Survival Analsysis

Artificial Intelligence for Digital Health (AID)

M.Sc. in Digital Health – University of Pisa

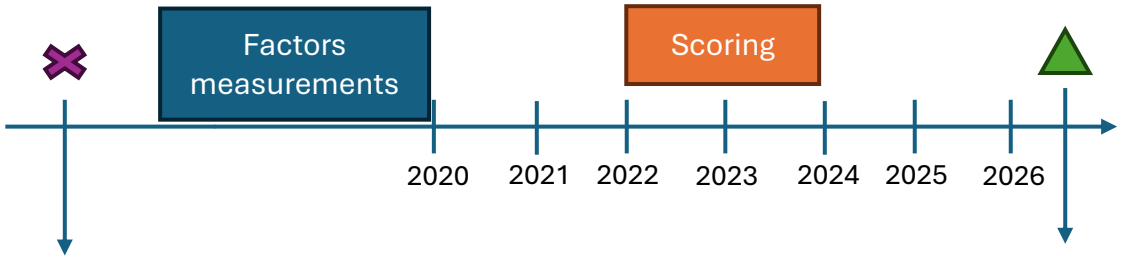
Davide Bacciu (davide.bacciu@unipi.it)

Lecture Outline

- Formalizing survival analysis as a regression problem
- Survival function estimation with baseline statistical models
  - Kaplan-Meier
  - Cox regression
- A broader view into machine learning for survival analysis
  - Neural networks for survival analysis
  - Survival trees

Previous lecture: risk stratification as (binary) classification

Right-censoring makes tackling the problem as a classification one specially challenging



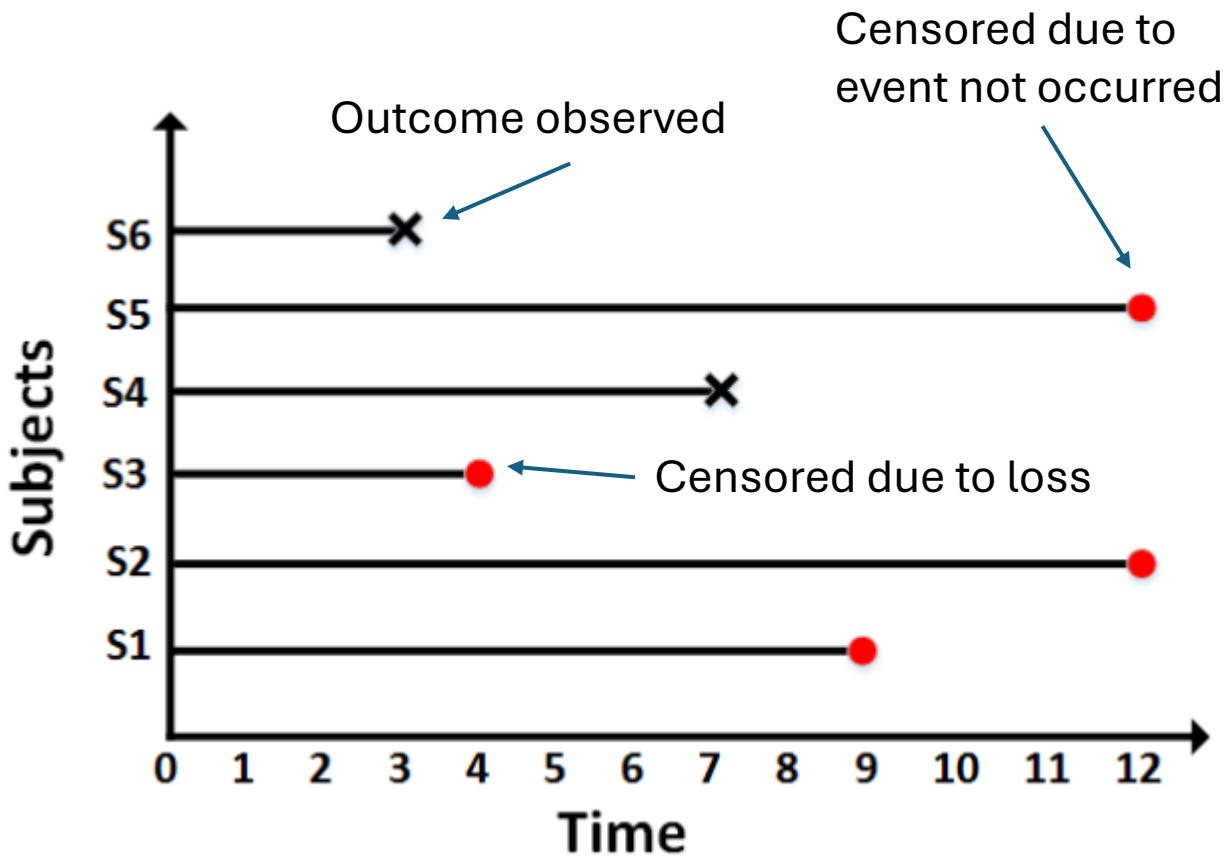
Left censoring

The observations of some factors for some subjects may not be available (e.g. it was collected in a different hospital)

Outcomes for some patient may not have materialized yet (so labelling of the outcome is not available)

Right censoring

Alternative framing: Survival Modelling



- Shift focus on predicting the time-to-event (or outcome) rather than event occurrence
- Change a classification problem with a regression one
- Advantages over classification
  - More training data retained
  - More fine-grained predictions

Survival Analysis Fundamentals

Time-to-Event Outcomes

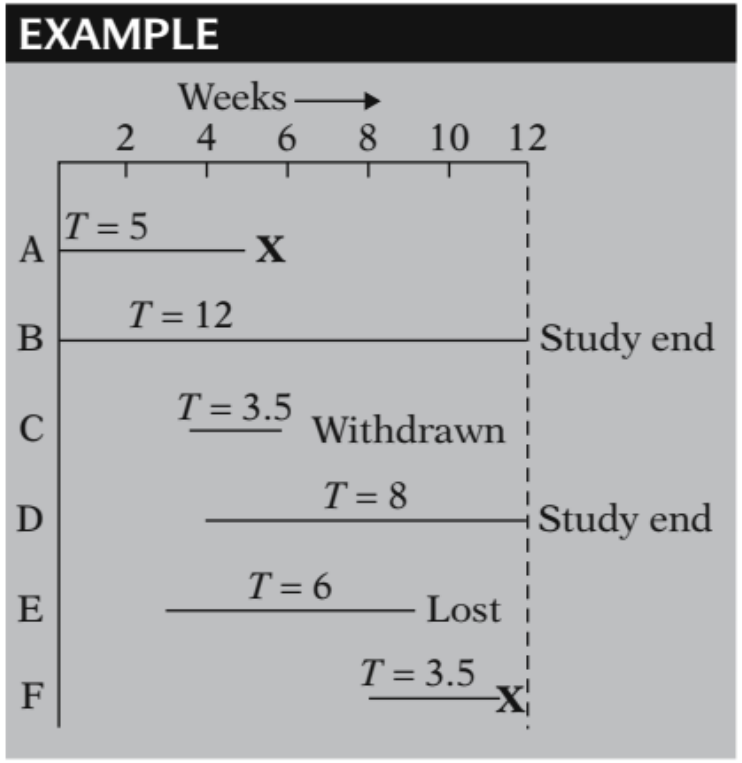
- In survival analysis the outcome variable of interest T is time until an event occurs
- Time: years, months, weeks, or days from the beginning of follow-up of a subject until an event occurs (sometimes also the age of an individual)
- Event: any designated experience of interest that may happen to an individual in our study (death, disease incidence, relapse from remission, recovery, …)

