Section 18

Lecture 9

Partial likelihood

From previous statistics courses you have learned about likelihood methods. You have used these for *parametric* models. However, because the Cox Model is semi-parametric, we cannot immediately apply likelihood methods for estimation. David Cox did a clever move to define the partial likelihood, and we shall see that we can mimic likelihood methods when we study the partial likelihood.

Partial likelihood

Consider the intensity function

$$\lambda_i(t) = Z_i(t)\alpha_0(t)e^{\boldsymbol{\beta}^T \boldsymbol{x}_i(t)}.$$

Let $N_\bullet = \sum_{l=1}^n N_l(t)$, and $\lambda_\bullet(t) = \sum_{l=1}^n \lambda_l(t) = \sum_{l=1}^n Z_l(t)\alpha_0(t)e^{\boldsymbol{\beta}^T \boldsymbol{x}_l(t)}$.
Note that $\lambda_i(t) = \lambda_\bullet(t)\pi(i \mid t)$ where

$$\pi(i \mid t) = \frac{\lambda_i(t)}{\lambda_\bullet} = \frac{Z_i(t)\cancel{\alpha_0(t)}e^{\boldsymbol{\beta}^T \boldsymbol{x}_i(t)}}{\sum_{l=1}^n Z_l(t)\cancel{\alpha_0(t)}e^{\boldsymbol{\beta}^T \boldsymbol{x}_l(t)}}.$$

Definition (Partial likelihood)

Consider the event times $T_1 < T_2 < \ldots$, and let i_j be the index of the individual who experiences an event at T_j. The partial likelihood is

$$L(\boldsymbol{\beta}) = \prod_{T_j} \pi(i_j \mid T_j) = \prod_{T_j} \frac{Z_{i_j}(T_j)e^{\boldsymbol{\beta}^T \boldsymbol{x}_{i_j}(T_j)}}{\sum_{l=1}^n Z_l(T_j)e^{\boldsymbol{\beta}^T \boldsymbol{x}_l(T_j)}}.$$

Interpretation: $\pi(i \mid t)$ is the probability of observing an event for individual i at time t, given the history until time t and that there is an event at time t.

Partial likelihood in a more familiar way

Let $\mathcal{R}_j = \{l : Z_l(T_j) = 1\}$, which is the risk set at T_j. Then, we can re-write the partial likelihood as something that looks familiar

$$L(\boldsymbol{\beta}) = \prod_{T_j} \frac{e^{\boldsymbol{\beta}^T \boldsymbol{x}_{i_j}(T_j)}}{\sum_{l \in \mathcal{R}_j} e^{\boldsymbol{\beta}^T \boldsymbol{x}_l(T_j)}}.$$

Definition (Maximum partial likelihood estimator)

The maximum partial likelihood estimator $\hat{\boldsymbol{\beta}}$ is the value of $\boldsymbol{\beta}$ that maximizes the partial likelihood.

Intuition on the partial likelihood

Because the non-parameteric form of the baseline hazard $\alpha_0(t)$, we pose no model assumptions on the time between the events $T_1, T_2 \ldots$. Thus, the partial likelihood is a function only of the *ranks* of the events. Indeed, the partial likelihood would be unchanged by any monotone transformation of the time scale.
Furthermore, the censoring times are only present in the risk sets. Under the independent censoring assumption, this is reasonable, because knowing when an individual is censored does not provide information about the hazard function.

Results on partial likelihood estimation

Theorem (Properties of the partial likelihood)

In large samples, $\hat{\boldsymbol{\beta}}$ is approximately normally distributed with $\mathbb{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}_0$ and variance $\mathcal{I}(\boldsymbol{\beta}_0)^{-1}$, where $\mathcal{I}(\beta) = -\frac{\delta^2}{\delta\beta_h, \delta\beta_j} logL(\beta)$ is the observed (and expected) partial information matrix.

$\mathcal{I}(\beta_0)$ is consistently estimated by $\mathcal{I}(\hat{\beta})$.
It took 9 years from the Cox model was proposed until the large sample properties of the Cox model were formally shown.
Later, it was realized that a very convenient way of proving this theorem is to use counting processes and martingales. I am not giving a rigorous proof here, but I am happy to direct you to sources if you want more detail: The idea is to show and use that the score $U(\beta) = \frac{\delta log(L(\beta))}{\delta\beta}$ is a martingale.[22]
The point is that – with the theory you have learned so far – you would be able to derive it, but the derivation is more tedious compared to what we have done until now.

[22] Per K Andersen et al. *Statistical models based on counting processes*. Springer Science & Business Media, 2012.

We can conduct tests of $\beta = \beta_0$ in the usual way

- Likelihood ratio statistic: $\chi^2_{LR} = 2\{\log L(\hat{\beta}) - \log L(\beta_0)\}$
- Wald test statistic: $\chi^2_W = (\hat{\beta} - \beta_0)^T I(\hat{\beta})(\hat{\beta} - \beta_0)$.
- Score test statistic: $\chi^2_{SC} = U(\beta_0)^T I(\beta_0)^{-1} U(\beta_0)$, *where* $U(\beta) = \frac{\delta}{\delta\beta} \log L(\beta)$. We shall see that this one is related to log rank test.

These statistics are χ^2 distributed with p degrees of freedom.

