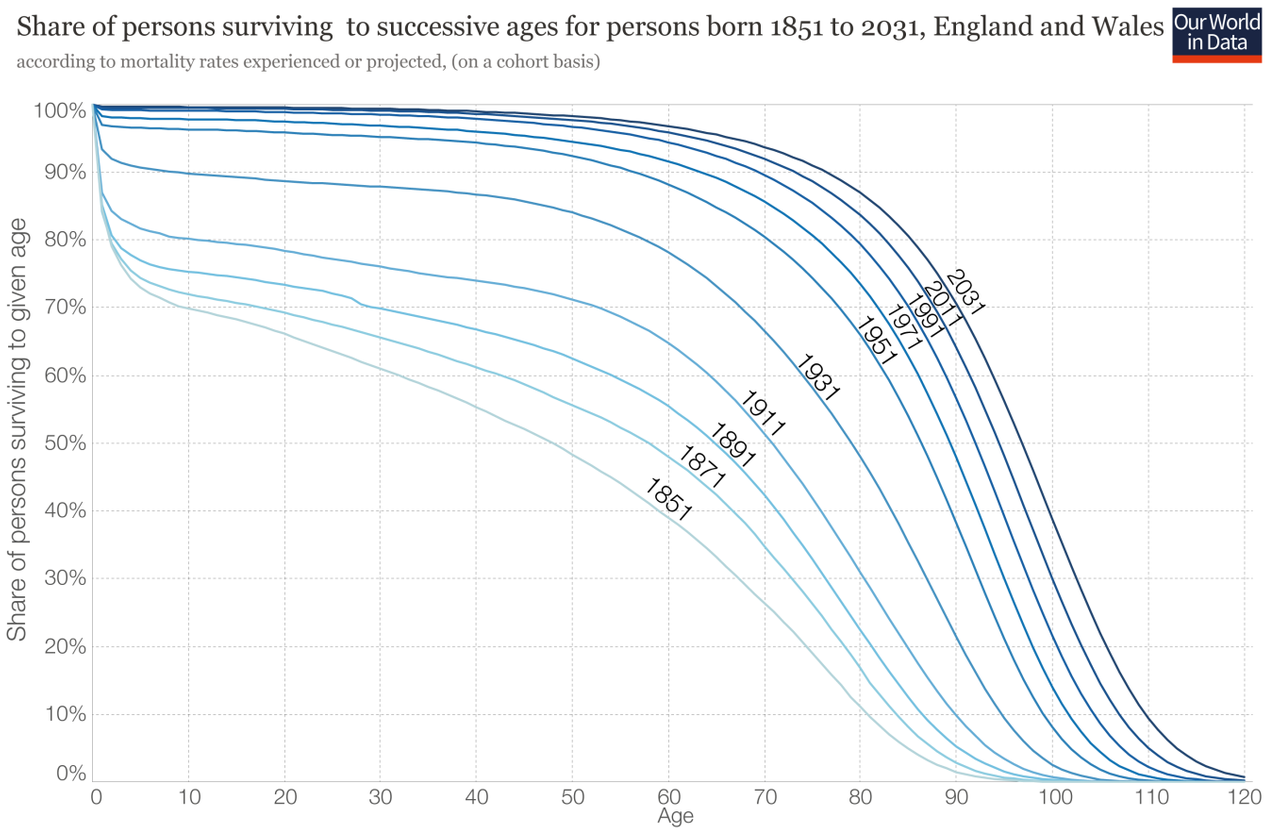
Survival Analysis, Censoring, Proportional Hazard Models

Peter Szolovits

- based on
  - Kleinbaum DG, Klein M. Survival analysis: a self-learning text. 2nd ed. New York, NY: Springer; 2005. Available (free) via MIT-Springer: https://link.springer.com/book/10.1007/0-387-29150-4 [beware typos]
  - lecture by David Sontag

Massachusetts Institute of Technology

You've Seen Survival Models Already



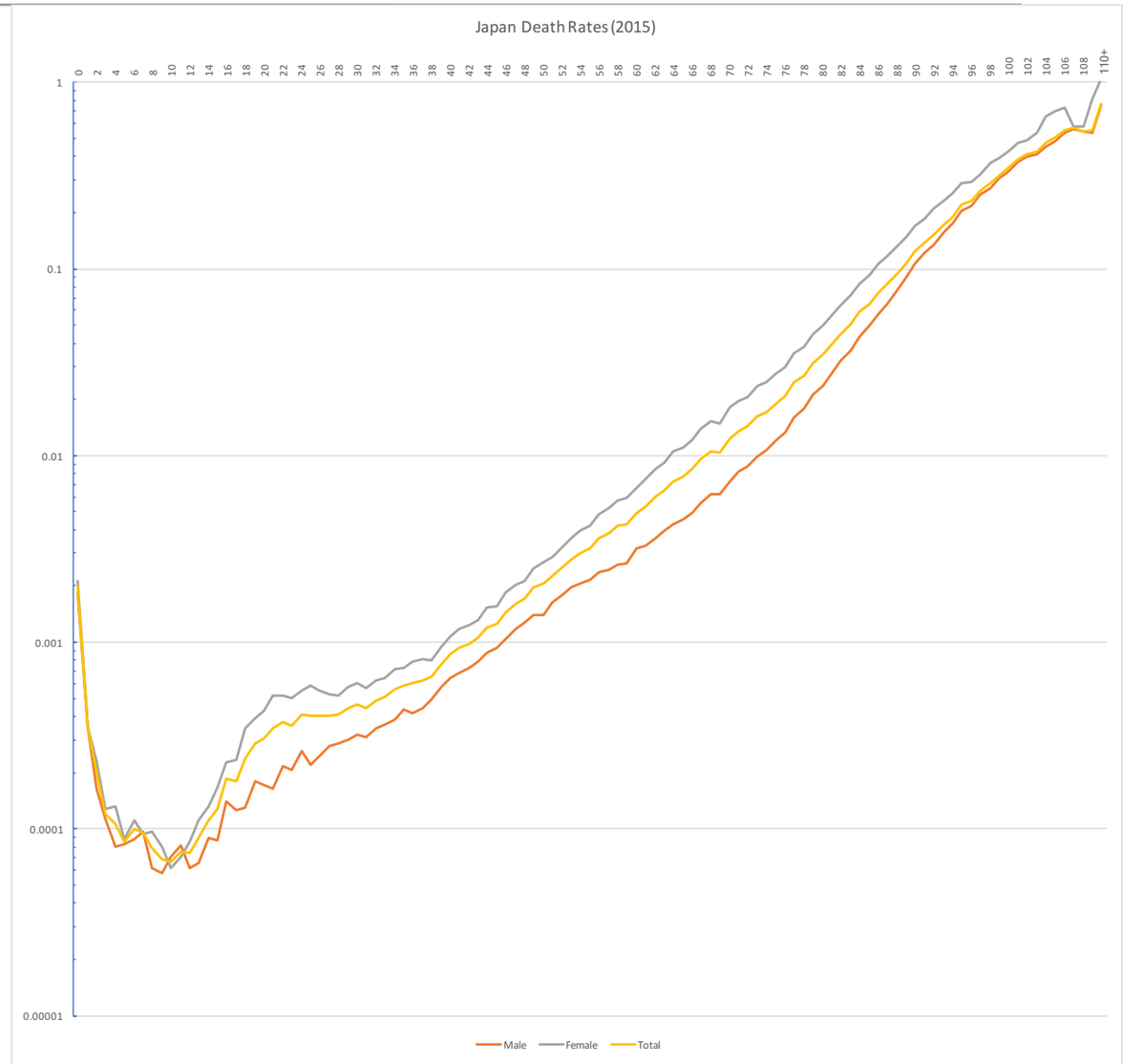
Share of persons surviving to successive ages for persons born 1851 to 2031, England and Wales

Our World in Data

according to mortality rates experienced or projected, (on a cohort basis)

Data source: Office for National Statistics (ONS). Note: Life expectancy figures are not available for the UK before 1951; for long historic trends England and Wales data are used

The interactive data visualization is available at OurWorldinData.org. There you find the raw data and more visualizations on this topic. Licensed under CC-BY-SA by the author Max Roser.