Examples

- Leukemia patients/time in remission (weeks)
- Disease-free cohort/time until heart disease (years)
- Heart transplants/time until death (months)

Right-censoring

Censoring occurs when we have some information about individual survival time, but we don't know the survival time exactly



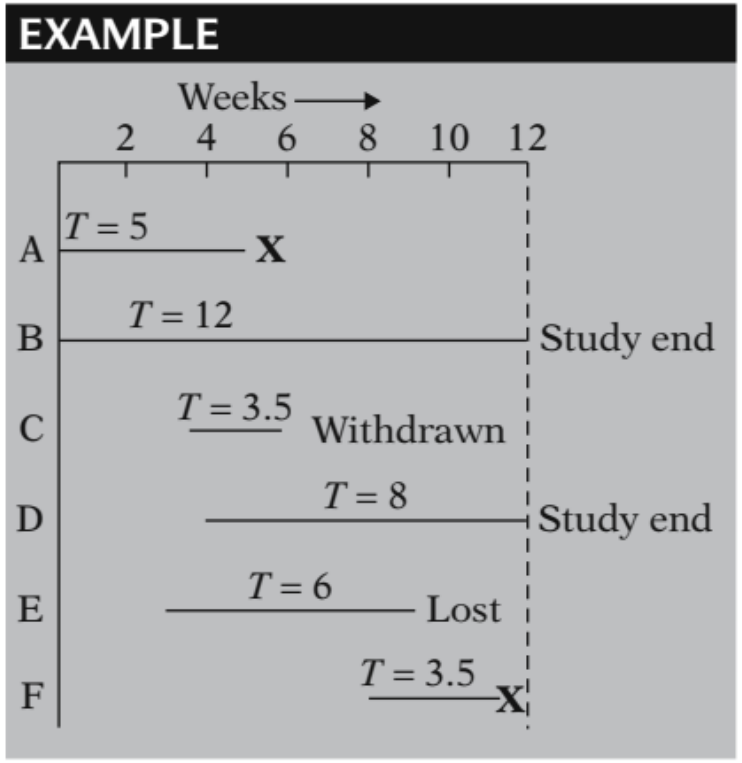
Source: Kleinbaum & Klein, 2005

Three main reasons for censoring:

1. Subject does not experience the event before the study ends
2. Subject is lost to follow-up during the study period
3. Subject withdraws from the study because of death (if death is not the event of interest) or some other competing risk (e.g., adverse drug reaction)

Right-censoring

Censoring occurs when we have some information about individual survival time, but we don't know the survival time exactly



Outcome data

In jargon, failed = outcome manifested

Person	T Survival time	Failed (1); censored (0)
A	5	1
B	12	0
C	3.5	0
D	8	0
E	6	0
F	3.5	1

Source: Kleinbaum & Klein, 2005

Terminology and Notation

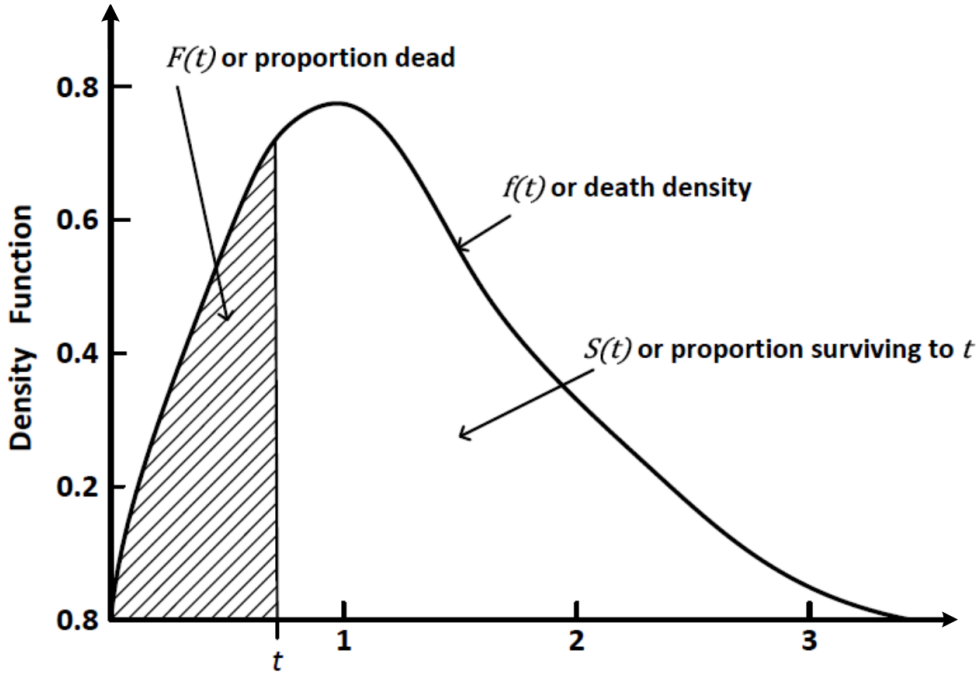
- $T \geq 0$ random variable for a subject survival time
- t is a specific value of time
- $f(t)$ and $F(t)$ denote the density and cumulative density (failure function) of T

$$F(t) = P(T \leq t) = \int_0^t f(\tau)d\tau$$

- We are interested in the survival (survivor) function

$$S(t) = P(T > t) = \int_t^\infty f(\tau)d\tau = 1 - F(t)$$

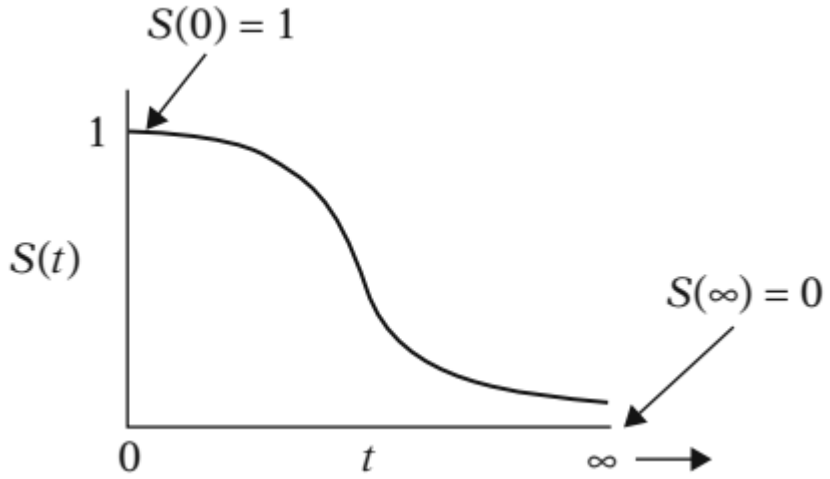
The probability that a subject survives beyond a particular time t



