Survival Analysis, Lecture 4 The Cox proportional hazards model

Ruth Keogh

Department of Medical Statistics
London School of Hygiene and Tropical Medicine

Aims

Broadly

- Introduce the Cox proportional hazards model
  - Also called simply the Cox model
  - Fitting this model is often called Cox regression
  - The most widely used model for analysis of survival data
- It is used specifically to estimate the effects of explanatory variables on survival
- It is a semi-parametric model and a special analysis is required
  - partial likelihood

Aims

Part 1

- Introduce the Cox proportional hazards model
- How we estimate the parameters of this model (hazard ratios) using a partial likelihood
- Interpreting the results from Cox proportional hazards models

Part 2

- Obtaining estimated survivor curves
- Handling tied event times

Part 3

- Assessing the assumptions of the Cox model
- Comparison with fully parametric models

Introduction to the Cox proportional hazards model

The Cox proportional hazards model

Regression Models and Life-Tables

By D. R. Cox

Imperial College, London

[Read before the Royal Statistical Society, at a meeting organized by the Research Section, on Wednesday, March 8th, 1972, Mr M. J. R. Healy in the Chair]

Summary

The analysis of censored failure times is considered. It is assumed that on each individual are available values of one or more explanatory variables. The hazard function (age-specific failure rate) is taken to be a function of the explanatory variables and unknown regression coefficients multiplied by an arbitrary and unknown function of time. A conditional likelihood is obtained, leading to inferences about the unknown regression coefficients. Some generalizations are outlined.

Reminder: proportional hazards models in general

Form of a proportional hazards model

- Under the assumption that explanatory variables act proportionally on the hazard, the hazard function for an individual with exposure x is

$$h(t|x) = h_0(t)e^{\beta x}$$

- $h_0(t)$ is the hazard function for a baseline individual
- β is a log hazard ratio (or more generally a vector of log hazard ratios)

What is the interpretation of e^{β} for a binary variable, and for a continuous variable?

Reminder: proportional hazards models in general

Form of a proportional hazards model

$$h(t|x) = h_0(t)e^{\beta x}$$

In fully parametric models a form is assumed for $h_0(t)$

- Exponential distribution: $h_0(t) = \lambda$
- Weibull distribution: $h_0(t) = \kappa\lambda t^{\kappa-1}$

Drawback of fully parametric approaches:

- We have to specify a form for the baseline hazard, which we may not wish to do

The Cox proportional hazards model

- Cox (1972) suggested that the baseline hazard could be left unspecified
- The model for the hazard of the form

$$h(t|x) = h_0(t)e^{\beta x}$$

is called the Cox proportional hazards model when no form is given for the baseline hazard

Semi-parametric

This is called a semi-parametric model because the effect of explanatory variables on the hazard is parameterized, but $h_0(t)$ is not parameterized.

Cox proportional hazards model: Analysis

The Cox proportional hazards model: definition

$$h(t|x) = h_0(t)e^{\beta x}$$

x_i : exposure for individual $i (i = 1, \ldots, n)$

t_i : event or censoring time for individual i

δ_i : indicator of event (1) or censoring (0)

Likelihood

$$\begin{aligned} L &= \prod_{i=1}^{n} f(t_i|x_i)^{\delta_i} S(t_i|x_i)^{1-\delta_i} = \prod_{i=1}^{n} h(t_i|x_i)^{\delta_i} S(t_i|x_i) \\ &= \prod_{i=1}^{n} \left\{ h_0(t_i)e^{\beta x_i} \right\}^{\delta_i} \exp\left\{ -e^{\beta x_i} \int_0^{t_i} h_0(u)du \right\} \end{aligned}$$

Cox proportional hazards model: Analysis

Likelihood

$$L = \prod_{i=1}^{n} \left\{ h_0(t_i) e^{\beta x_i} \right\}^{\delta_i} \exp\left\{ -e^{\beta x_i} \int_0^{t_i} h_0(u) du \right\}$$

Question:

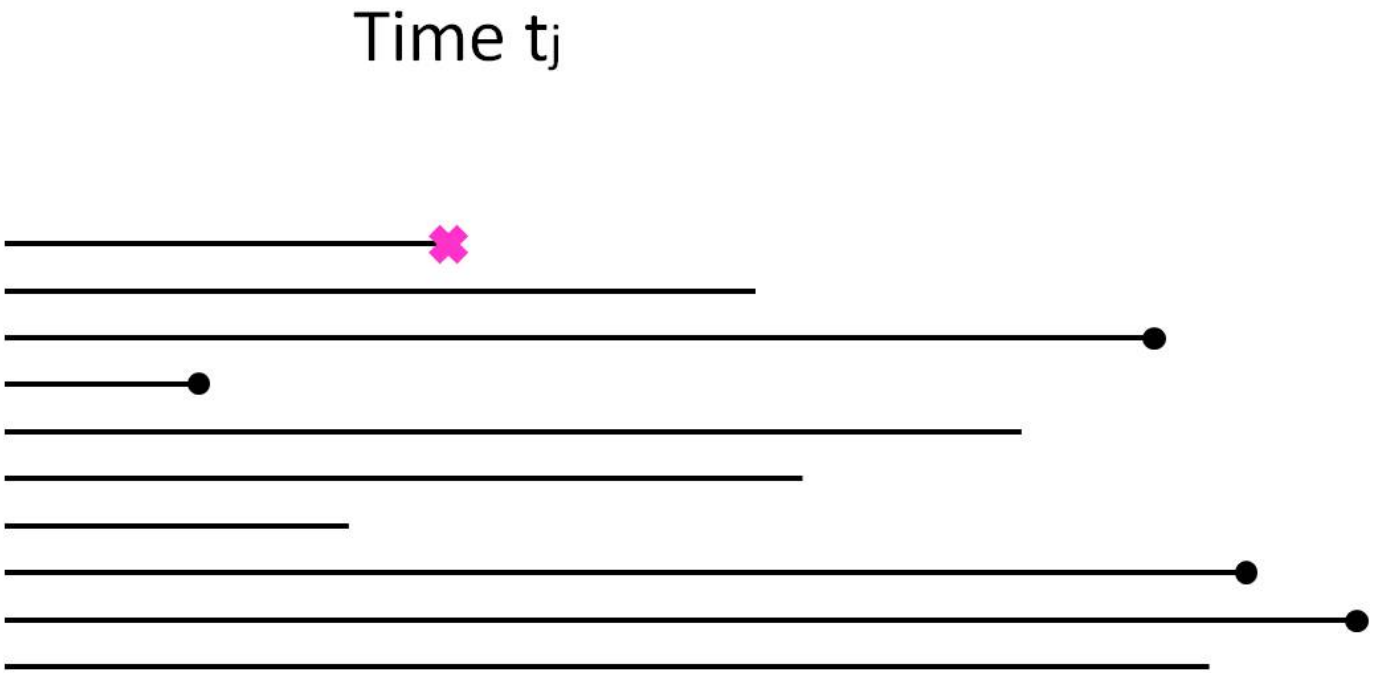
How can we use this likelihood without choosing a particular form for the baseline hazard $h_0(t)$?

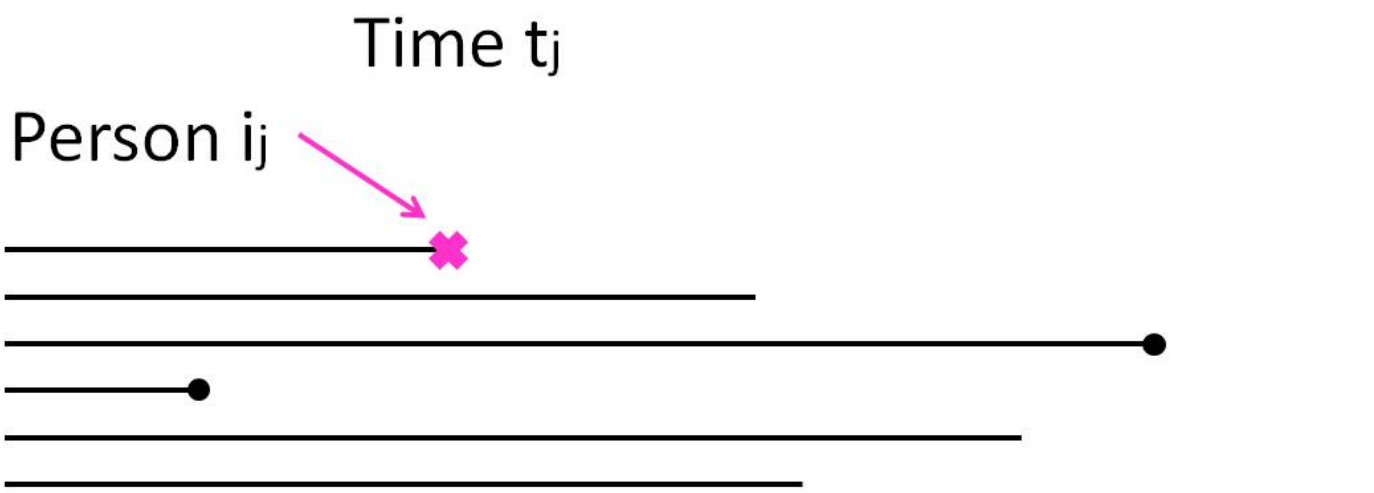
Answer:

We can't and so instead we use a special kind of analysis.