Example: Survival among Norwegian smokers and non-smokers

Table 4.3 Estimated relative risks (hazard rate ratios) with 95% confidence intervals (c.i.) based on a Cox regression analysis of the total mortality in three Norwegian counties.

Covariate	Hazard ratio	95% c.i.
Sex	0.58	0.48–0.70
Former smoker	1.37	1.05–1.78
1–9 cigarettes per day	2.44	1.83–3.25
10–19 cigarettes per day	2.45	1.91–3.14
20 or more cigarettes per day	2.96	2.19–4.00

From Odd Aalen, Ornulf Borgan, and Hakon Gjessing. *Survival and event history analysis: a process point of view*. Springer Science & Business Media, 2008

We are interested in the absolute risk (survival)

To estimate expected survival from the Cox model, we also need an estimate of the baseline (cumulative) hazard $H_0(t) = \int_0^t \alpha_0(u)du$. Consider again the aggregated survival process $N_\bullet = \sum_{l=i}^n N_i(t)$, which has intensity

$$\lambda_\bullet(t) = \alpha_0(t)\Big(\sum_{l=1}^n Z_l(t)e^{\boldsymbol{\beta}^T \boldsymbol{x}_l(t)}\Big).$$

Suppose we knew $\boldsymbol{\beta}$, then $\lambda_\bullet(t)$ would satisfy the multiplicative intensity model, which would mean that we could simply estimate $\hat{H}_0(t)$ by the Nelson-Aalen-like estimator

$$\hat{H}_0(t;\boldsymbol{\beta}) = \int_0^t \frac{dN_\bullet(u)}{\sum_{l=1}^n Z_l(u)e^{\boldsymbol{\beta}^T \boldsymbol{x}_l(u)}}.$$

This motivates the Breslow estimator, which is given on the next slide

Breslow estimator

Definition (Breslow estimator)

$$\hat{H}_0(t) = \int_0^t \frac{dN_\bullet(u)}{\sum_{l=1}^n Z_l(u) e^{\hat{\boldsymbol{\beta}}^T \boldsymbol{x}_l(u)}} = \sum_{T_j \leq t} \frac{1}{\sum_{l \in \mathcal{R}_j} e^{\hat{\boldsymbol{\beta}}^T \boldsymbol{x}_l(T_j)}}.$$

Thus, when the covariates are fixed, we get an estimator of the conditional (on $\boldsymbol{x}_0$) hazard

$$\hat{H}(t \mid \boldsymbol{x}_0) = \hat{H}_0(t) e^{\hat{\boldsymbol{\beta}}^T \boldsymbol{x}_0},$$

and then we use the product integral representation

$$S(t \mid \boldsymbol{x}_0) = \prod_{u \leq t} \{1 - dH(u \mid \boldsymbol{x}_0)\}$$

to motivate an estimator of the conditional survival, $\hat{S}(t \mid \boldsymbol{x}_0)$,

$$\hat{S}(t \mid \boldsymbol{x}_0) = \prod_{u \leq t} \{1 - d\hat{H}(u \mid \boldsymbol{x}_0)\} = \prod_{T_j \leq t} \{1 - \Delta \hat{H}(T_j \mid \boldsymbol{x}_0)\},$$

which is consistent and asymptotically normal.

Some comments on hazard modelling

- In my opinion, there is an unfortunate habit of reporting hazard ratios in the literature.
- Reporting parameters on the survival scale is, broadly speaking, more desirable.

  Yet, hazard models, including the Cox model, are useful when we estimate these other parameters.
- In the next slides, I will give some arguments why.