Source: Wang et al, 2017

Images from: Kleinbaum & Klein, 2005

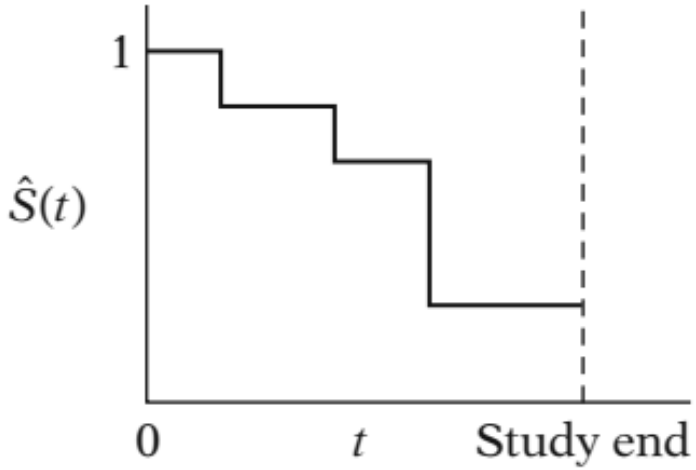
Survival Function



- Monotonic, nonincreasing
- Equals 1 at time 0 and decreases to 0 as time approaches infinity

Clearly, this is **purely theoretic**

- In practice, we typically obtain graphs that are step functions
- Study period is never infinite in length and the estimated survivor function may not go to zero at the end of the study



Hazard Function

Hazard $h(t)$ - Instantaneous "probability" per unit time that an event occurs exactly at time t given that the patient has survived at least until time t

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t}$$

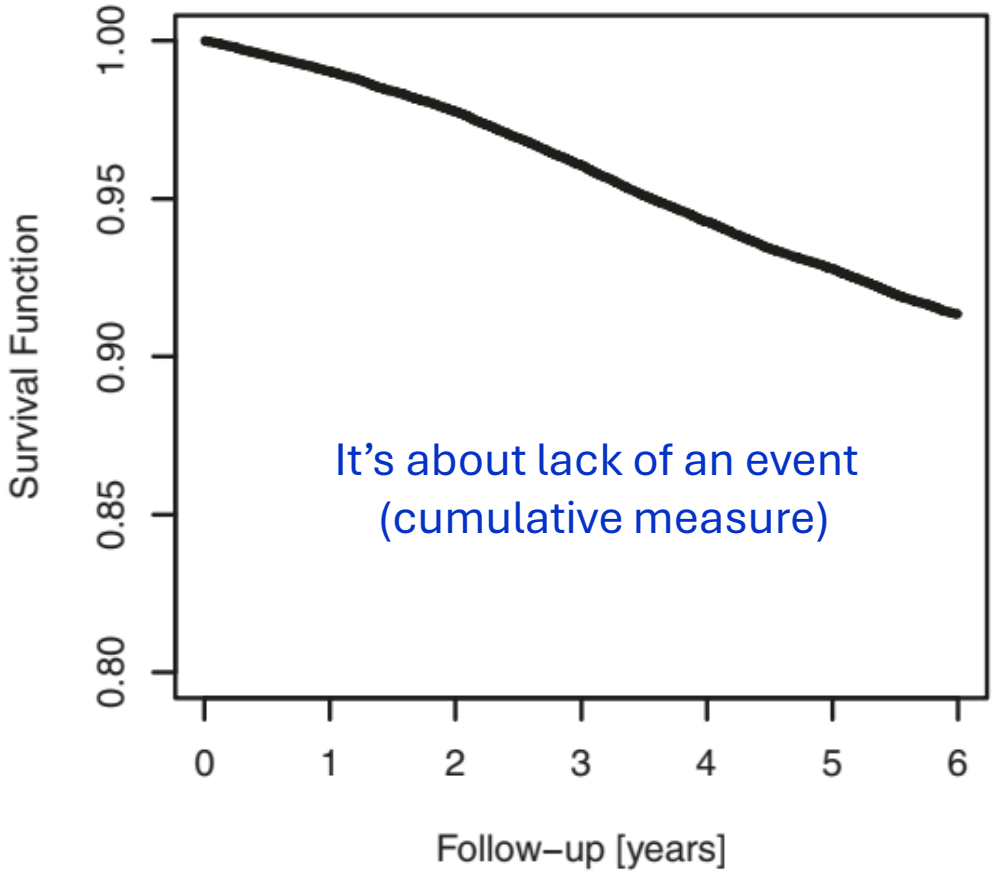
Also known as velocity of the failure function or conditional failure rate (scale for this ratio ranges in $[0, \infty)$, and depends on the unit of time being considered)

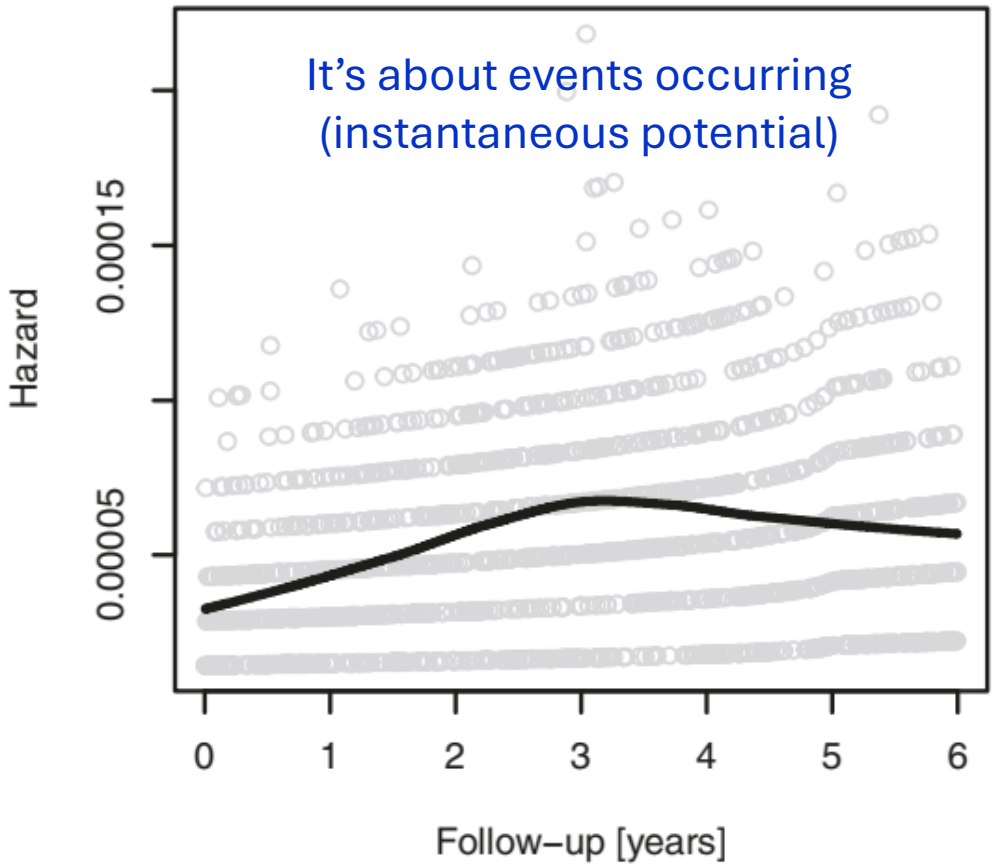
We also have the cumulative hazard

$$H(t) = \int_0^t h(\tau) d\tau \quad \text{s.t. } S(t) = \exp(-H(t))$$

Survival Vs Hazard

While the survivor function is more naturally appealing for analysis of survival data, the survival model is usually written in terms of the hazard function