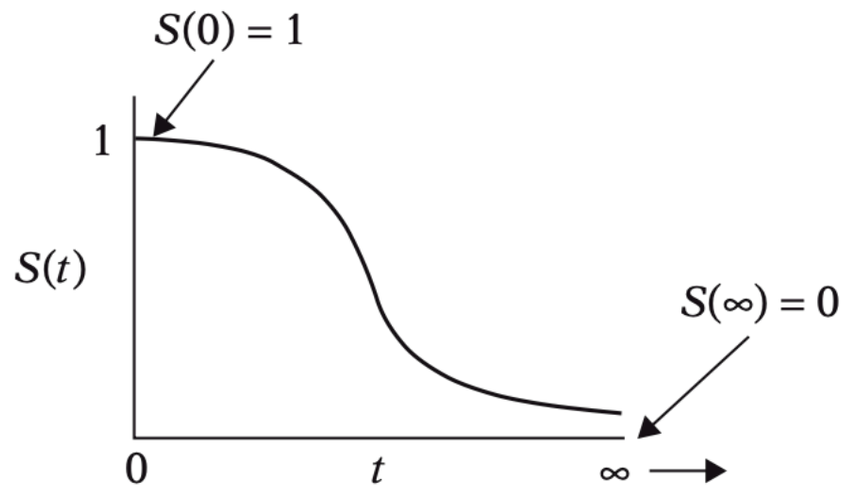
Survival Curve

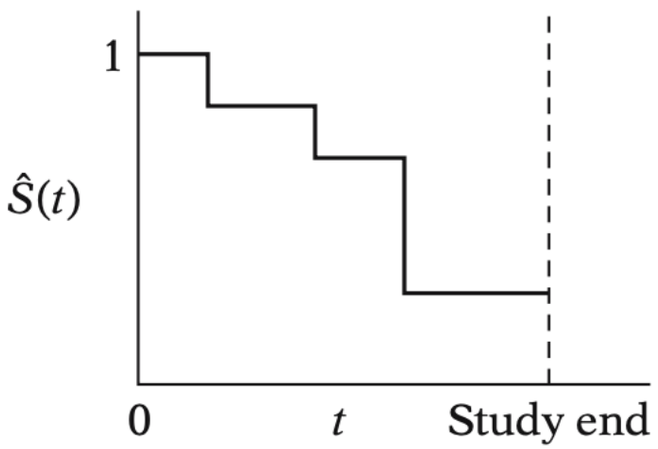
Hazard Curve

- Outcome can be good (recovery from surgery) or bad (death)

Survival and Hazard

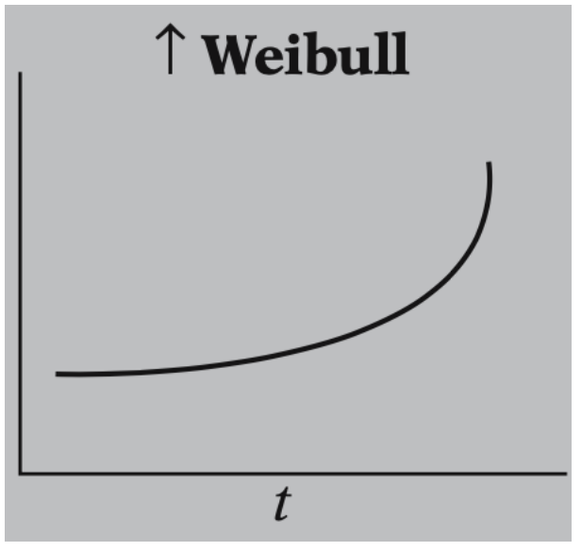
- In absence of censoring, survival is easy to compute:
  - Just fraction of subjects still “alive” at some time
- T is a random variable denoting the survival time of an individual; $T \geq 0$
  t is a specific value of T
- Survival, $S(t) = P(T > t) \in [0,1]$
- Hazard:
  - $h(t) = -\left[\frac{dS(t)/dt}{S(t)}\right] \in [0,\infty]$, or
  - $h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t}$
- Empirical survival curves $\hat{S}(t)$ are step functions
- If instantaneous risk of death (density of S) = $f(t)$,
  $S(t) = P(T > t) = \int_{u=t}^{\infty} f(u)du$

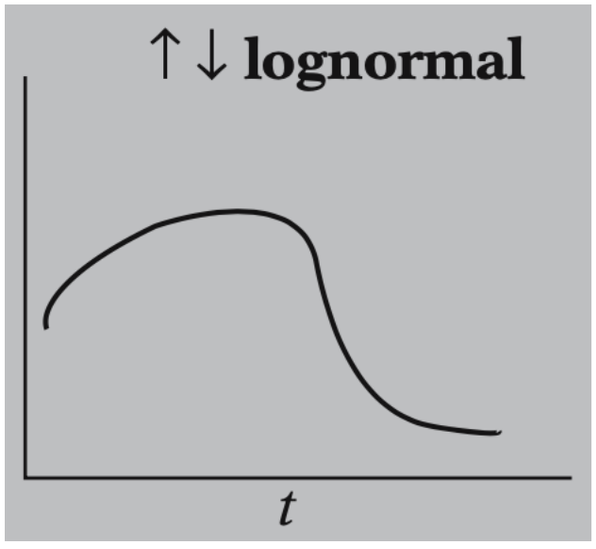


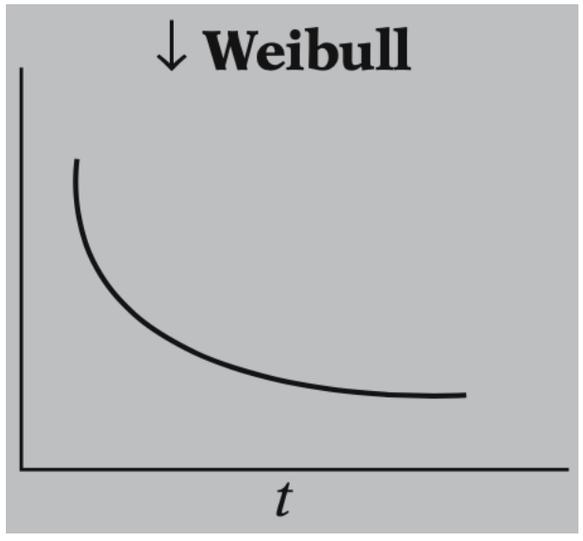


Parametric Hazard Functions
(if we know something about what to expect)

- Exponential: $h(t) = \lambda$, i.e., constant hazard
  - E.g., healthy subject, with constant risk of getting run over, murdered, etc.
  - $S(t) = e^{-\lambda t}$; density function of survival function: $f(t) = \lambda e^{-\lambda t}$
- Increasing Weibull: $h(t) = \lambda\phi t^{\phi-1}$; $S(t) = e^{-\lambda t^{\phi}}$
  - E.g., leukemia patient unresponsive to therapy
- Decreasing Weibull
  - E.g., patient recovering from surgery
- Lognormal: $h(t) = f(t)/S(t)$; $S(t) = 1 - \Phi\{(\ln t - \mu)/\sigma\}$
  - E.g., TB patient







Commonly used parametric survival models

Table IV: Density, Survival and Hazard functions for the distributions commonly used in the parametric methods in survival analysis.

Distribution	**PDF** $f(t)$	**Survival** $S(t)$	**Hazard** $h(t)$
Exponential	$\lambda exp(-\lambda t)$	$exp(-\lambda t)$	λ
Weibull	$\lambda k t^{k-1} exp(-\lambda t^k)$	$exp(-\lambda t^k)$	$\lambda k t^{k-1}$
Logistic	$\frac{e^{-(t-\mu)/\sigma}}{\sigma(1+e^{-(t-\mu)/\sigma})^2}$	$\frac{e^{-(t-\mu)/\sigma}}{1+e^{-(t-\mu)/\sigma}}$	$\frac{1}{\sigma(1+e^{-(t-\mu)/\sigma})}$
Log-logistic	$\frac{\lambda k t^{k-1}}{(1+\lambda t^k)^2}$	$\frac{1}{1+\lambda t^k}$	$\frac{\lambda k t^{k-1}}{1+\lambda t^k}$
Normal	$\frac{1}{\sqrt{2\pi}\sigma} exp(-\frac{(t-\mu)^2}{2\sigma^2})$	$1-\Phi(\frac{t-\mu)}{\sigma})$	$\frac{1}{\sqrt{2\pi}\sigma(1-\Phi((t-\mu)/\sigma))} exp(-\frac{(t-\mu)^2}{2\sigma^2})$
Log-normal	$\frac{1}{\sqrt{2\pi}\sigma t} exp(-\frac{(log(t)-\mu)^2}{2\sigma^2})$	$1-\Phi(\frac{log(t)-\mu}{\sigma})$	$\frac{\frac{1}{\sqrt{2\pi}\sigma t} exp(-(log(t)-\mu)^2/2\sigma^2)}{1-\Phi(\frac{log(t)-\mu}{\sigma})}$

We obtain **conditional** models $f(t \mid \boldsymbol{x}; \beta)$ by letting, e.g., $\lambda = \exp(\boldsymbol{\beta} \cdot \boldsymbol{x})$

Wang P, Li Y, Reddy CK. Machine Learning for Survival Analysis: A Survey. arXiv; 2017. Available from: http://arxiv.org/abs/1708.04649

Illustrative Example of Survival Analysis

- Freireich et al. The Effect of 6-Mercaptopurine on the Duration of Steroid-Induced Remissions in Acute Leukemia: A Model for Evaluation of Other Potentially Useful Therapy. Blood, 21: 699-716, 1963
  - Example from Kleinbaum DG, Klein M. Survival analysis: a self-learning text. 2nd ed. New York, NY: Springer; 2005
- All patients were induced into remission, then half were treated with 6-Mercaptopurine to see whether it helped maintain remission
- Example shows the week in which each patient failed (or was censored, indicated by "+")
- First, we consider Group 2 (placebo)
  - No censored data

EXAMPLE (continued)

Group 1 (Treatment) $n = 21$	Group 2 (Placebo) $n = 21$
6, 6, 6, 7, 10,	1, 1, 2, 2, 3,
13, 16, 22, 23,	4, 4, 5, 5,
6+, 9+, 10+, 11+,	8, 8, 8, 8,
17+, 19+, 20+,	11, 11, 12, 12,
25+, 32+, 32+,	15, 17, 22, 23
34+, 35+	