Example: Survival among Norwegian never-smokers and heavy smokers

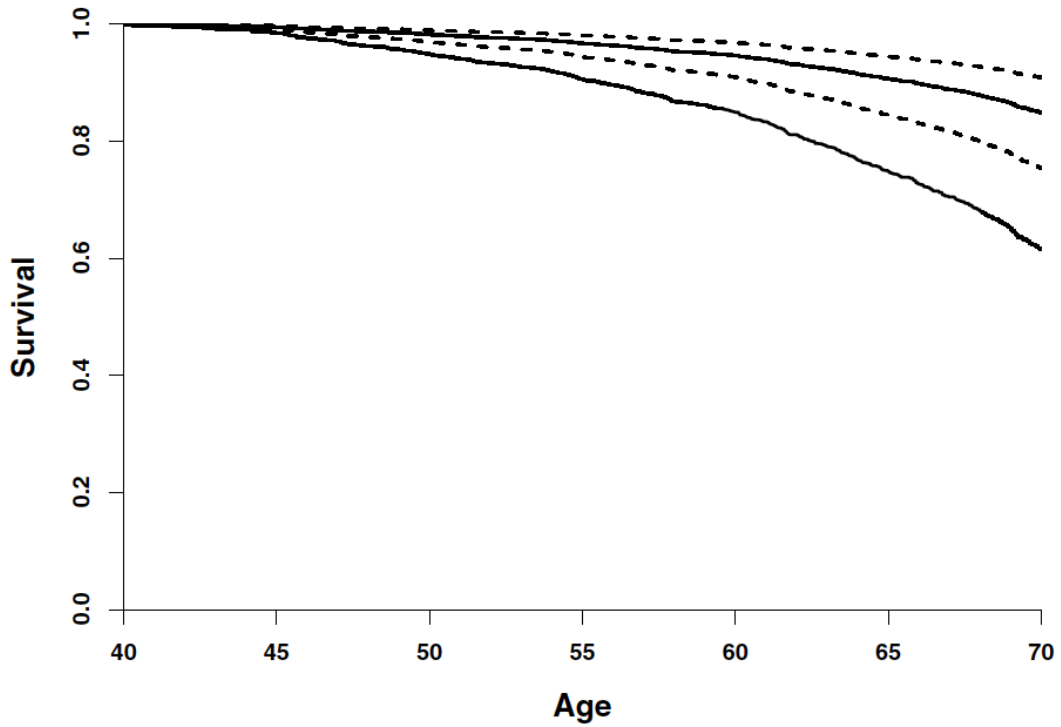
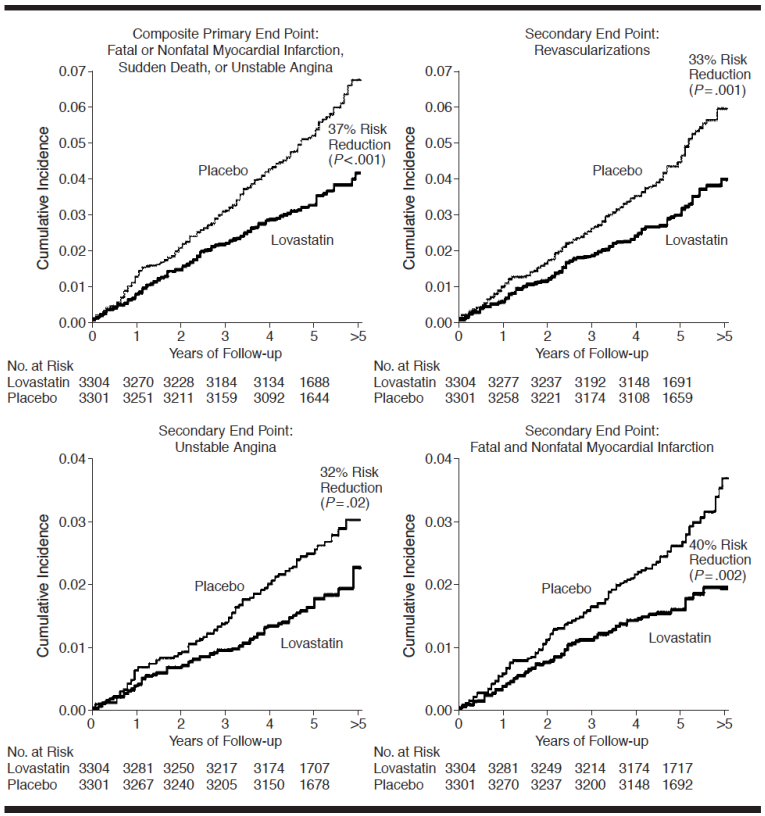


Fig. 4.3 *Estimated survival curves (given survival to 40 years) for some covariate values based on a Cox regression analysis of the total mortality in three Norwegian counties. Males: drawn lines; females: dashed lines. Upper lines: never smoked; lower lines: smokes 20 or more cigarettes per day.*

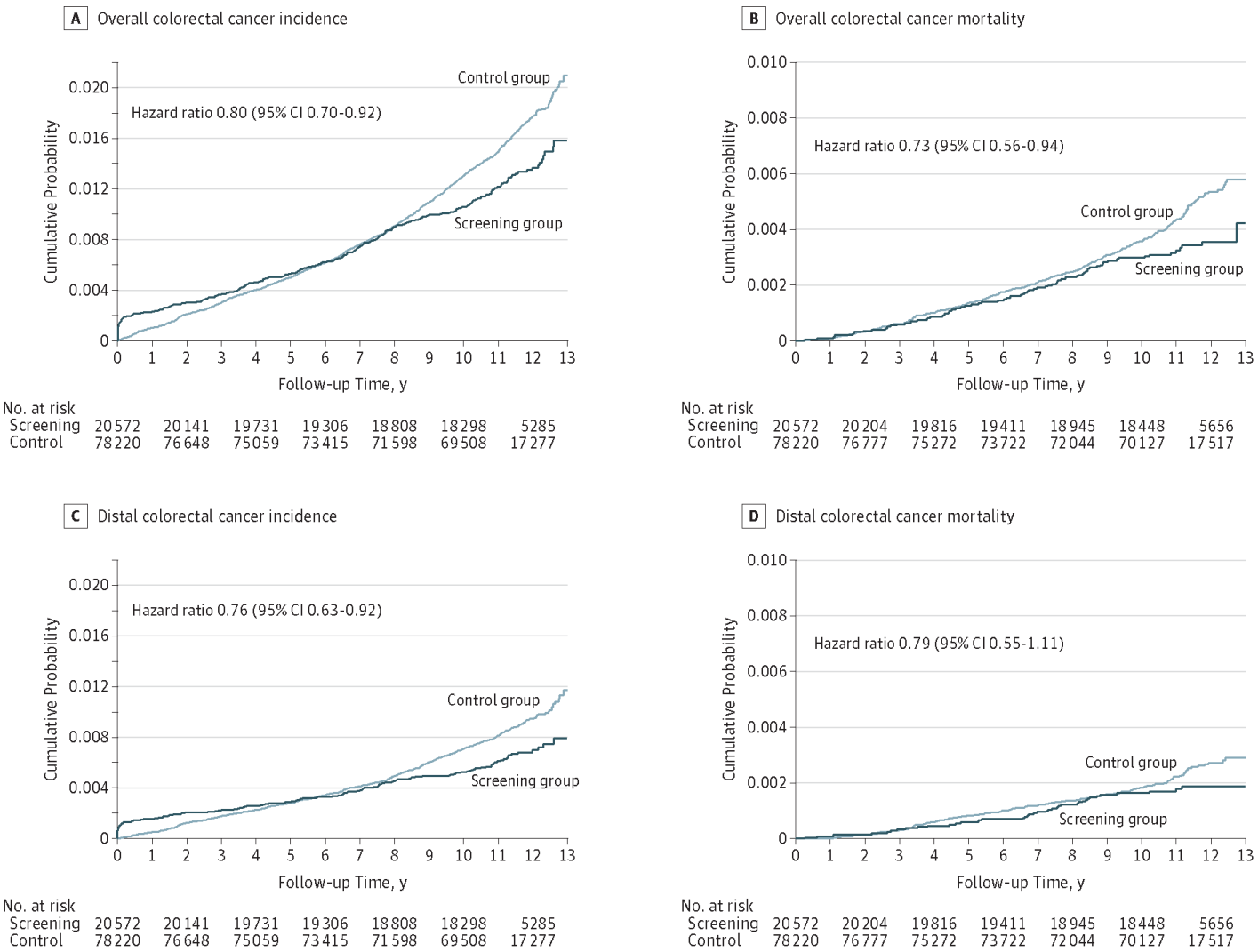
Arguably, the curves provide more information than the numbers in the previous table.