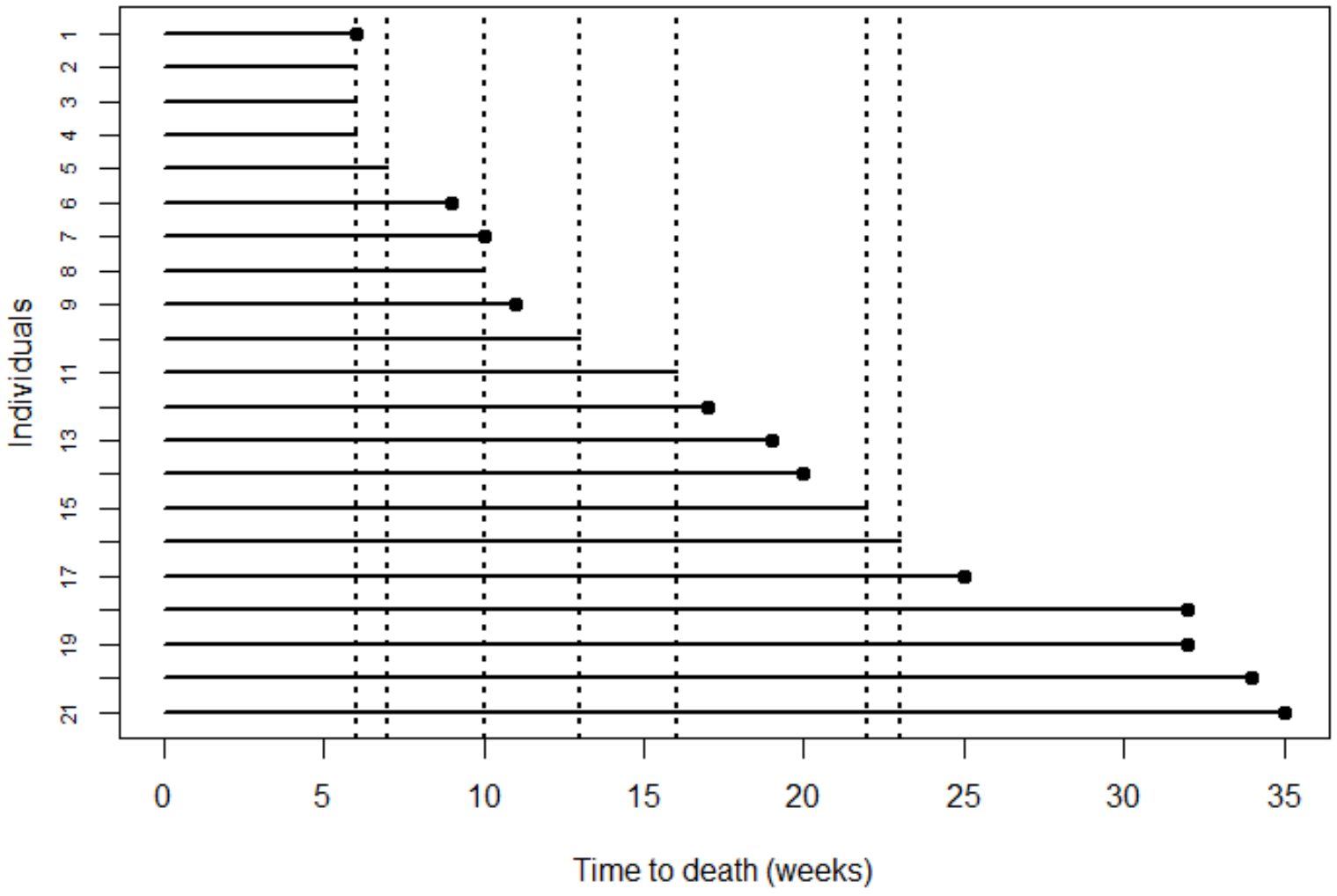
Partial likelihood

Definition: Risk sets





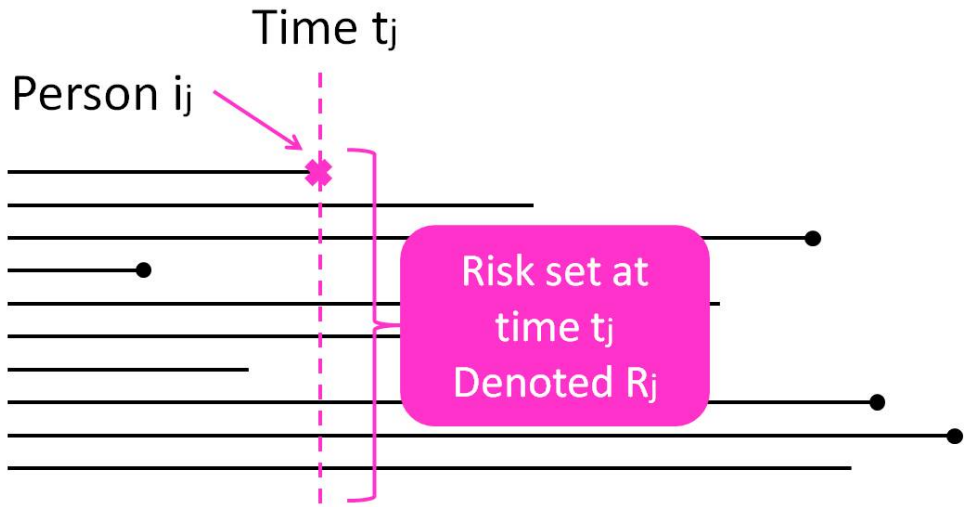
Example


Individuals
1
2
3
4
5
6
7
8
9
11
13
15
17
19
21
0
5
10
15
20
25
30
35
Time to death (weeks)

A conditional probability at each event time

Question

Given that the set of individuals R_j has survived up to time t_j without having the event or being censored, what is the probability that it was individual i_j with exposure x_{i_j} who had the event at time t_j when it might have been any one of the other individuals in the risk set R_j with their corresponding exposures?