	# failed	# censored	Total
Group 1	9	12	21
Group 2	21	0	21

Alternative Representations of the Data (for now, just Group 2, no censoring)

	Indiv. (#)	t (weeks)	δ (failed or censored)	X (Group)
GROUP 2	22	1	1	0
	23	1	1	0
	24	2	1	0
	25	2	1	0
	26	3	1	0
	27	4	1	0
	28	4	1	0
	29	5	1	0
	30	5	1	0
	31	8	1	0
	32	8	1	0
	33	8	1	0
	34	8	1	0
	35	11	1	0
	36	11	1	0
	37	12	1	0
	38	12	1	0
	39	15	1	0
	40	17	1	0
	41	22	1	0
	42	23	1	0

- All the placebo patients failed remission by 23 weeks of the trial
- Subjects ordered by failure time
- $\delta = 0$ means censored, $\delta = 1$ means failed
- $X = 0$ means placebo, $X = 1$ means treated by 6-Mercaptopurine
- E.g., patient 32 failed remission in week 8

#	t	δ	X_1	X_2	$\cdots$	X_p
1	t_1	δ_1	X_{11}	X_{12}	$\cdots$	X_{1p}
2	t_2	δ_2	X_{21}	X_{22}	$\cdots$	X_{2p}
$\vdots$						$\vdots$
j	t_j	δ_j	X_{j1}	X_{j2}	$\cdots$	X_{jp}
$\vdots$						$\vdots$
n	t_n	δ_n	X_{n1}	X_{n2}	$\cdots$	X_{np}

Transform to Show How Many Patients Fail (or Are Censored) at Each Time Anyone Fails

Ordered failure times t_j	# at risk n_j	# of failures m_j	# censored in $[t_j, t_{j+1})$ q_j
0	21	0	0
1	21	2	0
2	19	2	0
3	17	1	0
4	16	2	0
5	14	2	0
8	12	4	0
11	8	2	0
12	6	2	0
15	4	1	0
17	3	1	0
22	2	1	0
23	1	1	0
	0		

Group 2 (Placebo) $n = 21$

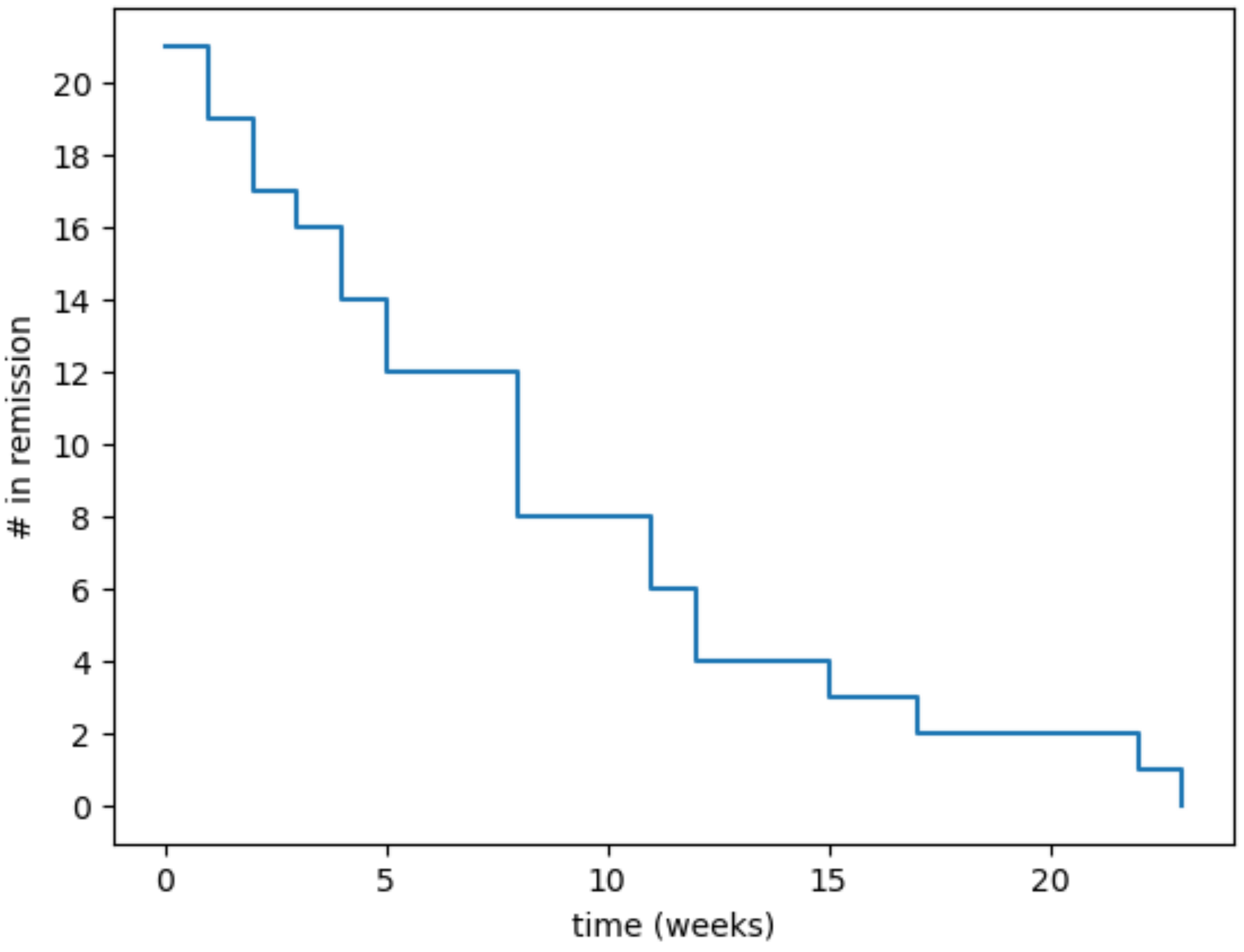
1, 1, 2, 2, 3,

4, 4, 5, 5,

8, 8, 8, 8,

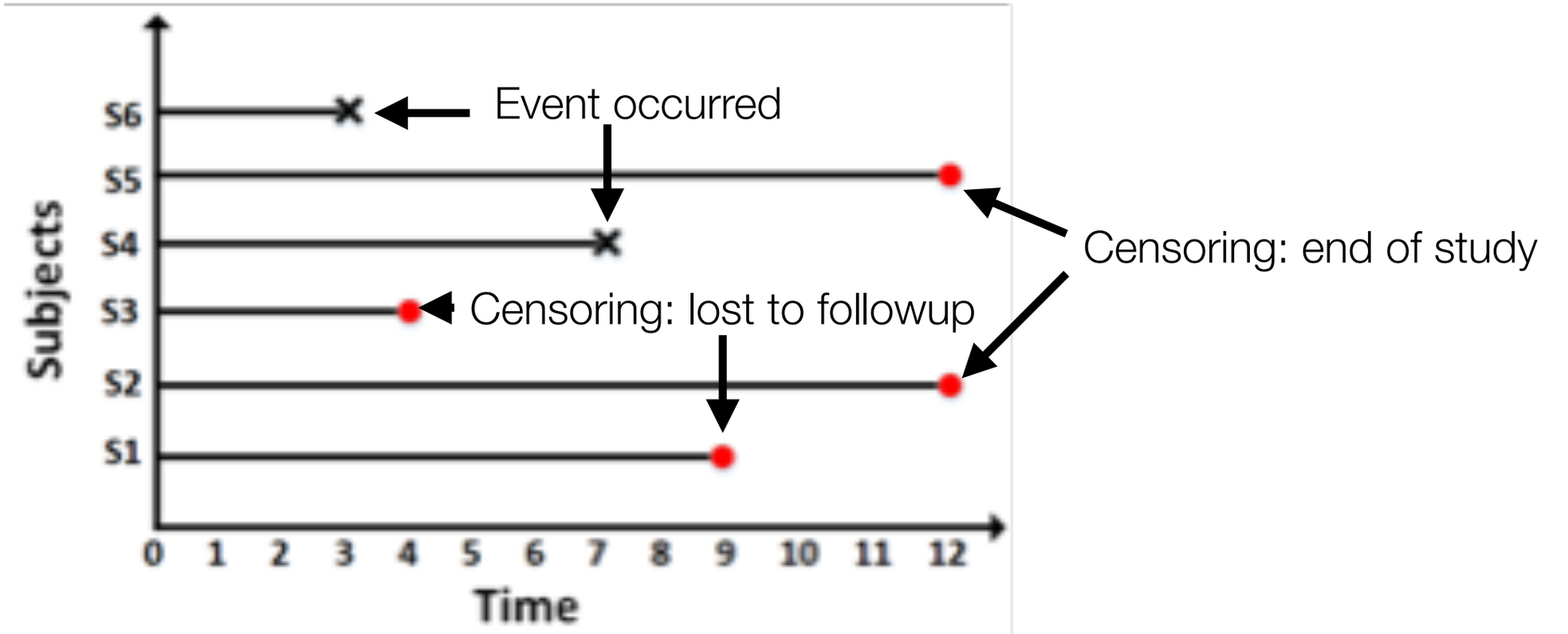
11, 11, 12, 12,

15, 17, 22, 23

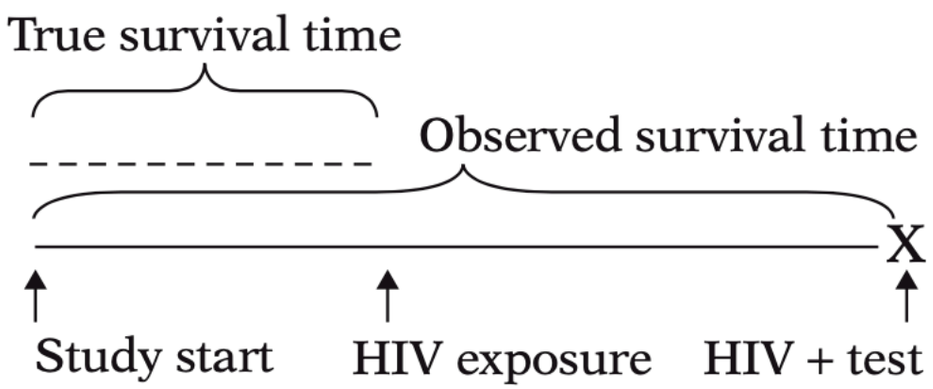

Survival in remission of Placebo group
in remission
0
2
4
6
8
10
12
14
16
18
20
0
5
10
15
20
time (weeks)