Source: Simon & Alifieris, 2024

Estimating the Survival Function

Baseline Survival Estimators

Ordered failure times t_j	# at risk n_j	# of failures d_j	# censored in $[t_j, t_{j+1})$ c_j
0	21	0	0
6	21	3	1
7	17	1	1
10	15	1	2
13	12	1	0
16	11	1	3
22	7	1	0
23	6	1	5
	0		

- We can compute the average or median survival time for our reference populations
  - Mean survival of placebo group $\rightarrow \frac{182}{21} = 8.7$ weeks
  - Mean survival of treatment group $\rightarrow \frac{359}{21} = 17.1$ weeks
- These estimates ignore censored subjects
  - They likely were in remission for even longer
  - Underestimates their remission duration
- We can use average hazard rate to account for censoring instead

$$\bar{h} = \frac{\sum_j \delta_j}{\sum_j T_j} \text{ where } \delta_j = \begin{cases} 1 \text{ if outcome occurred} \\ 0 \text{ if censored} \end{cases}$$

Kaplan-Meier (KM) Method

- Nonparametric estimator: more effective when no-assumptions on event time distribution or proportional hazard can be made
- Main intuition: survival probability at time t_j is a product of the same estimate up to the previous time t_{j-1} and the observed survival rate at t_j
- Let's put it into formulas

$$\hat{S}(t_j) = P(T > t_j) = \hat{S}(t_{j-1}) \times P(T > t_j | T > t_{j-1})$$

- This is a recursive formulation in time which, through some mathematical manipulation yields to the final KM estimator

$$\hat{S}(t_j) = \prod_{i \leq j} (1 - \hat{h}_i) = \prod_{i \leq j} \left(1 - \frac{d_i}{n_i}\right) \quad n_i = n_{i-1} - d_{i-1} - c_{i-1}$$

where d_i is the number of events at time t_i and n_i is the number of subjects at risk at time t_i and c_i is the number of censored

Computing KM on Censored Data

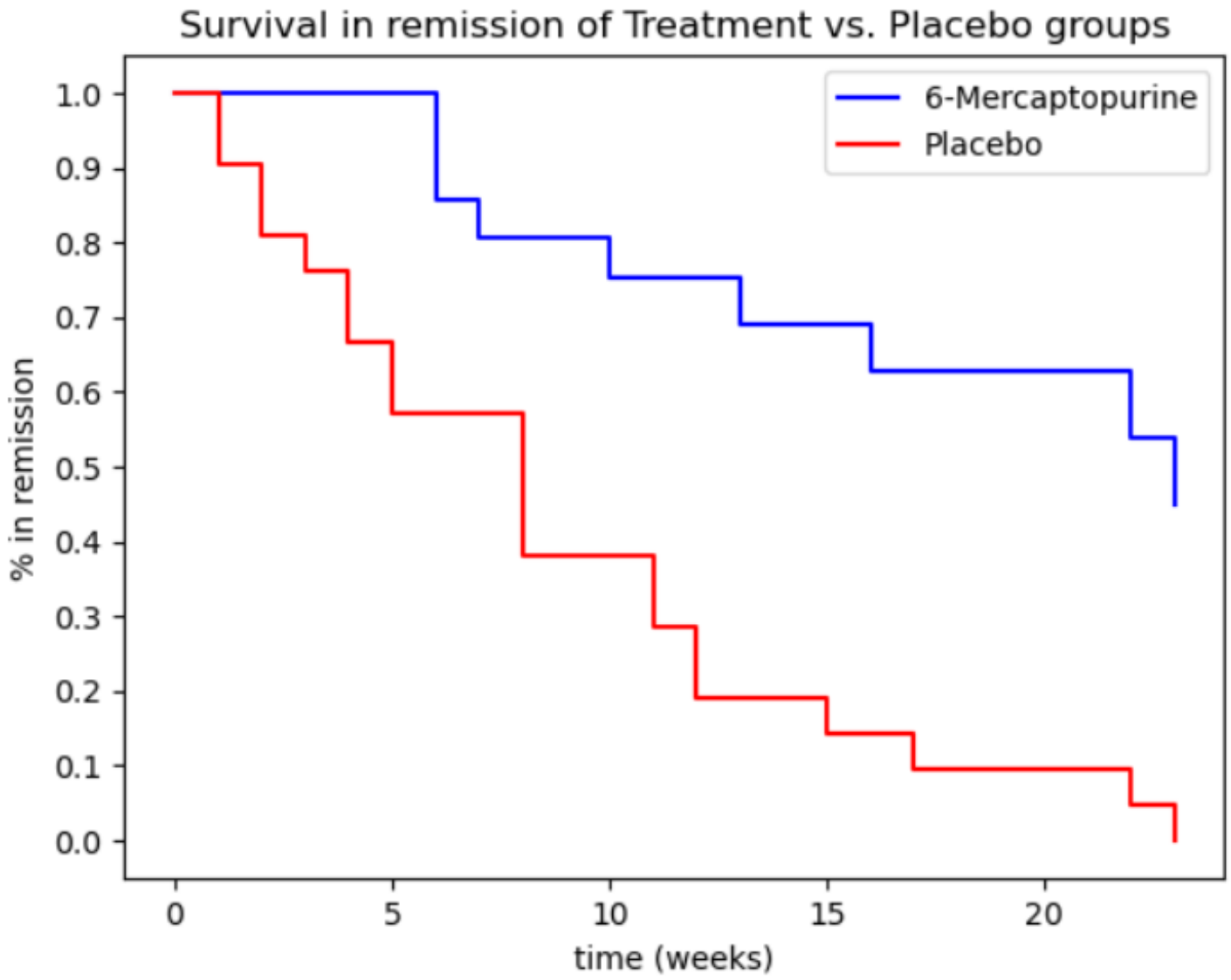
Ordered failure times t_j	# at risk n_j	# of failures d_j	# censored in $[t_j, t_{j+1})$ c_j	$\hat{S}(t_j)$
0	21	0	0	1
6	21	3	1	1 x 18/21 = .8571
7	17	1	1	.8571 x 16/17 = .8067
10	15	1	2	0.8067 x 14/15 = .7529
13	12	1	0	.7529 x 11/12 = .6902
16	11	1	3	0.6902 x 10/11 = .6275
22	7	1	0	.6275 x 6/7 = .5378
23	6	1	5	.5378 x 5/6 = .4482
	0			

$$\hat{S}(t_j) = \prod_{i \leq j} \left(1 - \frac{d_i}{n_i}\right)$$

$$n_i = n_{i-1} - d_{i-1} - c_{i-1}$$

- d_i number of events
- n_i number of subjects at risk
- c_i number of censored

Comparing Treatment Group Vs Placebo Survival



Plot credit: P. Szolovits @ MIT

Confidence intervals for the survival curves

- Greenwood's formula is a common method for directly estimating the confidence interval of the log-survival function (there are many more)