A conditional probability at each event time

Question

Given that the set of individuals R_j has survived up to time t_j without having the event or being censored, what is the probability that it was individual i_j with explanatory variables x_{i_j} who had the event at time t_j when it might have been any one of the other individuals in the risk set R_j with their corresponding explanatory variables?

$$\frac{\text{Probability individual } i_j \text{ had the event given in risk set } R_j}{\sum_{k \in R_j} \text{Probability individual } k \text{ had the event given in risk set } R_j}$$

$$\frac{h_0(t_j)e^{\beta x_{i_j}}}{\sum_{k \in R_j} h_0(t_j)e^{\beta x_k}} = \frac{e^{\beta x_{i_j}}}{\sum_{k \in R_j} e^{\beta x_k}}$$

Combining conditional probabilities

Conditional probability at time t_j

$$\frac{\text{Probability individual } i_j \text{ had the event given in } R_j}{\sum_{k \in R_j} \text{Probability individual } k \text{ had the event given in } R_j} = \frac{e^{\beta x_{i_j}}}{\sum_{k \in R_j} e^{\beta x_k}}$$

There is a conditional probability like this at each event time $t_1, t_2, \cdots, t_K$. We take the product of these over all event times.

Partial likelihood

$$L_P = \prod_{j=1}^{K} \frac{e^{\beta x_{i_j}}}{\sum_{k \in R_j} e^{\beta x_k}}$$

This is called a partial likelihood because it is not the likelihood for the full survival process, but for part of it.

Properties of the partial likelihood

Partial likelihood

$$L_P = \prod_{j=1}^{K} \frac{e^{\beta x_{i_j}}}{\sum_{k \in R_j} e^{\beta x_k}}$$

- The partial likelihood has the same asymptotic properties as a standard likelihood.
- So we can estimate β using maximum likelihood estimation, and that the variance of the MLEs is given by the inverse of the information matrix.

Estimating the parameters β in this way is referred to as Cox regression.

Example: Leukaemia patient trial data

- We fitted the Cox proportional hazards model to the leukaemia patient data
- The explanatory variable x is treatment group ($x = 1$) or control group ($x = 0$)

The model

$$h(t|\text{control}) = h_0(t), \qquad h(t|\text{treatment}) = h_0(t)e^{\beta}$$

- $\hat{\beta} = -1.51$, $SE(\hat{\beta}) = 0.410$, 95% CI $(-2.31, -0.71)$
- $e^{\hat{\beta}} = 0.22$, 95% CI $(0.10, 0.49)$

What is the interpretation of these results?

Example: Leukaemia patient trial data

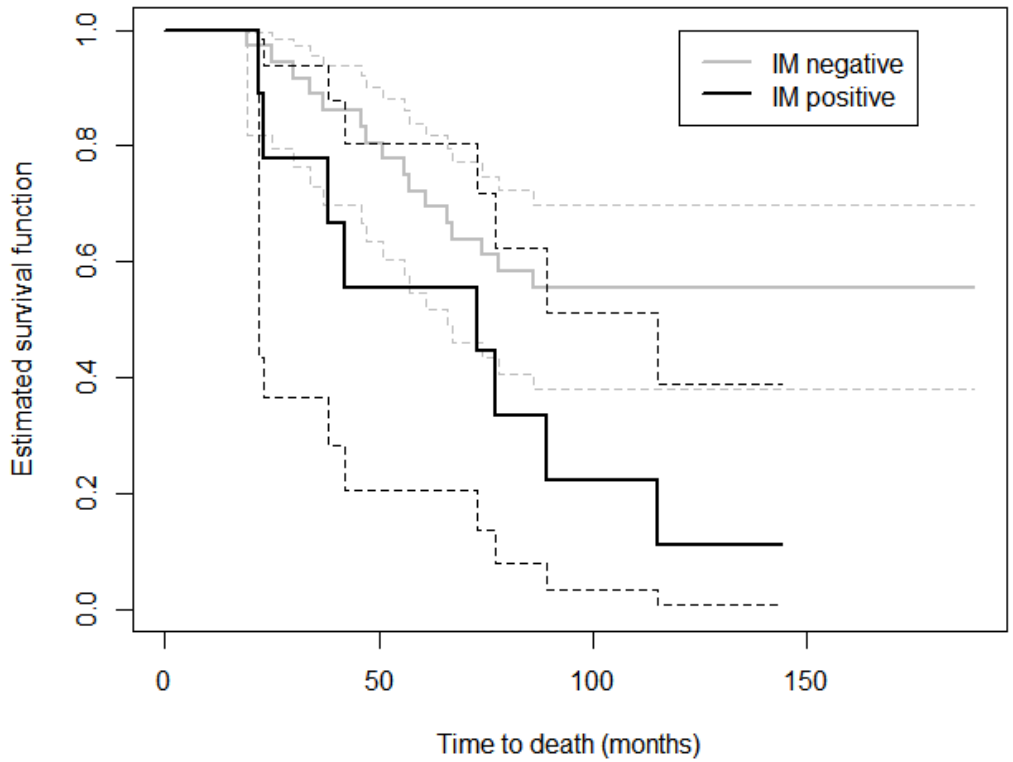
```
. stcox group

        failure _d:  death
  analysis time _t:  time

Cox regression -- Breslow method for ties

No. of subjects =           42                  Number of obs    =           42
No. of failures =           30
Time at risk    =          541
                                                LR chi2(1)       =        15.21
Log likelihood  =   -86.379622                  Prob > chi2      =       0.0001

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       group |   .2210887   .0905501    -3.68   0.000     .0990706    .4933877
------------------------------------------------------------------------------
```