Censoring

- Right-censoring: lost track of subject before the event occurs
  - switched insurance carrier, study ended before event, moved away, withdrew



- Left-censoring: event occurs before start of study/data collection/test



- Typical assumption: *censoring is independent of outcome given covariates*

How to Deal with Censored Data: Kaplan-Meier

Ordered failure times t_j	# at risk n_j	# of failures m_j	# censored in $[t_j, t_{j+1})$ q_j
0	21	0	0
6	21	3	1
7	17	1	1

Group 1 (Treatment) $n = 21$

6, 6, 6, 7, 10,

13, 16, 22, 23,

6+, 9+, 10+, 11+,

17+, 19+, 20+,

25+, 32+, 32+,

34+, 35+

How to Deal with Censored Data: Kaplan-Meier

Ordered failure times t_j	# at risk n_j	# of failures m_j	# censored in $[t_j, t_{j+1})$ q_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		

Group 1 (Treatment) $n = 21$

6, 6, 6, 7, 10,
13, 16, 22, 23,
6+, 9+, 10+, 11+,
17+, 19+, 20+,
25+, 32+, 32+,
34+, 35+

Aside: Overall Statistics Already Favor Treatment

- Ignoring censoring,
  - Mean survival of placebo group = 182/21 = 8.7 weeks
  - Mean survival of treatment group = 359/21 = 17.1 weeks
  - The treatment group stays in remission about twice as long
- Because censored patients very likely were in remission even longer, this underestimates their remission duration
- Average hazard rate is number of failures / total remission days
  - 21/182 for placebo = 0.115
  - 9/359 for treatment = 0.025
- So, this seems like a “no brainer”

- But we really should analyze survival using what we know about censoring

Kaplan-Meier Idea

- There is some conditional probability that a subject who has survived to time t_{j-1} will survive to t_j
- That conditional probability can be estimated by the empirical fraction of subjects who survive from t_{j-1} to t_j
  - But we don't count subjects who have been censored in $[t_{j-1}, t_j)$
- The probability that a subject survives at least to t_j is then the probability that they survive to t_{j-1} times that conditional probability
  - $\hat{S}(t_j) = \hat{S}(t_{j-1}) \times \hat{P}(T > t_j \mid T \geq t_j) = \prod_{i=1}^{j} \hat{P}(T > t_i \mid T \geq t_i)$

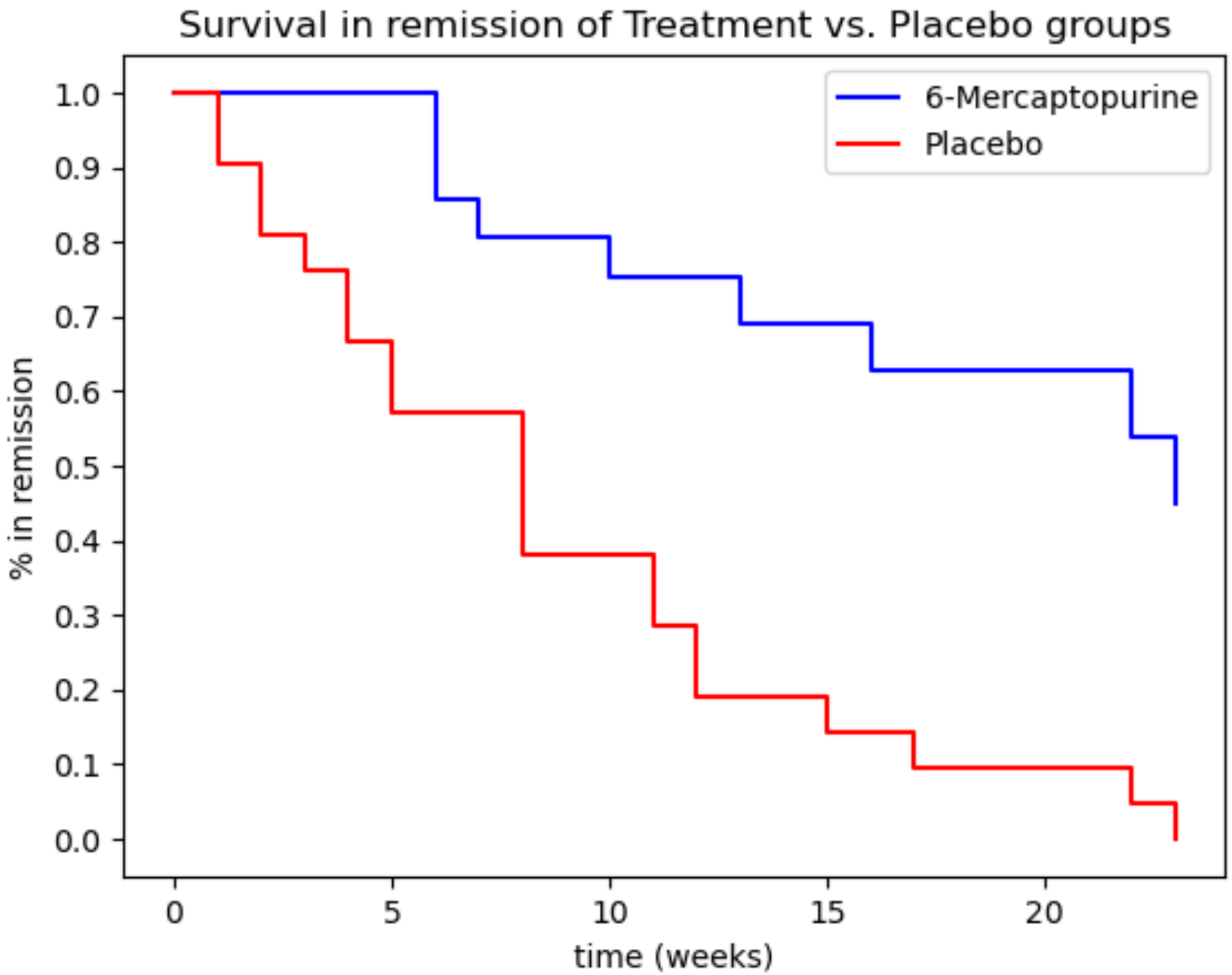
Calculating Survival for Censored Data (Kaplan-Meier)