Hazards are often not proportional, but hazard ratios are too often reported anyway

Statin therapy



Cancer screening



Clearly not proprtional hazards....

Table from the article on cancer screening

Table 3. Colorectal Cancer Incidence and Mortality in the Screening and Control Groups

	Group					
	Screening		Control			
	No.	Cases/100 000 Person-Years	No.	Cases/100 000 Person-Years	HR (95% CI)	*P* Value
Colorectal Cancer Incidence[a]						
Overall	253	112.6	1086	141.0	0.80 (0.70-0.92)	.001
Location						
Distal	137	60.9	621	80.1	0.76 (0.63-0.92)	.004
Proximal	112	49.8	424	55.5	0.90 (0.73-1.10)	.31
Sex						
Men	128	115.6	586	157.6	0.73 (0.60-0.89)	.002
Women	125	109.6	500	125.5	0.87 (0.72-1.06)	.18
Age group, y						
50-54	40	57.2	315	84.3	0.68 (0.49-0.94)	.02
55-64	213	140.6	771	169.6	0.83 (0.71-0.96)	.02
Screening modality						
Flexible sigmoidoscopy	114	101.9	1086	141.0	0.72 (0.59-0.87)	.001
Flexible sigmoidoscopy + FOBT	139	123.3	1086	141.0	0.88 (0.74-1.05)	.15
Colorectal Cancer Mortality[b]						
Overall	71	31.4	330	43.1	0.73 (0.56-0.94)	.02
Location						
Distal	39	17.2	168	21.8	0.79 (0.55-1.11)	.18
Proximal	30	13.4	139	18.3	0.73 (0.49-1.09)	.12
Sex						
Men	32	28.6	182	49.1	0.58 (0.40-0.85)	.005
Women	39	34.2	148	37.4	0.91 (0.64-1.30)	.62
Age group, y						
50-54	12	17.1	87	23.2	0.74 (0.40-1.35)	.32
55-64	59	38.7	243	53.1	0.73 (0.55-0.97)	.03
Screening modality						
Flexible sigmoidoscopy	41	36.4	330	43.1	0.84 (0.61-1.17)	.30
Flexible sigmoidoscopy + FOBT	30	26.5	330	43.1	0.62 (0.42-0.90)	.01
All-cause mortality	2183	969.0	7762	994.6	0.97 (0.93-1.02)	.28

Abbreviations: FOBT, fecal occult blood test; HR, hazard ratio.

[a] Person-years of observation: for the screening group, 221 429; and for the control group, 828 207.

[b] Person-years of observation: for the screening group, 222 677; and for the control group, 832 003.

Graphical model checking (this is just visual inspection...)

We consider a Cox model with fixed covariates, which is most used in practice. Then