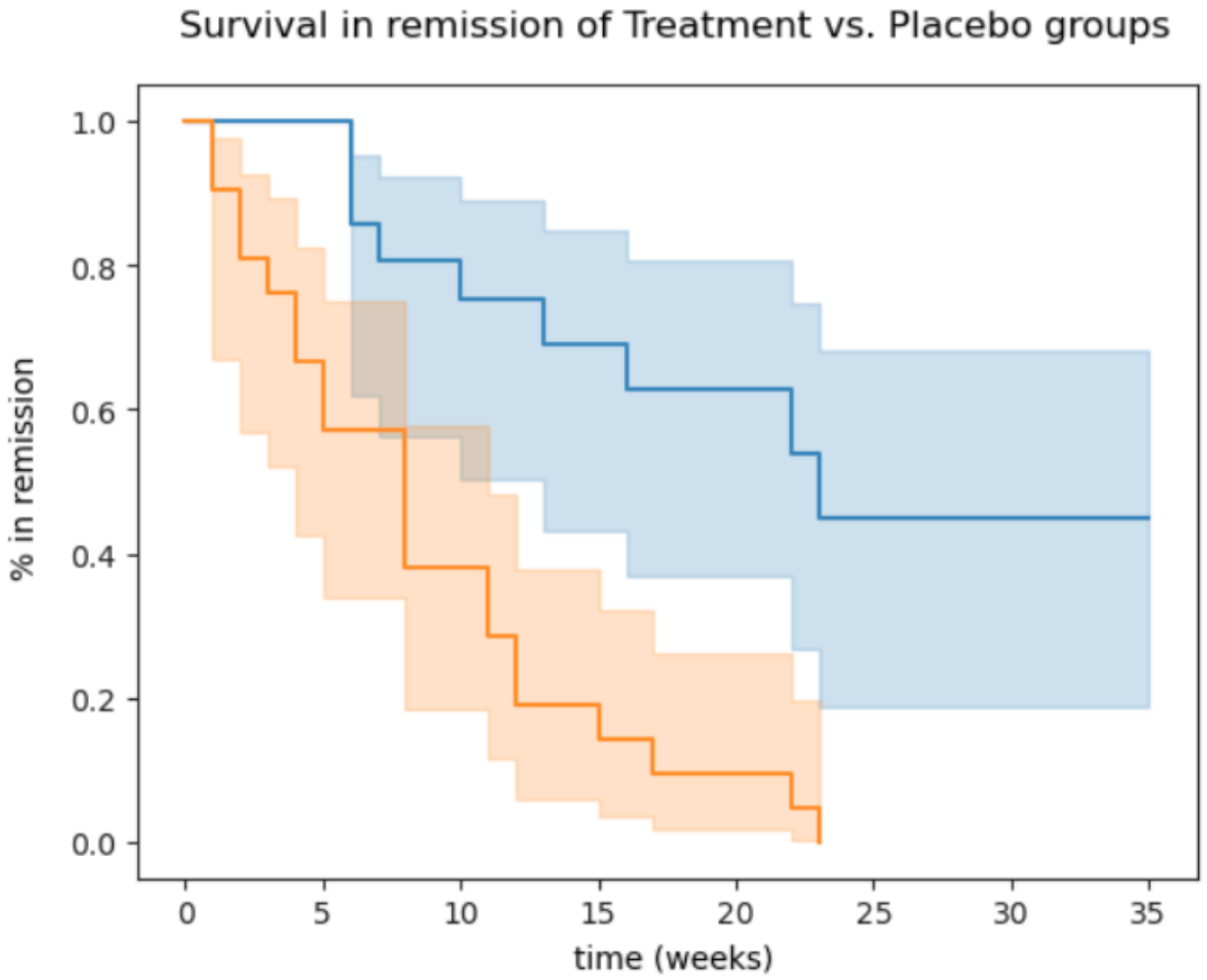
$$Var\left(\log \hat{S}(t_j)\right) = \sum_{i \leq j} \frac{d_i}{n_i(n_i - d_i)}$$

- Variance of the (non-log) survival function can be obtained by the delta method

$$\hat{S}(t_j) = z \sqrt{\hat{S}(t_j)^2 \sum_{i \leq j} \frac{d_i}{n_i(n_i - d_i)}}$$

where z is the normal quantile corresponding to the confidence level (ie. for 95% confidence $z = 1.96$)

Survival curves with confidence intervals



Plot credit: P. Szolovits @ MIT

Comparing Survival Curves

- Consider a group variable which divides the population into G groups: assess association between grouping and survival at each time t_j
- Consider for simplicity G=2, the expected number of events in group 1 at time t_j

$$e_{1j} = \frac{n_{1j}}{n_{1j} + n_{2j}}\left(d_{1j} + d_{2j}\right)$$

where n_{gj} denotes the number of subjects in group g at t_j (similarly for d_{gj})

- The log-rank test statistics is

$$Z = \frac{\left(O_g - E_g\right)^2}{Var\left(O_g - E_g\right)} \quad with\ O_g = \sum_j d_{gj}\ , E_g = \sum_j e_{gj}$$

under the null hypothesis "no difference between survival curves" then $Z \sim \chi^2$

Cox Proportional Hazards Model

- Let us reintroduce a regression model as it allows assessing the effect of covariates on the hazard and making predictions
- Cox proportional hazard is a semiparametric model multiplying a nonparametric baseline hazard h_0 (function of time) to the covariates/features $\boldsymbol{x}_i$ effect (time invariant)

$$h_i(t) = h_0(t)\exp(\boldsymbol{\theta}\boldsymbol{x}_i)$$

  - $h_i(t) \rightarrow$ hazard of the *i*th subject at time t
  - $h_0(t) \rightarrow$ baseline hazard (shared between subjects)
  - $\boldsymbol{\theta} \rightarrow$ regression coefficients

Cumulative Hazard and Survival

- Cumulative hazard can be obtained from integration of the hazard potential

$$H_i(t) = H_0(t)\exp(\boldsymbol{\theta}\boldsymbol{x}_i) = \int_0^t h_0(\tau)d\tau \exp(\boldsymbol{\theta}\boldsymbol{x}_i)$$

- The survival function then reads as

$$S_i(t) = \exp(H_0(t)\exp(\boldsymbol{\theta}\boldsymbol{x}_i)) = S_0(t)^{\exp(\boldsymbol{\theta}\boldsymbol{x}_i)}$$

where the baseline survival function is

$$S_0(t) = \exp\big(H_0(t)\big)$$

Estimating the Cox Model

- Computig the Cox proportional hazard requires (1) fitting the regression parameters $\boldsymbol{\theta}$ and (2) estimating the cumulative baseline hazard (nonparametric)

Estimating the Cox Model (1)

- Regression parameters are found as a minimization problem of the following loss (log-partial likelihood)