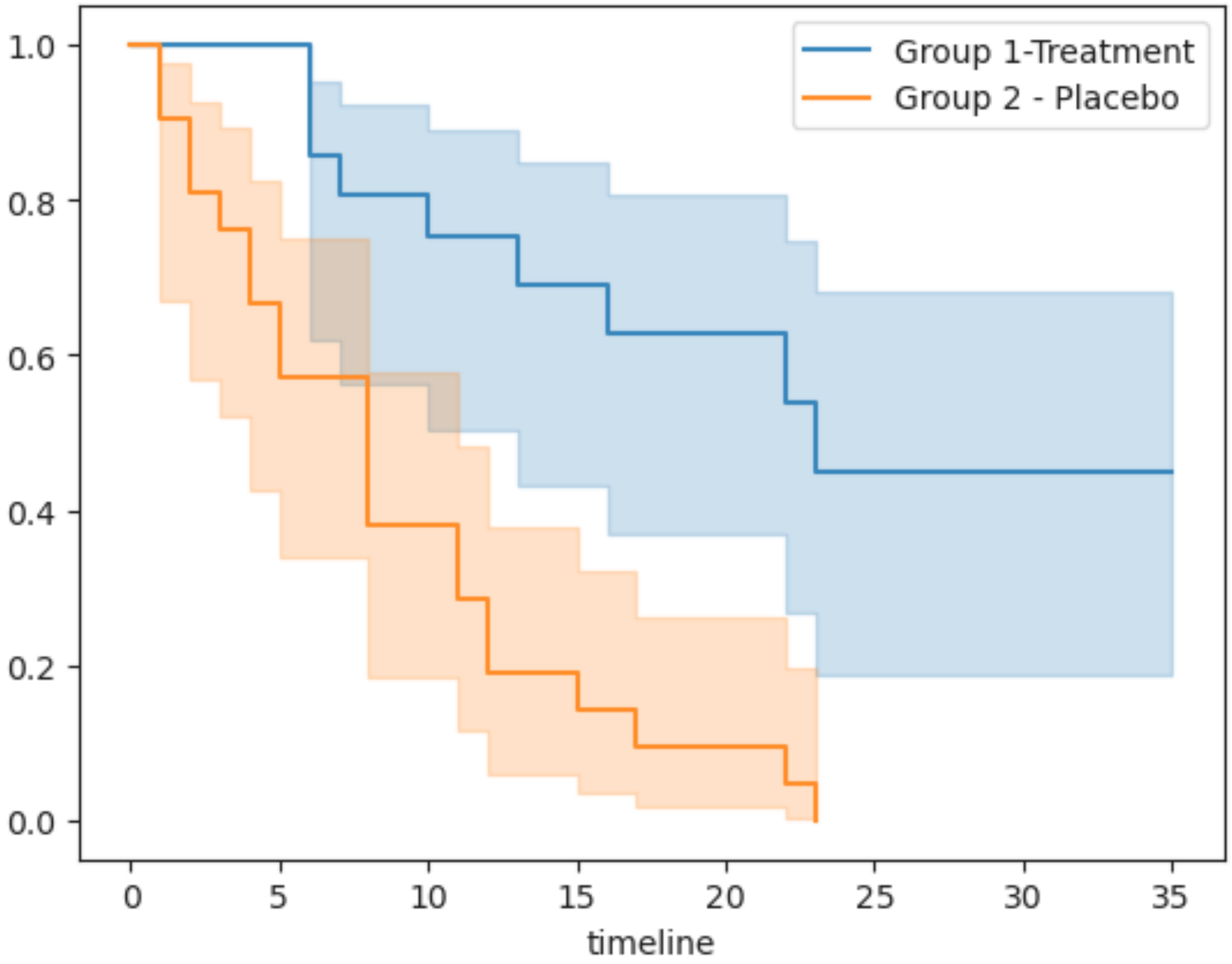
Ordered failure times t_j	# at risk n_j	# of failures m_j	# censored in $[t_j, t_{j+1})$ q_j	$\hat{S}(t_j)$
0	21	0	0	1
6	21	3	1	1 x 18/21 = .8571
7	17	1	1	.8571 x 16/17 = .8067
10	15	1	2	0.8067 x 14/15 = .7529
13	12	1	0	.7529 x 11/12 = .6902
16	11	1	3	0.6902 x 10/11 = .6275
22	7	1	0	.6275 x 6/7 = .5378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

Same Method also Works for Uncensored Data

Ordered failure times	# at risk n_j	# of failures m_j	# censored in $[t_j, t_{j+1})$ q_j	$\hat{S}(t_j)$	fraction surviving
0	21	0	0	1	1
1	21	2	0	1 x 19/21 = .9048	19/21
2	19	2	0	.9048 x 17/19 = .8095	17/21
3	17	1	0	.8095 x 16/17 = .7619	16/21
4	16	2	0	.7619 x 14/16 = .6667	14/21
5	14	2	0	.6667 x 12/14 = .5714	12/21
8	12	4	0	.5714 x 8/12 = .3810	8/21
11	8	2	0	.3810 x 6/8 = .2857	6/21
12	6	2	0	.2857 x 4/6 = .1905	4/21
15	4	1	0	.1905 x 3/4 = .1429	3/21
17	3	1	0	.1429 x 2/3 = .0952	2/21
22	2	1	0	.0952 x 1/2 = .0476	1/21
23	1	1	0	.0476 x 0/1 = 0	0/21
	0				

Comparison of Treatment vs. Placebo Groups




Group 1-Treatment
Group 2 - Placebo
1.0
0.8
0.6
0.4
0.2
0.0
0
5
10
15
20
25
30
35
timeline

Are Two Kaplan-Meier Curves Significantly Different? The Log-Rank Test

Remission data: $n = 42$

	# failures		# in risk set	
$t_{(j)}$	m_{1j}	m_{2j}	n_{1j}	n_{2j}
1	0	2	21	21
2	0	2	21	19
3	0	1	21	17
(4)	0	2	21	16
5	0	2	21	14
6	3	0	21	12
7	1	0	17	12
8	0	4	16	12
(10)	1	0	15	8
11	0	2	13	8
12	0	2	12	6
13	1	0	12	4
15	0	1	11	4
16	1	0	11	3
17	0	1	10	3
22	1	1	7	2
23	1	1	6	1

- A chi-square test using observed vs. expected cell counts for different categories of outcomes
- If the two curves are ***not*** different, the expected failure counts at each failure time should be in proportion to the number of patients at risk
  - $e_{1j} = \left(\frac{n_{1j}}{n_{1j} + n_{2j}}\right) \times (m_{1j} + m_{2j})$
  - $e_{2j} = \left(\frac{n_{2j}}{n_{1j} + n_{2j}}\right) \times (m_{1j} + m_{2j})$

Expanded Table (Remission Data)

		# failures		# in risk set		# expected		Observed–expected	
j	$t_{(j)}$	m_{1j}	m_{2j}	n_{1j}	n_{2j}	e_{1j}	e_{2j}	$m_{1j}-e_{1j}$	$m_{2j}-e_{2j}$
1	1	0	2	21	21	(21/42) × 2	(21/42) × 2	–1.00	1.00
2	2	0	2	21	19	(21/40) × 2	(19/40) × 2	–1.05	1.05
3	3	0	1	21	17	(21/38) × 1	(17/38) × 1	–0.55	0.55
4	4	0	2	21	16	(21/37) × 2	(16/37) × 2	–1.14	1.14
5	5	0	2	21	14	(21/35) × 2	(14/35) × 2	–1.20	1.20
6	6	3	0	21	12	(21/33) × 3	(12/33) × 3	1.09	–1.09
7	7	1	0	17	12	(17/29) × 1	(12/29) × 1	0.41	–0.41
8	8	0	4	16	12	(16/28) × 4	(12/28) × 4	–2.29	2.29
9	10	1	0	15	8	(15/23) × 1	(8/23) × 1	0.35	–0.35
10	11	0	2	13	8	(13/21) × 2	(8/21) × 2	–1.24	1.24
11	12	0	2	12	6	(12/18) × 2	(6/18) × 2	–1.33	1.33
12	13	1	0	12	4	(12/16) × 1	(4/16) × 1	0.25	–0.25
13	15	0	1	11	4	(11/15) × 1	(4/15) × 1	–0.73	0.73
14	16	1	0	11	3	(11/14) × 1	(3/14) × 1	0.21	–0.21
15	17	0	1	10	3	(10/13) × 1	(3/13) × 1	–0.77	0.77
16	22	1	1	7	2	(7/9) × 2	(2/9) × 2	–0.56	0.56
17	23	1	1	6	1	(6/7) × 2	(1/7) × 2	–0.71	0.71
Totals		9	21			19.26	10.74	–10.26	+10.26

Log-Rank Test

- $O_1 - E_1 = -10.26; O_2 - E_2 = 10.26$
  - Arbitrarily, choose the second distribution
- Log-rank statistic LR = $\dfrac{(O_2 - E_2)^2}{\mathrm{Var}(O_2 - E_2)}$
- $\mathrm{Var}(O_i - E_i) = \dfrac{\sum_j n_{1j} n_{2j} (m_{1j} + m_{2j})(n_{1j} + n_{2j} - m_{1j} - m_{2j})}{(n_{1j} + n_{2j})^2 (n_{1j} + n_{2j} - 1)}$

- Null hypothesis H_0: no difference between survival curves
- LR $\sim \chi^2$ with 1 degree of freedom under H_0
- For our example, using Python's `lifelines` package
  - $O_2 - E_2 = 10.26$, $\mathrm{Var}(O_2 - E_2) = 6.2685$, so LR = 16.793
  - $p < .0001$; in fact, $\log_2(p) = -14.55$
  - Thus, H_0 is soundly rejected, so the treatment is effective
- Alternatives to Log-Rank test:
  - Wilcoxon, Tarone-Ware, Peto, Flemington-Harrington, …

https://medium.com/analytics-vidhya/log-rank-test-kaplan-meier-survival-curve-python-code-3fc78da644d5

Leukemia Remission Data

Group 1($n = 21$)		Group 2($n = 21$)	
t(weeks)	log WBC	t(weeks)	log WBC
6	2.31	1	2.80
6	4.06	1	5.00
6	3.28	2	4.91
7	4.43	2	4.48
10	2.96	3	4.01
13	2.88	4	4.36
16	3.60	4	2.42
22	2.32	5	3.49
23	2.57	5	3.97
6+	3.20	8	3.52
9+	2.80	8	3.05
10+	2.70	8	2.32
11+	2.60	8	3.26
17+	2.16	11	3.49
19+	2.05	11	2.12
20+	2.01	12	1.50
25+	1.78	12	3.06
32+	2.20	15	2.30
32+	2.53	17	2.95
34+	1.47	22	2.73
35+	1.45	23	1.97

+ denotes censored observation

Dealing with Covariates

- So far, groups determined by a single factor; e.g., treatment vs. placebo
- Outcomes often depend on factors: demographics, comorbidities, lab data, geography, etc.
- How do we deal with these additional factors?
  - Confounding: X_2
  - Interaction: $X_3 = X_1 \times X_2$ for synergistic