$$\alpha(t \mid \boldsymbol{x}) = \alpha_0(t) e^{\beta^T \boldsymbol{x}}.$$

Thus,

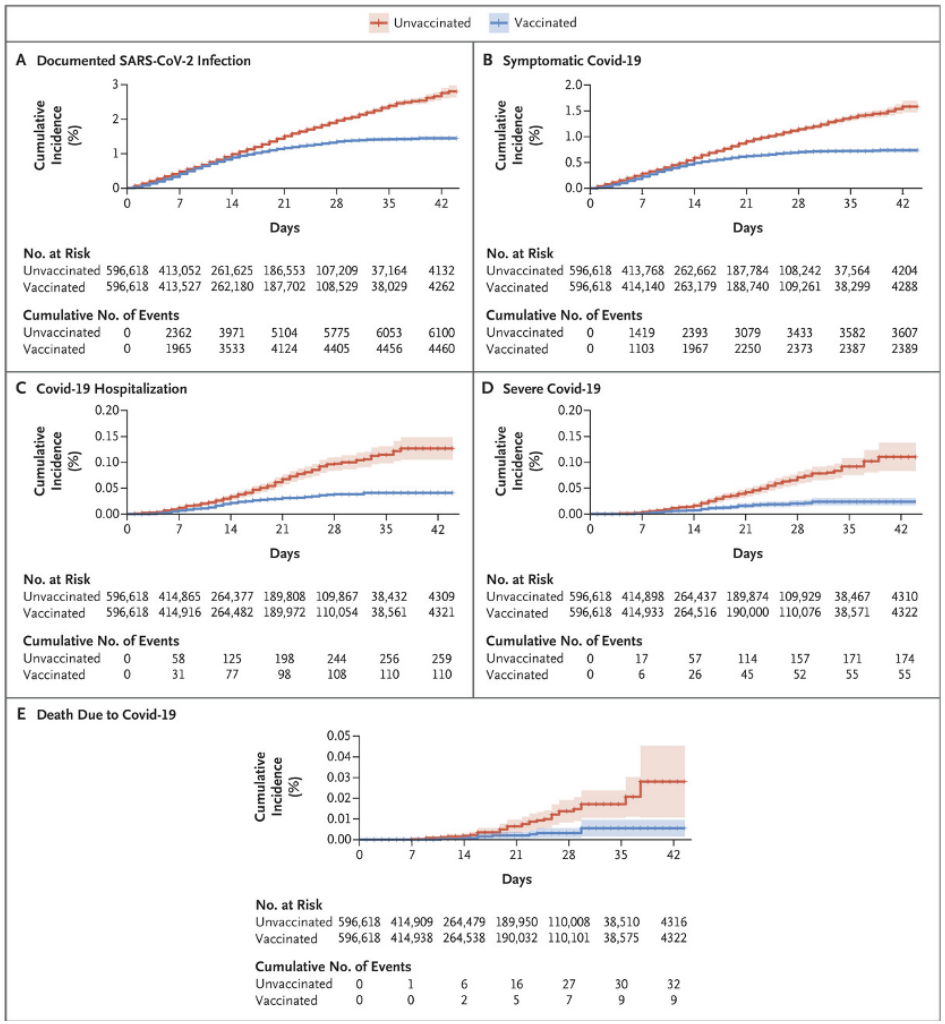
$$-\log(S(t \mid \boldsymbol{x})) = \int_0^t \alpha(s \mid \boldsymbol{x}) ds = \int_0^t \alpha_0(s) e^{\beta^T \boldsymbol{x}} ds.$$

The model implies linearity in the following sense,

$$\log\{-\log(S(t \mid \boldsymbol{x}))\} = \log\{\int_0^t \alpha(s \mid \boldsymbol{x}) ds\} = \log\{\int_0^t \alpha_0(s) ds\} + \beta^T \boldsymbol{x}.$$

Thus, $\log\{-\log(\hat{S}(t \mid \boldsymbol{x}_1))\}$ and $\log\{-\log(\hat{S}(t \mid \boldsymbol{x}_2))\}$ for any $\boldsymbol{x}_1, \boldsymbol{x}_2$ should be parallel in a plot with time on the horizontal axis. For a single binary covariate, we could e.g. look at the difference between two Kaplan-Meier estimators.

Example COVID in real-life



The authors did not assume proportional hazard here.

Estrogen trial continued: Does the effect change over time?

Table 2. Estrogen plus Progestin and the Risk of CHD, According to Year of Follow-up.*

Year of Follow-up	CHD		Hazard Ratio for CHD (95% CI)
	Estrogen-plus-Progestin Group	Placebo Group	
	no. of cases (annualized percentage)		
1	42 (0.50)	23 (0.29)	1.81 (1.09–3.01)
2	38 (0.45)	28 (0.35)	1.34 (0.82–2.18)
3	19 (0.23)	15 (0.19)	1.27 (0.64–2.50)
4	32 (0.39)	25 (0.32)	1.25 (0.74–2.12)
5	29 (0.41)	19 (0.28)	1.45 (0.81–2.59)
≥6	28 (0.37)	37 (0.56)	0.70 (0.42–1.14)

* CHD includes acute myocardial infarction (MI) necessitating hospitalization, silent myocardial infarction as determined by serial electrocardiography, and death due to CHD. There were nine silent myocardial infarctions (four in the estrogen-plus-progestin group and five in the placebo group). Hazard ratios are stratified according to age, presence or absence of a previous coronary event, and randomly assigned diet-modification group and are adjusted for previous coronary-artery bypass grafting or percutaneous transluminal coronary angioplasty. The z score for trend was –2.36 (P=0.02); the test for trend

So far I have talked a lot about experiments or randomized trials

Reminder:

Definition (Average causal effect)

A contrast of expected counterfactual outcomes in the same population of individuals under two different treatments (exposures).

- In biostatistics, we are very often interested in causal effects.
- In RCTs, adjustment of covariates is not necessary, in principle.