Example: Leukaemia patient trial data

```
> leukaemia.cox<- coxph(Surv(time=time,event=death)~as.factor(group),
                                  data=leukaemia,ties="breslow")
> summary(leukaemia.cox)
Call:
coxph(formula = Surv(time = time, event = death) ~ as.factor(group),
    data = leukaemia, ties = "breslow")

  n= 42, number of events= 30

                      coef exp(coef) se(coef)      z Pr(>|z|)
as.factor(group)1 -1.5092    0.2211   0.4096 -3.685 0.000229 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                    exp(coef) exp(-coef) lower .95 upper .95
as.factor(group)1    0.2211      4.523   0.09907    0.4934
```

Example: Breast cancer study data

- 45 female breast-cancer patients from a cancer registry followed for death
- The time origin is the date of breast cancer diagnosis. Time is measured in months since diagnosis.
- Exposure of interest: immuno-histochemical response (positive or negative).



Example: Breast cancer study data

```
. stcox im

        failure _d:  death
  analysis time _t:  time

Cox regression -- no ties

No. of subjects =           45                 Number of obs    =           45
No. of failures =           24
Time at risk    =         4425
                                               LR chi2(1)       =         4.45
Log likelihood  =   -81.520649                 Prob > chi2      =       0.0350

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          im |   2.664988   1.158975     2.25   0.024     1.136362    6.249912
------------------------------------------------------------------------------
```

Example: Breast cancer study data

```
> breastcancer.cox<- coxph(Surv(time=time,event=death)~as.factor(im),
                                  data=breastcancer,ties="breslow")
> summary(breastcancer.cox)
Call:
coxph(formula = Surv(time = time, event = death) ~ as.factor(im),
    data = breastcancer, ties = "breslow")

  n= 45, number of events= 24

                 coef exp(coef) se(coef)     z Pr(>|z|)
as.factor(im)2 0.9802    2.6650   0.4349 2.254   0.0242 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

               exp(coef) exp(-coef) lower .95 upper .95
as.factor(im)2     2.665     0.3752     1.136      6.25
```

Example: Alcohol intake and breast cancer in the EPIC-Norfolk cohort

- EPIC-Norfolk is a cohort of 25,639 individuals who were recruited in the 1990s aged about 40 and older.
- This cohort has been followed up for disease diagnoses and death.
- A large amount of data on lifestyle exposures.

This example

- We look at times to breast cancer diagnosis (as a first cancer) among 12,576 women in this cohort.
- Women were censored if they had another cancer prior to a breast cancer or if they died from other causes
- The time scale is age.

Example: Alcohol intake and breast cancer in the EPIC-Norfolk cohort

- We are interested in studying the association between alcohol intake and breast cancer risk.
- We want to adjust for any potential confounders
- For illustration, we consider adjustment for family history of breast cancer and smoking status.

The model

- Alcohol intake was measured using a questionnaire and is measured in this example in units of 10 grams per day.
- Family history is a binary variable (yes/no).
- Smoking status is a categorical variable with three categories: never smoker, former smoker, current smoker.

$$h(t|\mathbf{x}) = h_0(t)\exp\{\beta_1 x_{\text{alc}} + \beta_2 x_{\text{FH}} + \beta_3 x_{\text{former-smoker}} + \beta_4 x_{\text{current-smoker}}\}$$

Example: results

The model

$$h(t|\mathbf{x}) = h_0(t)\exp\{\beta_1 x_{\text{alc}} + \beta_2 x_{\text{FH}} + \beta_3 x_{\text{former-smoker}} + \beta_4 x_{\text{current-smoker}}\}$$

Variable	Hazard ratio	95% confidence interval	p-value
Alcohol intake (per 10 g/day)	1.140	(1.041.1.249)	0.005
Family history (yes vs no)	1.766	(1.337,2.334)	< 0.001
Former smoker	0.875	(0.714,1.047)	0.197
Current smoker	1.001	(0.749,1.337)	0.995