Similar to Linear Regression Models

Model 1:

	Coef.	Std. Err.	z	p > |z|	Haz. Ratio	[95% Conf.	Interval]
Rx	1.509	0.410	3.68	0.000	4.523	2.027	10.094

No. of subjects = 42 Log likelihood = −86.380 Prob > chi2 = 0.0001

Model 2:

	Coef.	Std. Err.	z	p > |z|	Haz. Ratio	[95% Conf.	Interval]
Rx	1.294	0.422	3.07	0.002	3.648	1.595	8.343
log WBC	1.604	0.329	4.87	0.000	4.975	2.609	9.486

No. of subjects = 42 Log likelihood = −72.280 Prob > chi2 = 0.0000

Model 3:

	Coef.	Std. Err.	z	p > |z|	Haz. Ratio	[95% Conf.	Interval]
Rx	2.355	1.681		161	10.537	0.391	284.201
log WBC	1.803	0.447	4.04	0.000	6.067	2.528	14.561
Rx x log WBC	−0.342	0.520	−0.66	0.510	0.710	0.256	1.967

No. of subjects = 42 Log likelihood = −72.066 Prob > chi2 = 0.0000

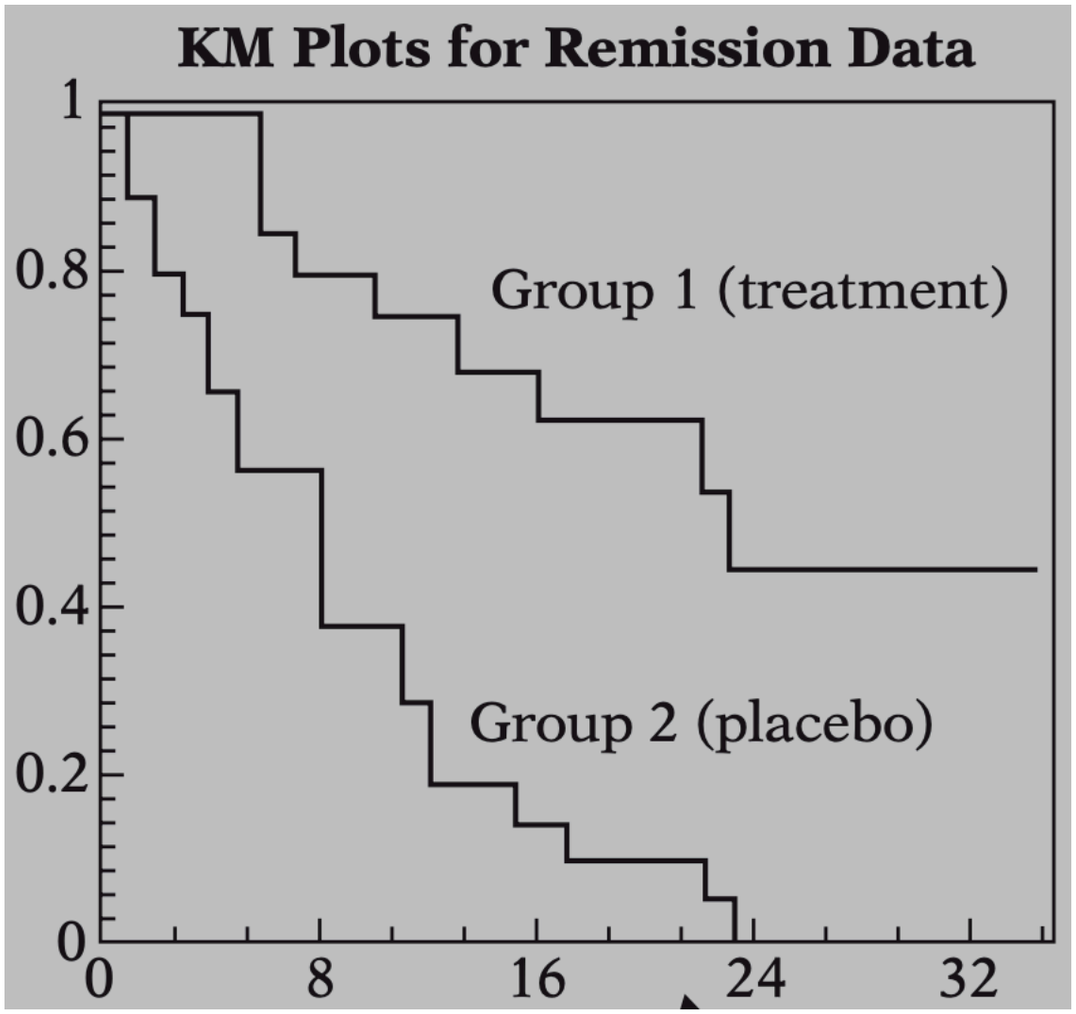
Checking for Confounding

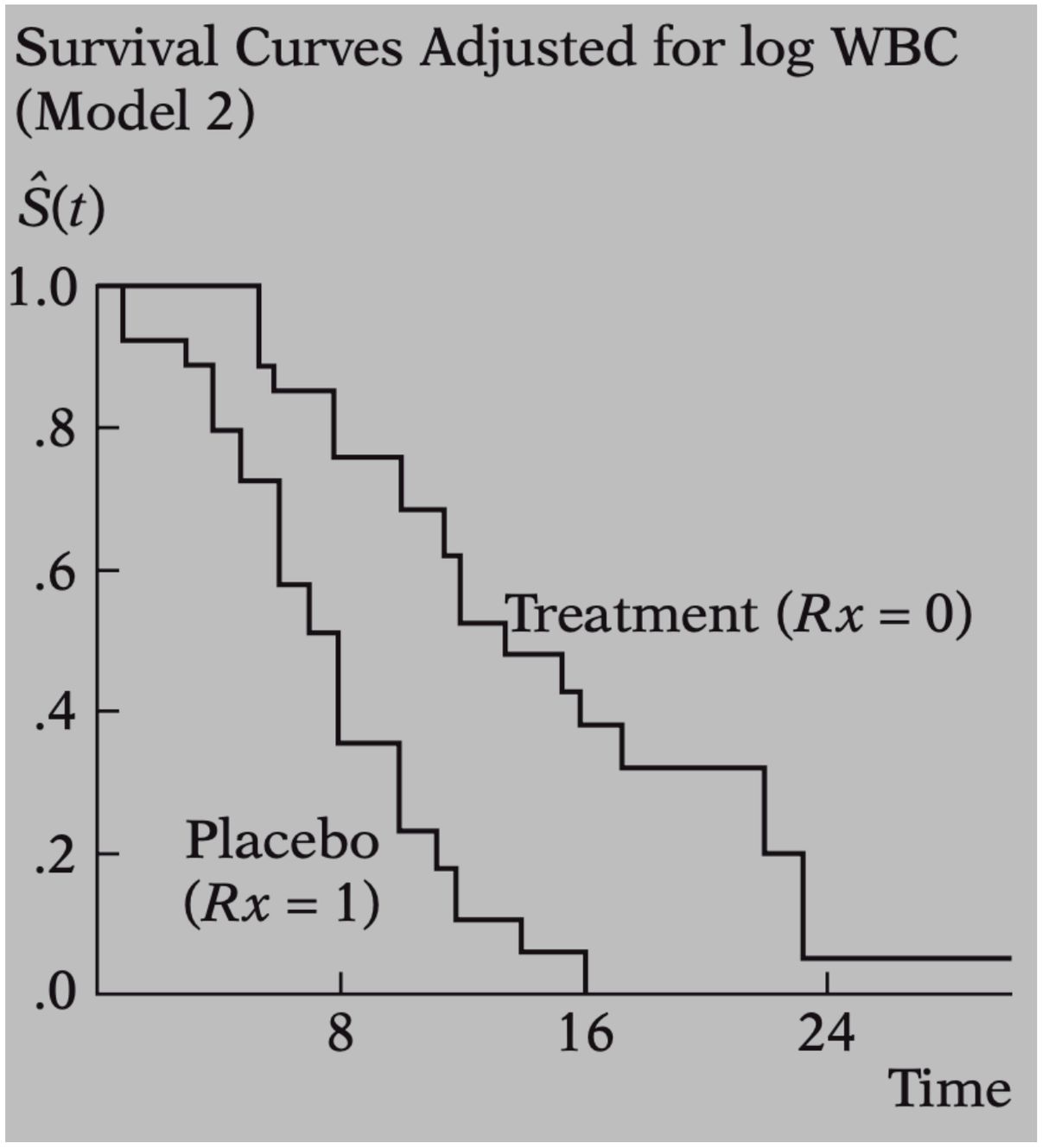
Model 2:

	Coef.	Std. Err.	z	p > |z|	Haz. Ratio	[95% Conf. Interval]	
Rx	1.294	0.422	3.07	0.002	3.648	1.595	8.343
log WBC	1.604	0.329	4.87	0.000	4.975	2.609	9.486