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{j=1}^{N} \delta_j \left(\boldsymbol{\theta}\boldsymbol{x}_j - \log \left(\sum_{i \in R_j} \exp(\boldsymbol{\theta}\boldsymbol{x}_i) \right) \right)$$

Sums over ordered times t_j

Dirac for censoring

- The update rule for parameters $\boldsymbol{\theta}$ are derived by minimization of $\mathcal{L}(\boldsymbol{\theta})$ (e.g. by some form of gradient descent)
- We can add to $\mathcal{L}(\boldsymbol{\theta})$ all regularization strategies we have seen so far

Estimating the Cox Model (2)

- Compute the cumulative baseline hazard using the regression parameters $\boldsymbol{\theta}$
- **Breslow estimator** (1)

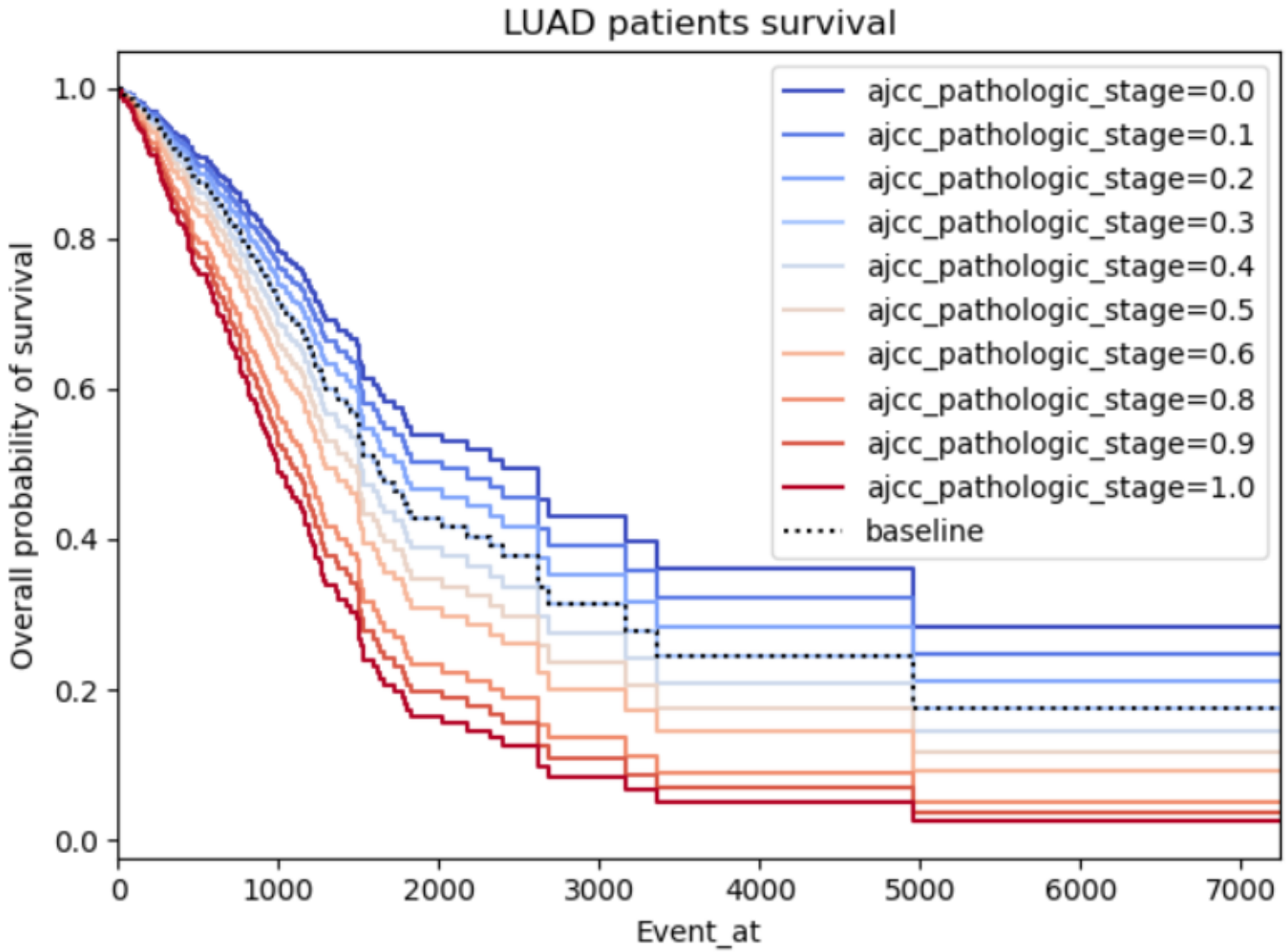
$$H_0(t_j) = \sum_{l \le j} \hat{h}_0(t_l)$$

$$\hat{h}_0(t_j) = \begin{cases} d_j \setminus \left(\sum_{i \in R_j} (\exp(\boldsymbol{\theta} \boldsymbol{x}_i)) \right)^{-1} & \text{if } t_j \text{ is an event time} \\ 0 & \text{otherwise} \end{cases}$$

R_j is the set of subjects at risk at time t_j

$\boldsymbol{x}_i$ is vector for the subject with outcome at time t_i

Cox Survival Curves



Allow to assess the effect of risk factors on survival

Time-Dependent Cox Model

- We can relax the vanilla Cox assumption about $\boldsymbol{x}$ features being time invariant
- The time-dependent Cox model

$$h_i(t) = h_0(t) \exp\left(\sum_{k=1}^{K1} \theta_k x_{ik}(t) + \sum_{k\prime=1}^{K2} \theta_{k\prime} x_{ik\prime} \right)$$

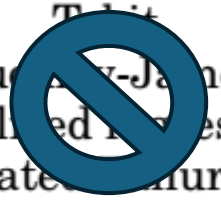
Time-dependent features and their parameters

Time-invariant features and their parameters

Modern Survival Analysis Landscape

Wrapping-up on Statistical Methods

Type	Advantages	Disadvantages	Specific methods
Non-parametric	More efficient when no suitable theoretical distributions known.	Difficult to interpret; yields inaccurate estimates.	Kaplan-Meier Nelson-Aalen Life-Table
Semi-parametric	The knowledge of the underlying distribution of survival times is not required.	The distribution of the outcome is unknown; not easy to interpret.	Cox model Regularized Cox CoxBoost Time-Dependent Cox
Parametric	Easy to interpret, more efficient and accurate when the survival times follow a particular distribution.	When the distribution assumption is violated, it may be inconsistent and can give sub-optimal results.	Tobit Buckley-James Penalized Regression Accelerated Failure Time