Simple example continued

In a simple randomized experiment (RCT),
$T^a \perp\!\!\!\perp A$ for $a = 0, 1$, and $T = AT^{a=1} + (1 - A)T^{a=0}$.
Suppose that the Cox model is correctly specified as

$$\alpha(t \mid a) = \alpha_0(t)e^{\beta^T a}.$$

This implies that $P(T > t \mid A = 1) = P(T > t \mid A = 0)^{\exp(\beta)}$, and thus

$$\begin{aligned}\exp(\beta) &= \frac{\log\{P(T > t \mid A = 1)\}}{\log\{P(T > t \mid A = 0)\}} \\ &= \frac{\log\{P(T^{a=1} > t)\}}{\log\{P(T^{a=1} > t)\}}.\end{aligned}$$

This is not a simple interpretation. Not clear why we should communicate effects on the log scale.

Example continues

Alternatively give an interpretation on the hazard scale,

$$\begin{aligned}\exp(\beta) &= \frac{\lim_{h\to 0} P(t+h > T \geq t \mid T \geq t, A=1)}{\lim_{h\to 0} P(t+h > T \geq t \mid T \geq t, A=0)} \\ &= \frac{\lim_{h\to 0} P(t+h > T^{a=1} \geq t \mid T^{a=1} \geq t)}{\lim_{h\to 0} P(t+h > T^{a=0} \geq t \mid T^{a=0} \geq t)}.\end{aligned}$$

The right hand shows side that the hazard functions with and without intervention for two separate groups of individuals; those who survive time t with treatment ($T^{a=1}$) and those who survive time t without treatment ($T^{a=0}$).

Problems with hazard ratios?

- Even in a perfect RCT, at any $t > 0$ the hazard ratio compares two groups of patients with potentially different characteristics: those who would survive to time t if assigned to $A = 1$ are not the necessarily the same as those who would survive to time t if assigned to $A = 0$.
- Hazards ratios are rate ratios, $\frac{\alpha_1(t)}{\alpha_2(t)}$, not risk ratios, $\frac{1-S_1(t)}{1-S_2(t)}$.

Simple example from randomized experiment

Simple reminder about causal inference.

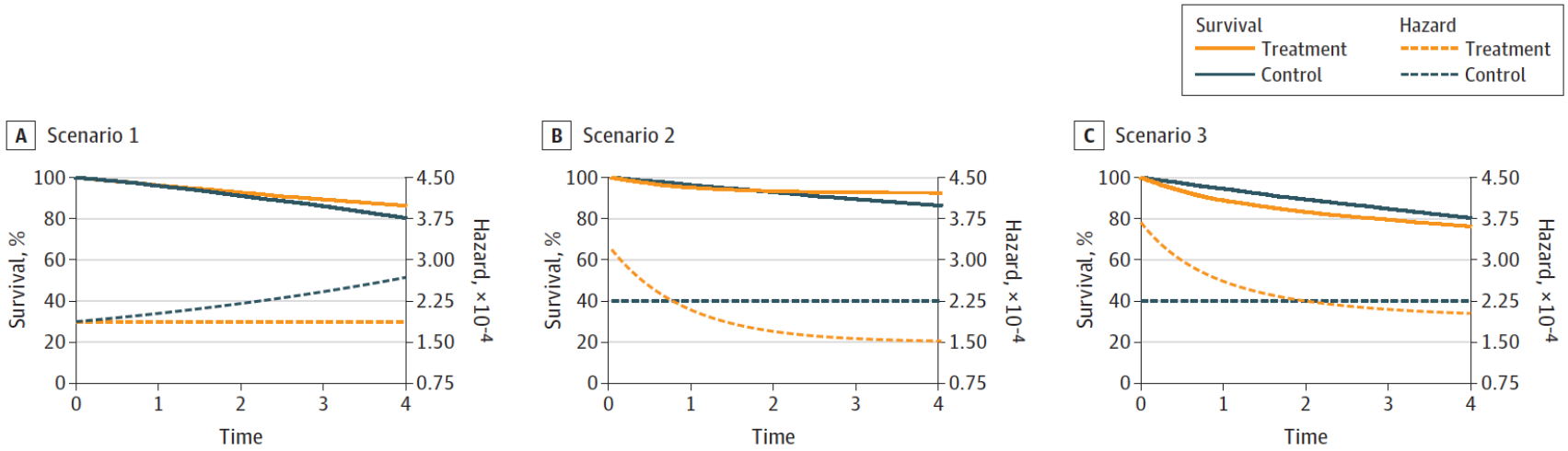
- Consider a simple binary covariate A that is randomly assigned.
- Now we define some simple counterfactual survival times $T^{a=1}$ and $T^{a=0}$.
- $T^a \in [0, \infty)$.
  The outcome variable that would have been observed under the treatment value a.
- Often we will specifically instantiate a, i.e. set a to a value:

  $T^{a=0} \in [0, \infty)$.
  The outcome variable that would have been observed under the treatment value $a = 0$.

  $Y^{a=1} \in [0, \infty)$.
  The outcome variable that would have been observed under the treatment value $a = 1$.

We suppose that: $T = AT^{a=1} + (1 - A)T^{a=0}$ (Consistency).