No. of subjects = 42 Log likelihood = –72.280 Prob > chi2 = 0.0000

- Is the treatment still a significant effect?
  - Yes; $p < .002$
- What is the best estimate for that effect?
  - $\widehat{\mathrm{HR}} = e^{1.294} = 3.648$
  - Confidence interval does not include 1.0
- $\widehat{\mathrm{HR}}$s are different for *Rx* in Model 1 (4.523) and Model 2 (3.648)
  - log WBC "explains away" part of the effect of *Rx*, so Model 2 should be used
  - Confidence interval for $\widehat{\mathrm{HR}}$ in Model 2 is (a little) narrower than in Model 1


KM Plots for Remission Data
1
0.8
0.6
0.4
0.2
0
0
8
16
24
32
Group 1 (treatment)
Group 2 (placebo)


Survival Curves Adjusted for log WBC (Model 2)
$\hat{S}(t)$
1.0
.8
.6
.4
.2
.0
8
16
24
Time
Treatment ($Rx = 0$)
Placebo ($Rx = 1$)

Cox Proportional Hazard Model

- The hazard depends on the covariates
- $h(t, \mathbf{X}) = h_0(t)e^{\sum_{i=1}^{p} \beta_i X_i}$ for $\mathbf{X} = (X_1, X_2, \ldots, X_p)$ as explanatory/predictor variables
  - Multiplicative contribution of each X_i; this term is parametric
- $h_0(t)$ is the time-dependent baseline hazard, not dependent on $\mathbf{X}$
  - Its form is not specified; thus Cox model is semi-parametric
- $e^{\sum_{i=1}^{p} \beta_i X_i}$ is the adjustment for covariates, which are time-independent
  - Therefore, time-dependent hazards due to covariates make this model inappropriate
  - There is an "extended Cox model" that allows time-dependent $\mathbf{X}$
- Can estimate the β_i without specifying the form of $h_0(t)$
- Contrast with Weibull model, which is parametric:
  - $h(t, \mathbf{X}) = \lambda p t^{p-1}$ where $\lambda = e^{\sum_{i=1}^{p} \beta_i X_i}$
  - $h_0(t) = pt^{p-1}$

How to Train a Cox Model?

- Cox model: $h(t, \mathbf{X}) = h_0(t)e^{\sum_{i=1}^{p} \beta_i X_i}$
- Estimate for our leukemia remission example
  - $\hat{h}(t, (X)) = \hat{h}_0(t)e^{1.294\,Rx + 1.604\,\log\mathrm{WBC}}$
- ML estimate: maximize likelihood function
  - L = joint probability of observed data = $L(\beta)$

	Coef.	Std.Err.	p > |z|	Haz. Ratio
Rx	1.294	0.422	0.002	3.648
log WBC	1.604	0.329	0.000	4.975

No. of subjects = 42 Log likelihood = –72.280

- For each failure time, we compute likelihood of the data L_i

$$L = L_1 \times L_2 \times \cdots \times L_k = \prod_{j=1}^{k} L_j \text{ for } k \text{ failure times}$$

- L_i considers only subjects who fail, but censored subjects are used in computing L_j for $j < i$
- Iterative solution over p parameters for $\dfrac{\partial \ln L}{\partial \beta_i} = 0$

What is the Cox Likelihood Function?

- Likelihood at each failure time that we see the events in the data given the estimated hazard function
  - Adjust the betas of that function to maximize the likelihood
- Simple example:

ID	TIME	STATUS	SMOKE
Barry	2	1	1
Gary	3	1	0
Harry	5	0	0
Larry	8	1	1

$h(t) = h_0(t)e^{\beta_1 SMOKE}$

ID	Hazard
Barry	$h_0(t)e^{\beta_1}$
Gary	$h_0(t)e^{0}$
Harry	$h_0(t)e^{0}$
Larry	$h_0(t)e^{\beta_1}$

- L does not depend on $h_0(t)$, or t
- Only the order of events matters

$$L = L_1 \times L_2 \times L_3$$

$$L_1 = \left[\frac{h_0(t)e^{\beta_1}}{h_0(t)e^{\beta_1} + h_0(t)e^{0} + h_0(t)e^{0} + h_0(t)e^{\beta_1}}\right]$$

$$L_2 = \left[\frac{h_0(t)e^{0}}{h_0(t)e^{0} + h_0(t)e^{0} + h_0(t)e^{\beta_1}}\right]$$

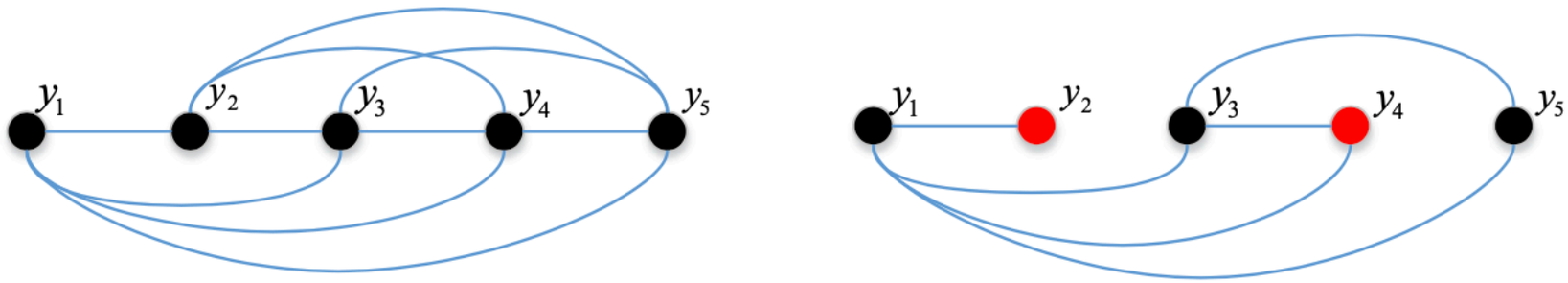
$$L_3 = \left[\frac{h_0(t)e^{\beta_1}}{h_0(t)e^{\beta_1}}\right]$$

$$L = \left[\frac{e^{\beta_1}}{e^{\beta_1} + e^{0} + e^{0} + e^{\beta_1}}\right] \times \left[\frac{e^{0}}{e^{0} + e^{0} + e^{\beta_1}}\right] \times \left[\frac{e^{\beta_1}}{e^{\beta_1}}\right]$$

Evaluation for survival modeling

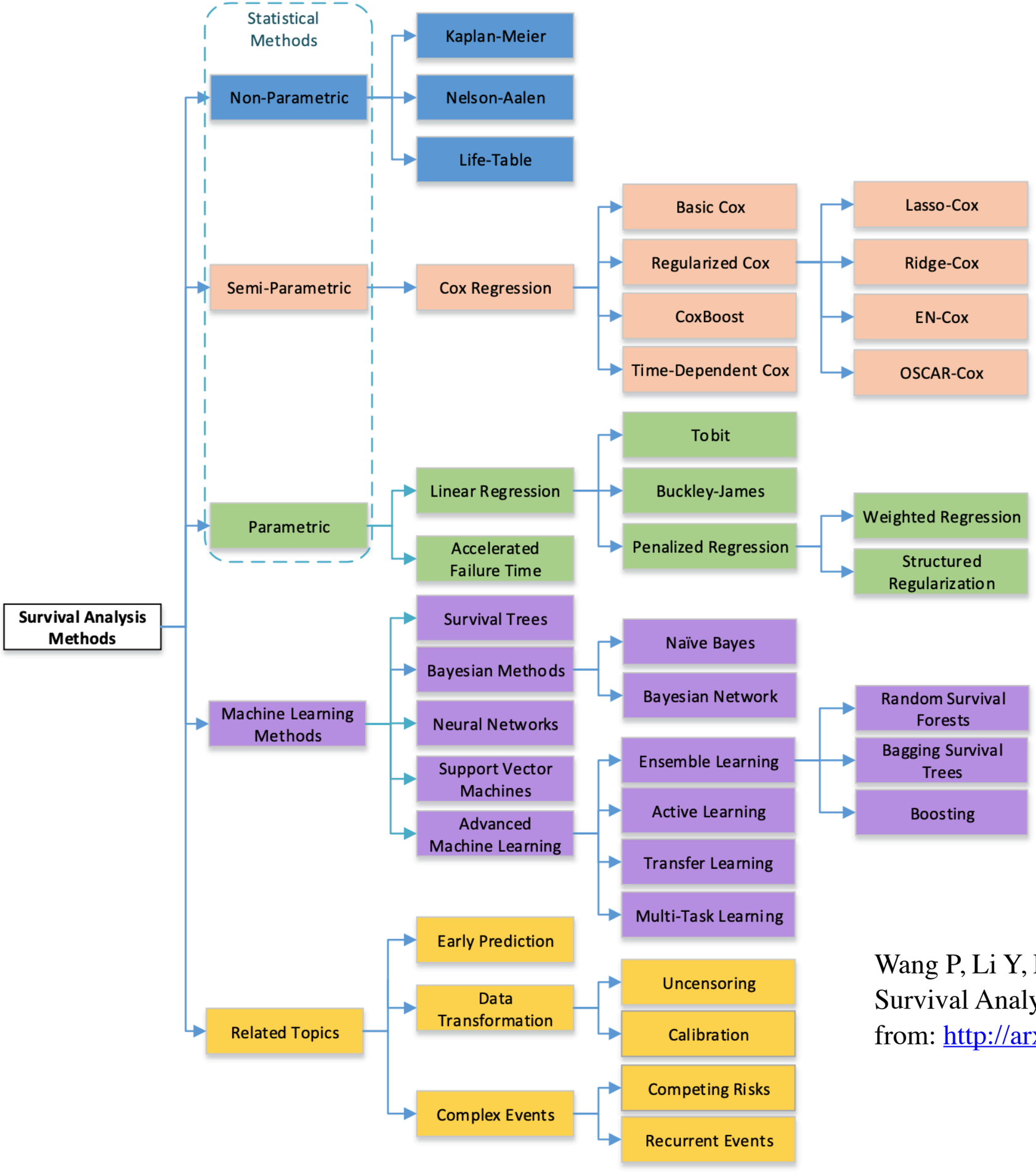
- Concordance-index (also called C-statistic): look at model's ability to predict relative survival times (notation here uses y_i instead of t_i):

$$\hat{c} = \frac{1}{n_c} \sum_{i:d_i=1} \sum_{t_i<t_j} \mathbf{1}[\hat{y}_i < \hat{y}_j] \text{ where } n_c = \sum_{i:d_i=1} \sum_{t_i<t_j} 1 \text{ and } \hat{y}_i = \mathbb{E}_{f(T|x_i;\beta)}[T]$$



