(t_k)$
0	5	0	1
6	5	0	1
7	4	1	$1 \times 0.75 = 0.75$
7.5	2	1	$0.75 \times 0.5 = 0.375$
9.5	1	0	0.375



In practice

```r
    km = survfit(Surv(stop, recidive) ~ 1,
                    data = bladder_v1)
    km
```

Q: What is the median survival time? How should we interpret it?

Comparing survival curves: The Log-rank test

Test Hypotheses

- H_0: Equality of survival functions vs.
  H_1: At least one survival function differs from others
- Compares observed events in each group to what's expected under the null hypothesis
- Key assumption: survival curves do not cross

Let O_i and E_i be the observed and expected number of events in group i. The test statistic is:

$$\frac{(O_A - E_A)^2}{E_A} + \frac{(O_B - E_B)^2}{E_B}$$

For k groups, the p-value is obtained from a χ^2 distribution with $k - 1$ degrees of freedom

In practice

```r
km = survfit(Surv(stop, recidive) ~ treatment,
                   data = bladder_v1)
km
```

```r
plot(km, lty = c(1,2))
legend("topright", c("Placebo", "thiotepa"), lty = 1:2)
```

Q: What does this plot tell us?

```r
survdiff(Surv(stop, recidive) ~ treatment, data = bladder_v1)
```

Q: Compare the survival functions and draw conclusions.

Cumulative hazard function estimation with Nelson-Aalen (1995)

The Nelson-Aalen estimator estimates the cumulative hazard using a step function that increases at each ordered event time:

$$\hat{\Lambda}_{NA}(t) = \sum_{k=1}^{K} \frac{d_k}{n_k} \tag{7}$$

where $k : t_k \leq t$. This represents the cumulative sum of estimated instantaneous hazard rates at each event time.

The Kaplan-Meier and Nelson-Aalen estimators are related by:

$$\hat{S}_{KM}(t) = \prod_{u \leq t} (1 - \Delta\hat{\Lambda}_{NA}(u)) \tag{8}$$

where the product is over all unique event times u, $u \leq t$, and $\Delta\hat{\Lambda}_{NA}(u)$ is the increment in the Nelson-Aalen estimator at time u.

Non-parametric estimations

Limitations

Non-parametric estimates can:

- Identify single prognostic factors (treatment assignment, patient characteristics)
- Cannot address individual patient data questions
- Do not account for multiple patient characteristics simultaneously

Semi-parametric estimations

Introduction to Cox Proportional Hazards Model (1972)

Let $Z = (Z_1, ..., Z_n)$ where $Z_i = (Z_{i1}, ..., Z_{ip})^T$ represent:

- p different covariates (predictors)
- Measured on n individuals
- Each Z_i is a vector of p characteristics for individual i

The Cox model is semi-parametric as it combines:

- Non-parametric component: baseline hazard $\lambda_0(t)$
- Parametric component: relative risk function $\exp(\beta^T Z)$

$$\lambda(t|Z) = \lambda_0(t) \cdot \exp(\sum_{j=1}^{p} \beta_j Z_j) \tag{9}$$

Model components

Baseline hazard $\lambda_0(t)$

- Hazard function when all covariates equal zero $\lambda(t|Z_{i1} = 0, ..., Z_{ip} = 0) = \lambda_0(t)$
- Left unspecified (non-parametric)
- Changes with time but same for all subjects

Parametric $\exp(\sum_{j=1}^{p} \beta_j Z_j)$

- Multiplicative effect on hazard
- Time-independent
- β_j: log hazard ratio for one-unit increase in Z_j
- $\exp(\beta_j)$: hazard ratio for one-unit increase in Z_j

Hazard Ratios

For individuals i and $\tilde{i}$:

$$\frac{\lambda(t|Z_i)}{\lambda(t|Z_{\tilde{i}})} = \frac{\lambda_0(t) \cdot \exp(\beta^T Z_i)}{\lambda_0(t) \cdot \exp(\beta^T Z_{\tilde{i}})} = \frac{\exp(\beta^T Z_i)}{\exp(\beta^T Z_{\tilde{i}})}. \tag{10}$$

Properties:

- Independent of baseline hazard
- Constant over time (proportional hazards)
- Interpretable as relative risk

Hazard Ratios

Single Covariate Change
If all the values of Z_i and $Z_{\tilde{i}}$ are equal with the exception of the kth value, where $Z_{ik} = Z_{\tilde{i}k} + 1$ and $k \in \{1, ..., p\}$, then for unit increase in covariate k:

$$\frac{\lambda(t|Z_i)}{\lambda(t|Z_{\tilde{i}})} = \exp(\beta^T(Z_i - Z_{\tilde{i}})) = \exp(\beta_k) \tag{11}$$

- Effect isolated to single variable
- All other covariates held constant
- Direct interpretation of coefficient

Hazard Ratio – Interpretation

For a treatment effect with $trt = Z_1 = 1$ (treated) vs. $Z_1 = 0$ (control), $HR_{trt} = HR_{Z_1} = exp(\beta_1)$:

- $HR_{trt} < 1 \rightarrow$ **Protective factor**: instantaneous risk in treated group is lower than in control group
- $HR_{trt} = 1 \rightarrow$ **No effect**: instantaneous risk in treated group equals that of control group
- $HR_{trt} > 1 \rightarrow$ **Risk factor**: instantaneous risk in treated group is higher than in control group