Illustrative examples



Mats J Stensrud and Miguel A Hernán. "Why test for proportional hazards?" In: *Jama* 323.14 (2020), pp. 1401–1402

Random censoring changes the hazard ratio

Scenario	Censoring	Hazard ratio (95% CI), Cox proportional hazards model	3-year survival difference, % (95% CI), Kaplan-Meier estimator
1	No	0.69 (0.66 to 0.72)	3.2 (2.6 to 3.8)
	Yes	0.71 (0.67 to 0.74)	3.1 (2.5 to 3.8)
2	No	0.51 (0.48 to 0.54)	3.6 (3.1 to 4.1)
	Yes	0.62 (0.58 to 0.66)	3.6 (3.0 to 4.1)
3	No	1.27 (1.22 to 1.32)	–5.2 (–5.8 to –4.5)
	Yes	1.34 (1.28 to 1.40)	–5.2 (–5.9 to –4.5)

Random censoring changes the magnitude of the hazard ratio from the Cox model (but not the magnitude of the Kaplan-Meier estimator). Be aware that the hazard ratios here are derived from mis-specified (the parameterization is clearly not correct), whereas the the Kaplan-Meier estimator is correctly specified.

Problems with the hazard ratio

- Non-collapsible (you will see this in your homework).
- Hazards ratios are rate ratios, $\frac{\alpha_1(t)}{\alpha_2(t)}$, not risk ratios, $\frac{1-S_1(t)}{1-S_2(t)}$.
- Does not say anything about the absolute risk, which is often of interest.
- Depends on censoring.

Collapsibility (heuristically)

A measure of association (such as the risk difference or the risk ratio) is said to be collapsible if the marginal measure of association is equal to a weighted average of the stratum-specific measures of association.

Collapsibility

Definition (Collapsibility of an association parameter at t)

Let $g[f(T, A)](t)$ be a function that describes the association between T and A in the joint distribution $f(T, A)$. We say that g is collapsible on a covariate V with weights $w_v(t)$ if
$\frac{\sum_v \{g[f(T,A|V=v)](t) \times w_v(t)\}}{\sum_v w_v(t)} = g[f(T, A)](t)$.

Definition (Collapsibility of a causal effect at t)

Let $h[f(T^{a=0}, T^{a=1})](t)$ be a function of $T^{a=0}$ and $T^{a=1}$ in the joint distribution $f(T^{a=0}, T^{a=1})$. We say that h is collapsible on a pre-treatment variable V with weights $w_v(t)$ if
$\frac{\sum_v \{h[f(T^{a=0},T^{a=1}|V=v)](t) \times w_v(t)\}}{\sum_v w_v(t)} = h[f(T^{a=0}, T^{a=1})](t)$.

Survival difference

Weights $P(V = v)$ because

$$\begin{aligned}
&\sum_v [P(T^{a=1} > t | V = v) - P(T^{a=0} > t | V = v)] \times P(V = v) \\
&\quad = \sum_v P(T^{a=1} > t | V = v) \times P(V = v) \\
&\qquad - \sum_v P(T^{a=0} > t | V = v) \times P(V = v) \\
&\quad = P(T^{a=1} > t) - P(T^{a=0} > t).
\end{aligned}$$

Survival ratio

Weights $P(V = v \mid T^{a=0} > t)$ because

$$\sum_v \frac{P(T^{a=1} > t \mid V = v) \times P(V = v \mid T^{a=0} > t)}{P(T^{a=0} > t \mid V = v)}$$

$$= \sum_v \frac{P(T^{a=1} > t \mid V = v) \times P(T^{a=0} > t \mid V = v) \times P(V = v)}{P(T^{a=0} > t \mid V = v) \times P(T^{a=0} > t)} \quad \text{(Bayes T)}$$

$$= \sum_v \frac{P(T^{a=1} > t \mid V = v) \times P(V = v)}{P(T^{a=0} > t)}$$

$$= \frac{\sum_v P(T^{a=1} > t \mid V = v) \times P(V = v)}{P(T^{a=0} > t)}$$

$$= \frac{P(T^{a=1} > t)}{P(T^{a=0} > t)}.$$

Counterexample for hazard ratios

Suppose that treatment $A \in \{0, 1\}$ is randomly assigned and that Z is a baseline covariate. Consider an absolutely continuous event time T with conditional hazards given by

$$\alpha(t \mid A = 0, Z) = Z\alpha(t), \quad \alpha(t \mid A = 1, Z) = rZ\alpha(t),$$

where $Z \in [0, \infty)$, $r > 0$, $\alpha(t) > 0 \ \forall t > 0$. Remember that the Laplace transform of Z is defined as

$$\mathcal{L}(c) = \mathbb{E}(e^{-cZ}).$$

for $c \in \mathbb{C}$. Thus,

$$S(t \mid A = 0) = \mathcal{L}(H(t)).$$

Taking derivatives,

$$\alpha(t \mid A = 0) = -\alpha(t)\frac{\mathcal{L}'(H(t))}{\mathcal{L}(H(t))}.$$

Counterexample continues

Suppose that Z is gamma distributed with mean 1 and variance δ, which has the Laplace transform