Use of the Cox model

The top 100 papers. Nature explores the most-cited research of all time. https://www.nature.com/news/the-top-100-papers-1.16224

The top 100 papers

< >

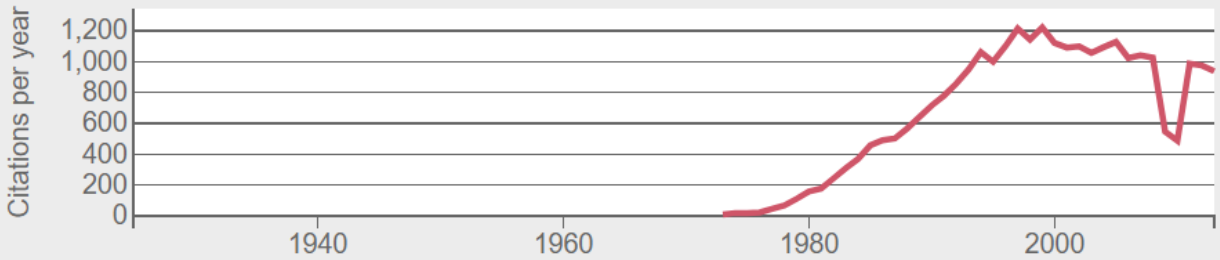
Click through to explore the Web of Science's all-time top-cited papers. (Data provided by Thomson Reuters, extracted on 7 October 2014).

Rank: **24** Citations: **28,439**

Regression models and life-tables.

Cox, D. R.

J. R. Stat. Soc., B **34**, 187–220 (1972).



Estimating the survivor curve

Estimating survivor curves

- Very often when we perform an analysis of survival data, the thing we are most interested in is the association between explanatory variables and survival.
  - Inference about this association comes from the hazard ratio estimates $e^{\hat{\beta}}$ and the corresponding p-values/CIs
- But sometimes we are interested in describing the survival of people in our study according to their explanatory features
  - to make predictions about the survival of an individual with particular explanatory variables
  - to compare survival probabilities for individuals with different characteristics
- This requires information about the baseline hazard

Estimating the survivor function

Recall the relationship

$$S(t|x) = \exp\left\{-H_0(t)e^{\beta x}\right\}$$

The estimated survivor curve from a Cox proportional hazards model for a person with explanatory variables x is

$$\widehat{S}(t|x) = \exp\left\{-\widehat{H}_0(t)e^{\hat{\beta} x}\right\}$$

Estimating survivor curves from a Cox ph model

- Consider ordered event times $t_1 < t_2 < \cdots < t_K$.
- The key to estimating the survivor curve is to first estimate the baseline cumulative hazard
  - this is the cumulative hazard at reference values for the explanatory variables
- e.g. reference values x^*. Often $x^* = 0$

Estimated baseline cumulative hazard

$$\widehat{H}_0(t) = \sum_{j=1}^{K} \frac{d_j}{\sum_{k \in R_j} \exp\{\hat{\beta}(x_k - x^*)\}} \qquad t_k \leq t < t_{k+1}$$