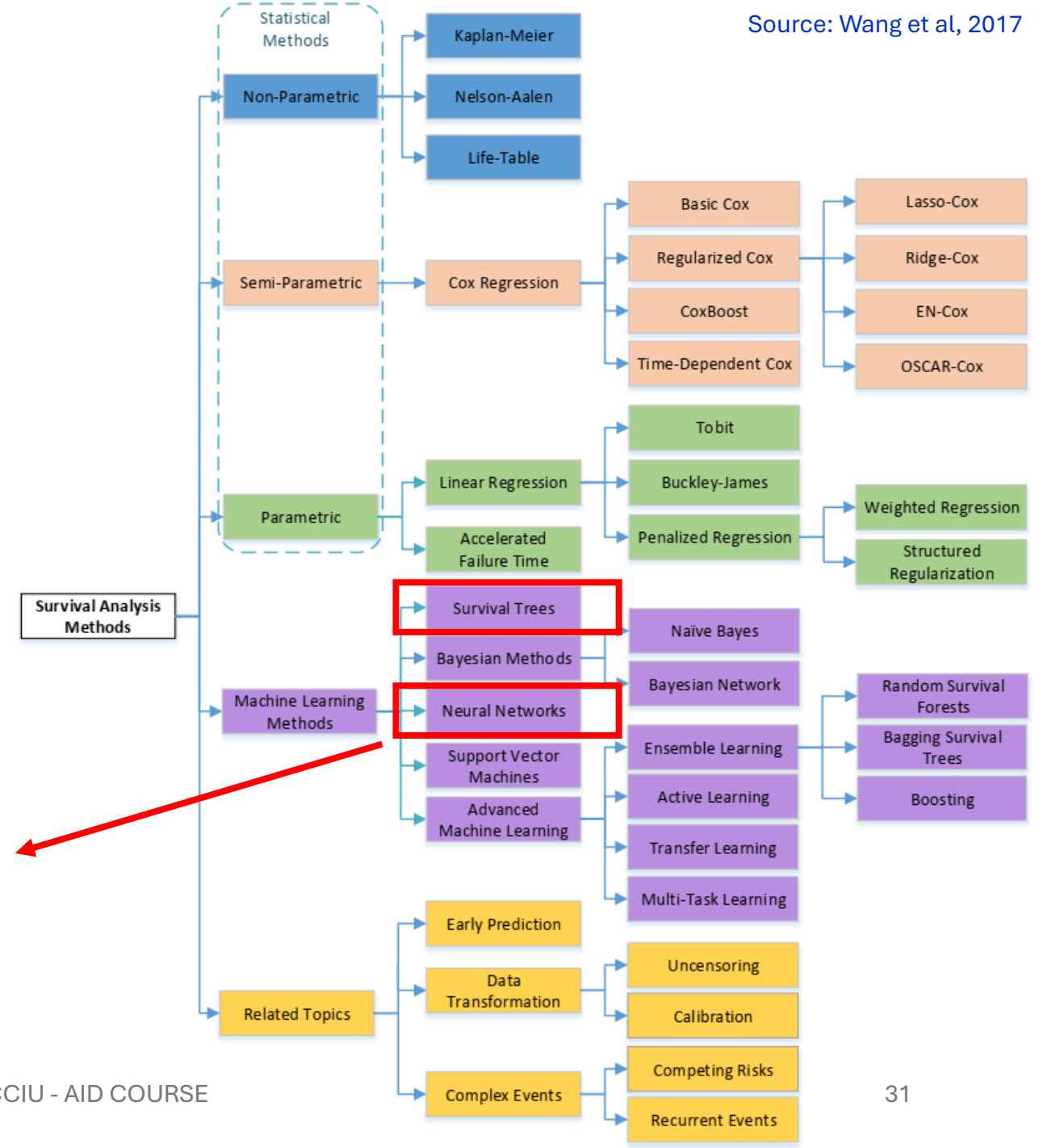


Source: Wang et al, 2017

Taxonomy of survival analysis methods

There are many more survival analysis approaches if we look into machine learning methodologies

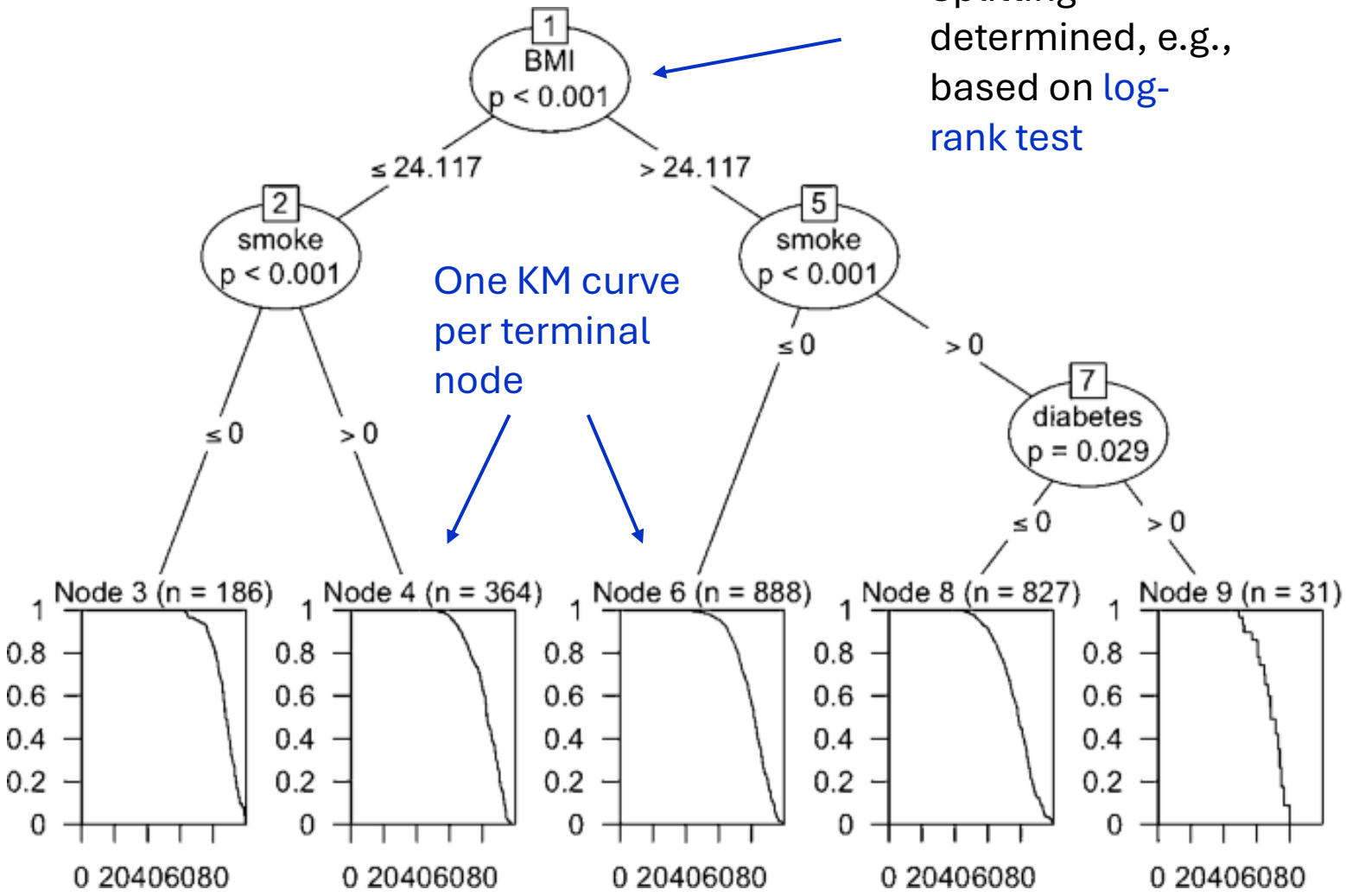
Intuition: take the $\boldsymbol{\theta x}$ regression term in Cox and replace it with a non-linear parameterized function $\mathrm{f}_{\boldsymbol{\theta}}(\boldsymbol{x})$