Parameter estimation

We want to find $\mathbb{P}(Z_i|t)$: probability that individual i experiences the event at time t

Individual i's contribution to model likelihood:

$$\mathcal{L}_i(\beta) = \mathbb{P}_\beta(Z_i|t_i) = \frac{\lambda(t_i|Z_i)}{\sum_{j:t_j \geq t_i} \lambda(t_i|Z_j)} = \frac{\exp(\beta * Z_i)}{\sum_{j:t_j \geq t_i} \exp(\beta * Z_j)}$$

- Numerator: instantaneous hazard for individual i at time t_i
- Denominator: sum of instantaneous hazards for all at-risk patients at t_i

Parameter estimation

We want to find $\mathbb{P}(Z_i|t)$: probability that individual i experiences the event at time t

Individual i's contribution to model likelihood:

$$\mathcal{L}_i(\beta) = \mathbb{P}_\beta(Z_i|t_i) = \frac{\lambda(t_i|Z_i)}{\sum_{j:t_j \geq t_i} \lambda(t_i|Z_j)} = \frac{\exp(\beta * Z_i)}{\sum_{j:t_j \geq t_i} \exp(\beta * Z_j)}$$

- Numerator: instantaneous hazard for individual i at time t_i
- Denominator: sum of instantaneous hazards for all at-risk patients at t_i

Partial likelihood function (for non-censored patients $\delta_i = 1$):

$$\mathcal{L}(\beta) = \prod_{i:\delta_i=1} \mathcal{L}_i(\beta) = \prod_{i:\delta_i=1} \frac{\exp(\beta Z_i)}{\sum_{j:t_j \geq t_i} \exp(\beta Z_j)}$$

Parameter estimation

Partial likelihood function

$$\mathcal{L}_p(\beta) = \prod_{i=1}^{n} \Big(\frac{\exp(\beta^T Z_i)}{\sum_{l \in R^{Cox}(T_i)} \exp(\beta^T Z_l)} \Big)^{\delta_i} \tag{12}$$

- δ_i: event indicator (1 = event, 0 = censored)
- $R^{Cox}(T_i)$: risk set at time T_i and $R^{Cox}(t) := \{l, l = 1, ..., n : T_l \geq t\}$

Properties for $R^{Cox}(t)$:

- Includes all subjects still at risk
- Dynamic - changes over time
- Accounts for censoring

Partial log-likelihood for maximization $l(\beta) = \log(\mathcal{L}(\beta))$ with Breslow algorithm (1970)

In practice

```r
bladder_v1$treatment <- factor(bladder_v1$treatment)
summary(coxph(Surv(stop, recidive) ~ treatment + number,
          data = bladder_v1))
```

Q: What is the treatment effect? Can we quantify it?

Model assumptions

Non-parametric baseline hazard:

- No distributional assumptions
- Can take any form
- Common to all subjects

Covariate Effects:

- Additive on log-hazard scale
- Linear relationship
- Time-independent

In practice

```r
resMart <- residuals(coxph(Surv(stop, recidive) ~ treatment + number,
            data = bladder_v1),
            type = "martingale")
plot(bladder_v1$number,
    resMart,
    main = "Martingale-residuals for number",
    xlab = "Number",
    ylab = "Residus",
    pch = 20)
lines(loess.smooth(bladder_v1$number, resMart), lwd = 2, col = "blue")
```

Q : What do you think?

Proportional Hazards Assumption

The Cox model is known as a **proportional hazards model**, which assumes that the ratio of risks remains constant over time

$$\frac{\lambda(t|Z_1, ..., Z_j, ...Z_p)}{\lambda(t|Z_1, ..., 0, ...Z_p)} = \exp(\beta_j Z_j) \tag{13}$$

i.e., the rate is constant over time

Verifications:

- Schoenfeld residuals (1982)
- Grambsch-Therneau test (1994)

In practice

```
        cox.zph(coxph(Surv(stop, recidive) ~ treatment + number
            ,
        data = bladder_v1))
```

Q : What can we conclude about the proportionality of risks?

Proportional Hazards Assumption

If PHA is not met:

- **Stratify** the model on the variable which does not respect the hypothesis by using the strata option
- Include an **interaction** between time and the variable which does not respect the hypothesis

Parametric estimations

Parametric estimations

Useful when:

- Prior information about event time distribution exists
- Extrapolation is needed

Exponential Model

- Constant hazard: $\lambda(t) = \lambda$
- Piecewise constant: $\lambda(t) = \lambda_j$ for $s_{j-1} \leq t < s_j$
- Intervals: $0 = s_0 < s_1 < ... < s_J = \infty$
- Basis for simple occurrence/exposure rates

Weibull Model

- Time-varying hazard: $\lambda(t) = \lambda \alpha t^{\alpha-1}$
- More mathematically flexible
- Can model:
  - Increasing hazard
  - Constant hazard
  - Decreasing hazard

Overall

Survival estimators – Overview

Type	Method	Key Characteristics
Non-Parametric	Kaplan-Meier	No distribution assumptions Directly estimated from data
Semi-Parametric	Cox Proportional Hazards Model	Includes multiple covariates No baseline hazard specification
Parametric	Exponential Weibull	Assumes specific distribution Constant or changing hazard rate