This is called Breslow's Estimate

Example: Leukaemia patient trial data

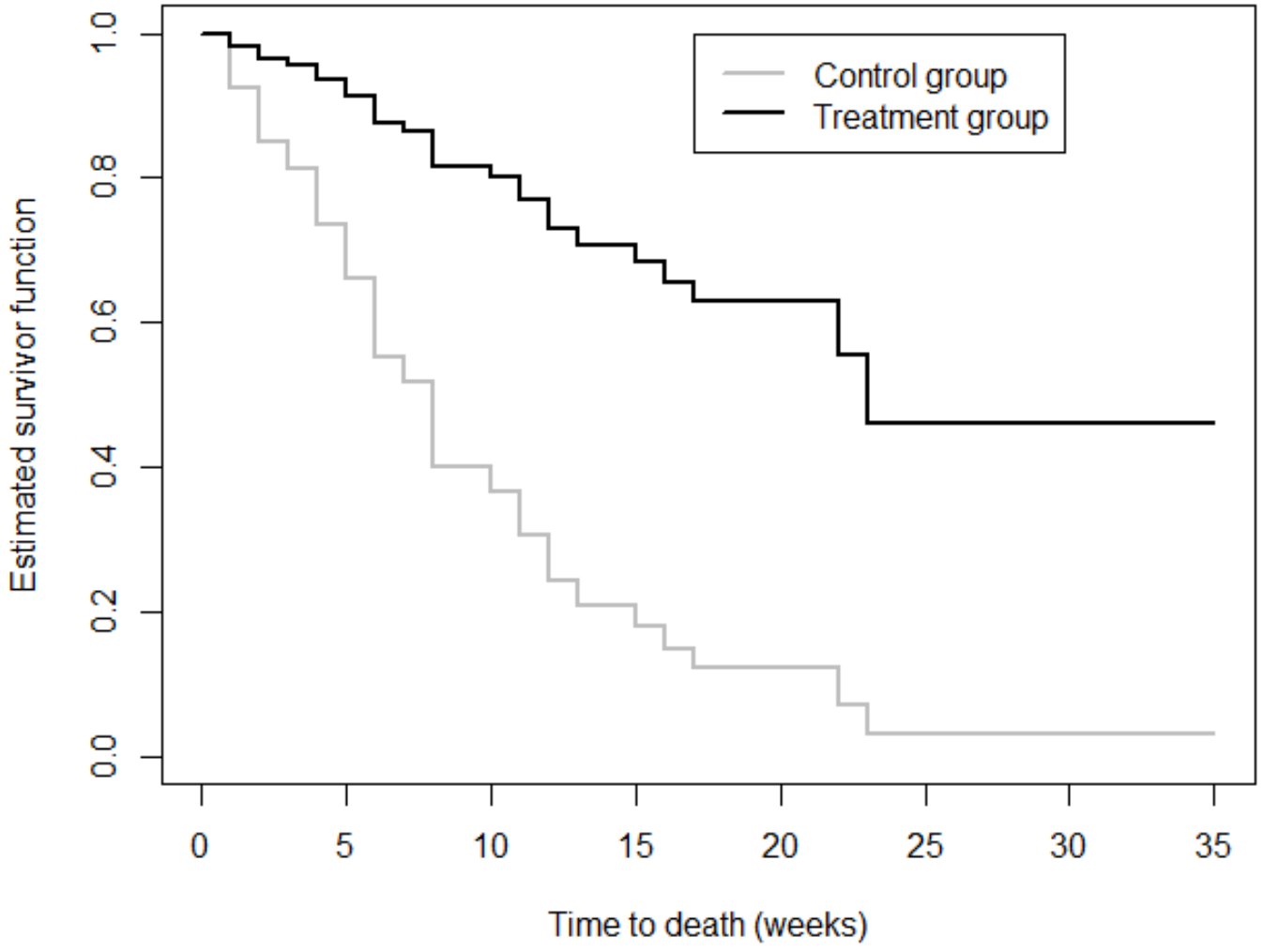
- We fitted the Cox proportional hazards model to the leukaemia patient data
- The explanatory variable x is treatment group ($x = 1$) or control group ($x = 0$)

The model

$$h(t|\text{control}) = h_0(t), \qquad h(t|\text{treatment}) = h_0(t)e^{\beta}$$

- $\hat{\beta} = -1.51$, $SE(\hat{\beta}) = 0.410$, 95% CI $(-2.31, -0.71)$
- $e^{\hat{\beta}} = 0.22$, 95% CI $(0.10, 0.49)$

Example: Leukaemia patient data: estimated survivor curves



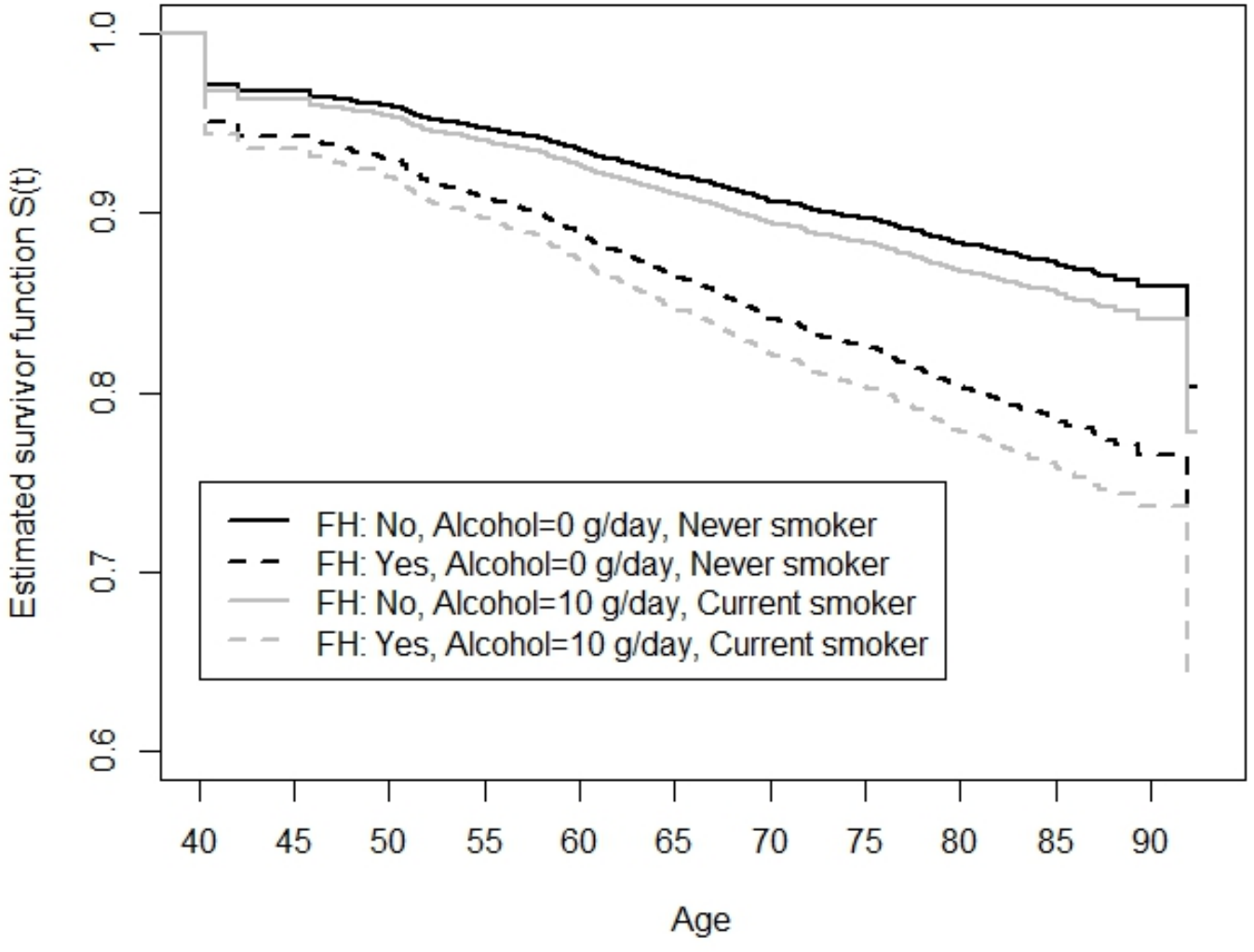
Example: Breast cancer risk in the EPIC-Norfolk cohort