- Black dots are observed, red at censored; compare only
  - all pairs of observed events
  - censored events that come after observed events
- Equivalent to AUC for binary variables and no censoring

Wang P, Li Y, Reddy CK. Machine Learning for Survival Analysis: A Survey. arXiv; 2017. Available from: http://arxiv.org/abs/1708.04649



Taxonomy of the methods developed for survival analysis

Wang P, Li Y, Reddy CK. Machine Learning for Survival Analysis: A Survey. arXiv; 2017. Available from: http://arxiv.org/abs/1708.04649

Deep Cox Mixture Model

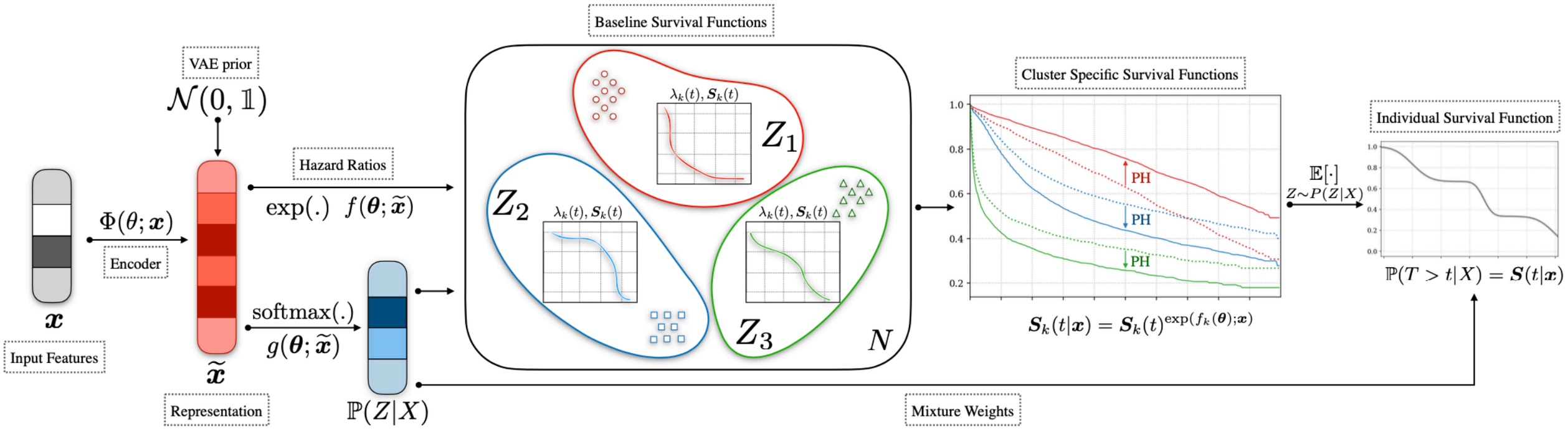
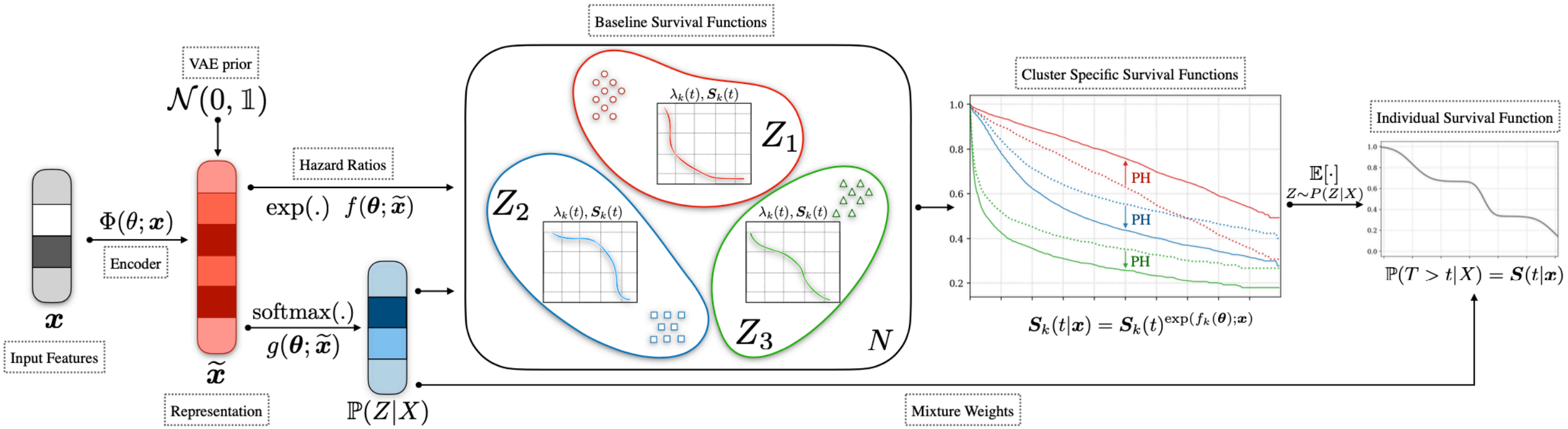


Figure 1: **Deep Cox Mixtures**: Representation of the individual covariates $\boldsymbol{x}$ are generated using an encoding neural network. The output representation $\widetilde{\boldsymbol{x}}$ then interacts with linear functions f and g that determine the proportional hazards within each cluster $Z \in \{1, 2, ...K\}$ and the mixing weights $\mathbb{P}(Z|X)$ respectively. For each cluster, baseline survival rates $\boldsymbol{S}_k(t)$ are estimated non-parametrically. The final individual survival curve $S(t|\boldsymbol{x})$ is an average over the cluster specific individual survival curves weighted by the mixing probabilities $\mathbb{P}(Z|X = \boldsymbol{x})$.

Nagpal C, Yadlowsky S, Rostamzadeh N, Heller K. Deep Cox Mixtures for Survival Regression. In: Proceedings of the 6th Machine Learning for Healthcare Conference, PMLR; 2021 p. 674–708. Available from: https://proceedings.mlr.press/v149/nagpal21a.html

Deep Cox Mixture Model



$$\mathscr{L}(\boldsymbol{\theta}, \boldsymbol{\Lambda}_k) = \prod_{i=1}^{|\mathscr{D}|} \int_Z (\boldsymbol{\lambda}(u_i|\boldsymbol{x}_i))^{\delta_i} \, \boldsymbol{S}_k(u_i|\boldsymbol{x}_i) \mathbb{P}(Z = k|\boldsymbol{x}_i).$$

$$\text{where, } \boldsymbol{\lambda}(u_i|\boldsymbol{x}_i) = \boldsymbol{\lambda}_k(u_i) \exp\left(f_k(\boldsymbol{\theta}, \boldsymbol{x}_i)\right), \quad \boldsymbol{S}_k(u_i|\boldsymbol{x}_i) = \boldsymbol{S}_k(u_i)^{\exp\left(f_k(\boldsymbol{\theta}; \boldsymbol{x}_i)\right)}$$

$$\text{and, } \mathbb{P}(Z = k|X = \boldsymbol{x}_i) = \text{softmax}\big(g(\boldsymbol{\theta}; \boldsymbol{x}_i)\big)$$

Nagpal C, Yadlowsky S, Rostamzadeh N, Heller K. Deep Cox Mixtures for Survival Regression. In: Proceedings of the 6th Machine Learning for Healthcare Conference, PMLR; 2021 p. 674–708. Available from: https://proceedings.mlr.press/v149/nagpal21a.html

Final Thoughts and References

- Strong assumption (censoring time independent of survival time) allow us to develop (relatively) simple solutions
  - But how realistic is this?
  - Can we relax this assumption?

- *Recommended starting place:* Kleinbaum & Klein. Survival Analysis: A Self-Learning Text. Springer Statistics for biology and Health, 2005
- *Additional detail:* Kalbfleisch & Prentice, The Statistical Analysis of Failure Time Data, Wiley 2002 [MIT proxy]
- Ishwaran et al., Random Survival Forests. The Annals of Applied Statistics, 2008
- Alaa and van der Schaar. Deep multi-task gaussian processes for survival analysis with competing risks. NeurIPS, 2017