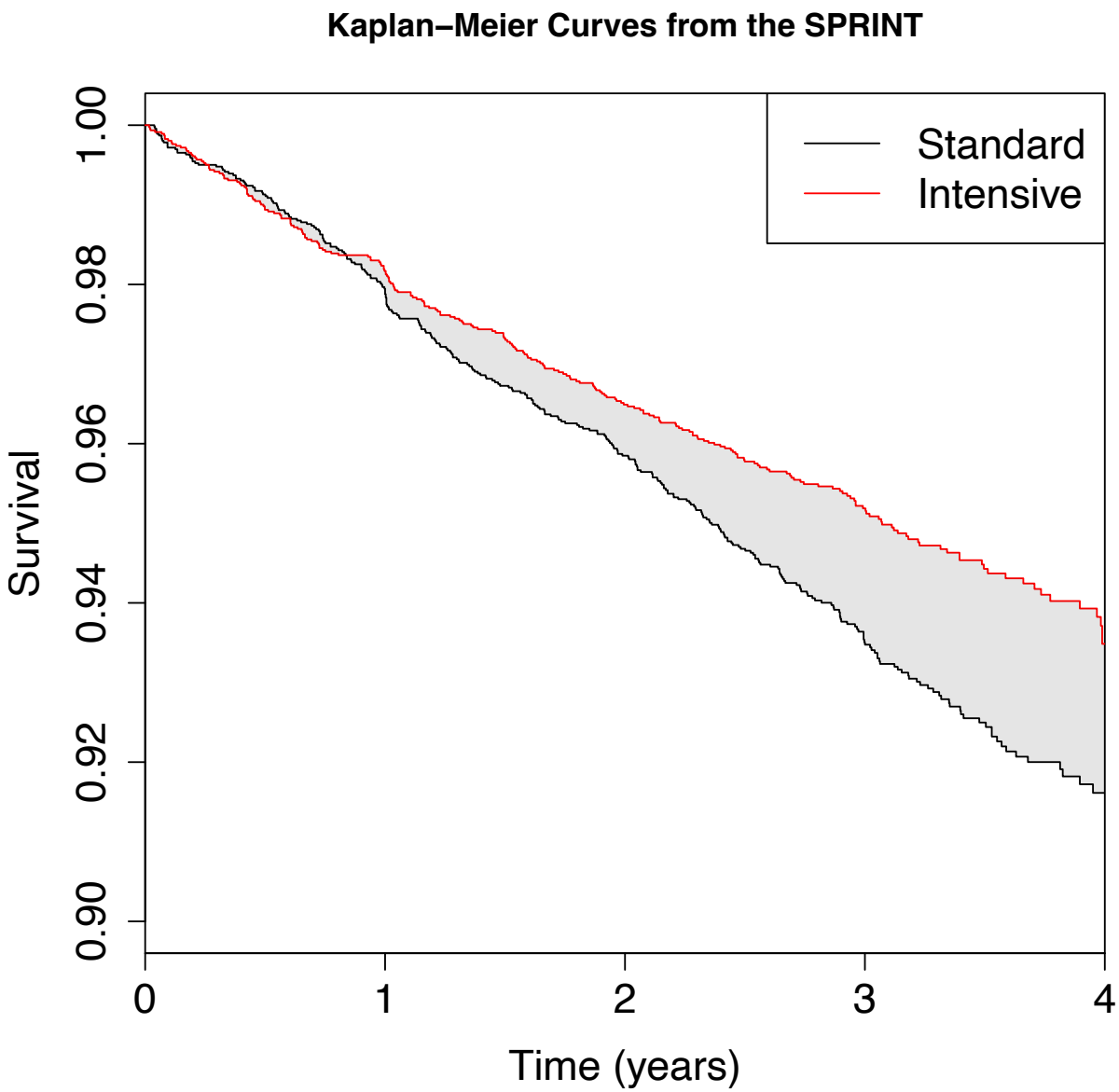
$$\mathcal{L}(c) = \{1 + \delta c\}^{\frac{-1}{\delta}},$$

where c is a complex number. Thus, $S(t) = \{1 + \delta H(t)\}^{\frac{-1}{\delta}}$, and $\alpha(t \mid A = 0) = \frac{\alpha(t)}{1+\delta H(t)}$. An identical argument gives that $\alpha(t \mid A = 1) = \frac{r\alpha(t)}{1+\delta r H(t)}$.
To finish the counterexample, select $\delta = 1$ (exponential distribution) and $r > 1$. This means that the hazard ratio is r conditional on $Z = z$, $\forall z \in [0, \infty]$. Then, for all $t > 0$,

$$\frac{\alpha(t \mid A = 1)}{\alpha(t \mid A = 0)} = r\frac{1 + H(t)}{1 + rH(t)} < r.$$

This counterexample shows that hazard ratios are not collapsible.

SPRINT study



SPRINT study

SPRINT		
Measure	**Value**	**Comments**
Hazard Ratio (HR)	0.75 (0.64, 0.89)	The clinical relevance of the treatment effect is hard to evaluate from the hazard ratio alone.
RMST difference 1 year of follow up	0.08 (-1.3, 1.1)	During the first year of follow-up, there is no significant difference in the outcome between intensive therapy and standard therapy.
RMST difference 4 year of follow up	13.2 (3.7, 22.6)	During the initial 4 years of follow-up, intensive treatment significantly delays the time to the outcome: The intensive treatment group are free of major adverse coronary event 13.2 days longer.

The hazard ratio does give a different impression of the effect compared to thethe restricted mean survival (RMST, as defined in a previous slide), which is estimated as

$$\hat{\mu}_t = \int_0^t \hat{S}(u)du,$$

and is the gray area between the curves in the previous plot.