Survival Trees

- Decision tree adapted for survival analysis
- The goal is to find the best feature and threshold to split the data into homogeneous survival groups
- Can be generalized to Random Survival Forests

Image source: Bertsimas et al, ML 2022



Splitting determined, e.g., based on log-rank test

One KM curve per terminal node

Internal Node Splitting Test

- Compares survival distributions between two or more groups
- Tests if survival curves differ significantly at each potential split
- Selects the split with the largest (log-rank) statistic, ensuring maximal survival time separation
- Statistical hypothesis testing
  - Null Hypothesis (H_0): There is no difference in survival between the groups
  - Alternative Hypothesis (H_1): There is a difference in survival between the groups

Splitting Using Log-Rank Test

1. Calculate the Observed O_i and Expected Events E_i

- For each time point where an event occurs, count
  - The number of individuals at risk in each group
  - The observed number of events in each group
  - The expected number of events under the assumption that survival is the same across groups

2. Compute the Log-Rank statistic based on the difference between the observed O_i and expected events E_i at each time point

- This follows a χ^2 distribution with 1 degree of freedom (for two groups)

3. Determine Significance

- Compare the test statistic to a χ^2 distribution to get a p-value, with a low p-value (e.g., <0.05) indicating a significant difference in survival between the groups

Log Rank Example

Group splitting	N_i	O_i	E_i
$BMI \leq 24.117$	310	161	143,39
$BMI > 24.117$	304	123	140.61
Total	614	284	284

Log-rank test: $\chi_1^2 = 8.2 \rightarrow$ p-value = 0.0042

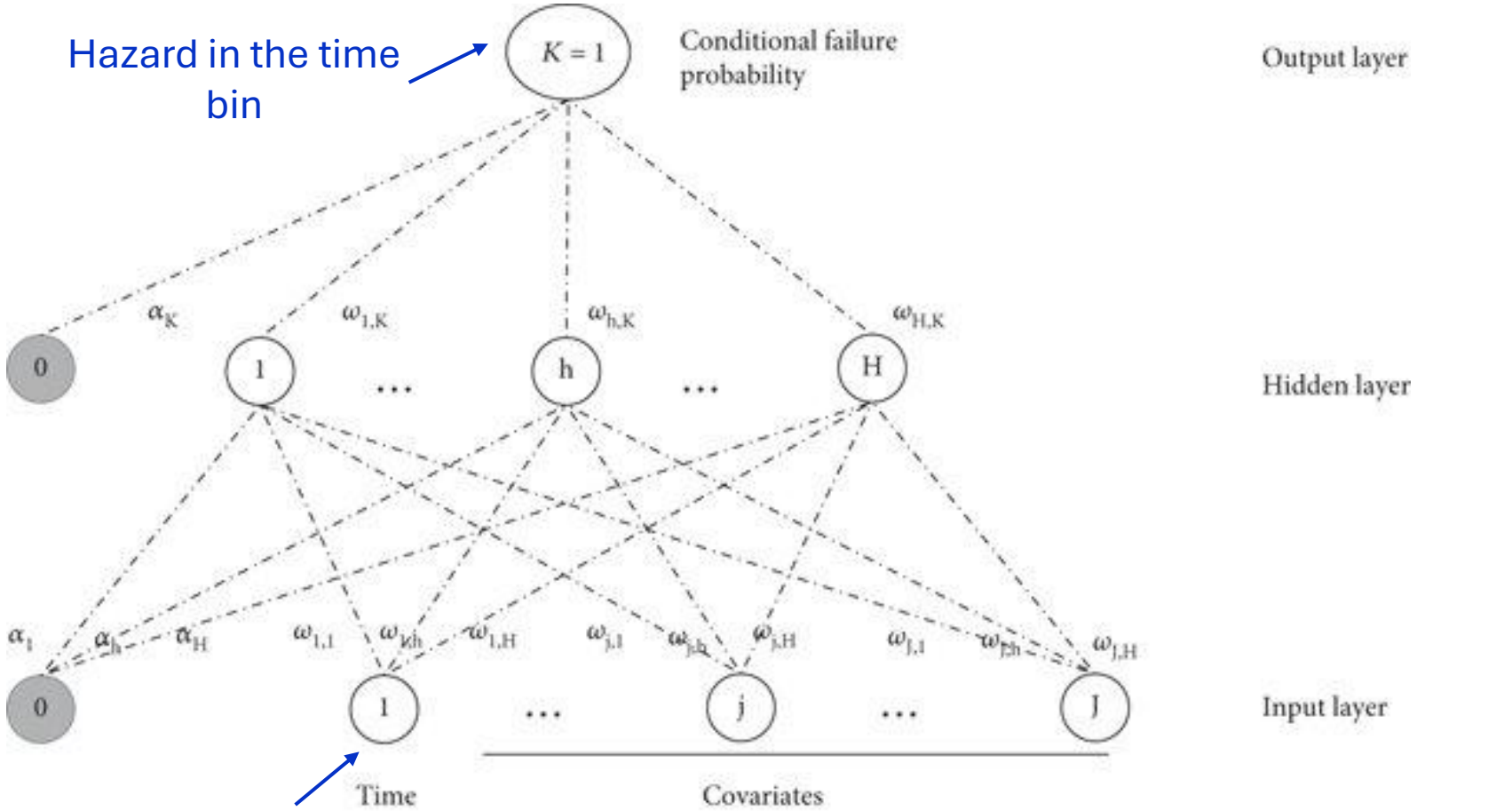
Partial logistic artificial neural networks (PLANN)

Image source: Biganzoli et al, 1998

One of the first (feedforward) NN-based approaches to survival analysis

- Nonlinear model

Later models introduced recurrent and deep architectures



Wrap-up

Take home lessons

- Survival analysis focuses on time-to-event data to understand outcomes
  - Shift focus on predicting the time-to-event rather than event occurrence
- Challenges with survival data are associated to (right) censoring
  - Outcomes may not materialize within the observation window
  - Competing risks further complicate the picture by "masking" outcomes
- Two baseline statistical estimators of survival
  - Kaplan-Meier (non-parametric): useful when no assumption on underlying distribution can be made
  - Cox regression (semi-parametric): allows introducing subject information in a non-parametric baseline
- More recently a broad set of non-linear survival analysis methods based on neural networks have been proposed
  - But there are also survival trees, survival forests, survival support vector machines
  - All nicely implemented for you (in the R language and else)

Next Lecture(s)

Introduction to Bayesian networks

- Graphical formalism
  - Structure and components of Bayesian networks
  - Random variables and conditional independence
  - Factorized distributions
- Relevant graphical substructures
- Reasoning graphically on conditional independence
- Learning in Bayesian Networks
- Applications in healthcare