The model

$$h(t|\mathbf{x}) = h_0(t)\exp\{\beta_1 x_{\text{alc}} + \beta_2 x_{\text{FH}} + \beta_3 x_{\text{former-smoker}} + \beta_4 x_{\text{current-smoker}}\}$$

Variable	Hazard ratio	95% confidence interval	p-value
Alcohol intake (per 10 g/day)	1.140	(1.041.1.249)	0.005
Family history (yes vs no)	1.766	(1.337,2.334)	< 0.001
Former smoker	0.875	(0.714,1.047)	0.197
Current smoker	1.001	(0.749,1.337)	0.995

Example: Breast cancer in the EPIC-Norfolk cohort



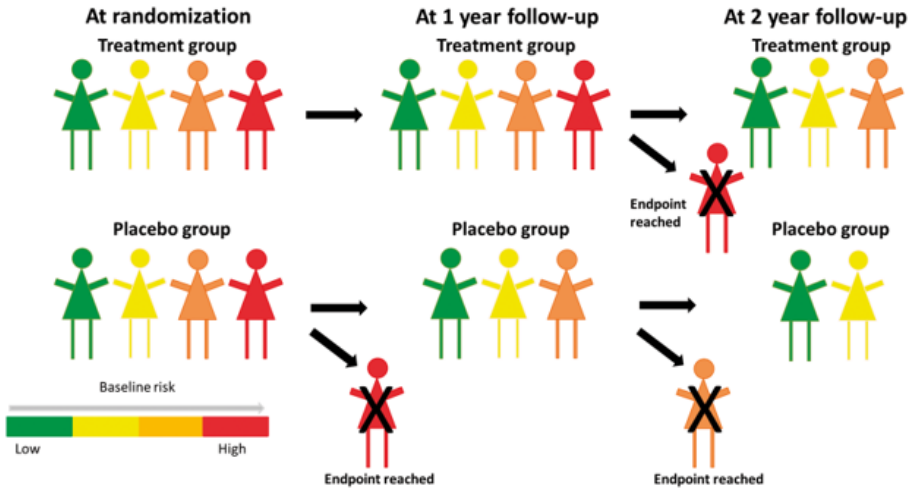
Beyond the hazard ratio

- The results from a Cox regression are typically presented in terms of the hazard ratio
- Over the past 10 years or so, a series of papers have been written which explain that a hazard ratio does not have a straightforward interpretation in terms of a ‘causal effect’.
- Hernan MA. The hazards of hazard ratios. Epidemiology 2010; 21; 13-15.

Beyond the hazard ratio

Stensrud MJ, Aalen JM, Aalen OO, Valberg M. Limitations of hazard ratios in clinical trials. European Heart Journal, 2019; 40: 1378–1383.

People have different disease risks. Given an effective treatment, high-risk individuals are likely to be depleted from the study population, and the depletion is expected to be larger in the placebo group than in the treatment group. This schematic drawing illustrates the built-in selection bias in population-level hazard ratios: by definition, the population-level hazard ratio at a given time point is based on individuals who survived up to that time point, thereby it is a comparison between the unbalanced groups.



Beyond the hazard ratio

A solution is to quantify the effects of covariates not in terms of a hazard ratio, but in terms of another quantity which does not suffer the above issues.

Risk difference at time t^*

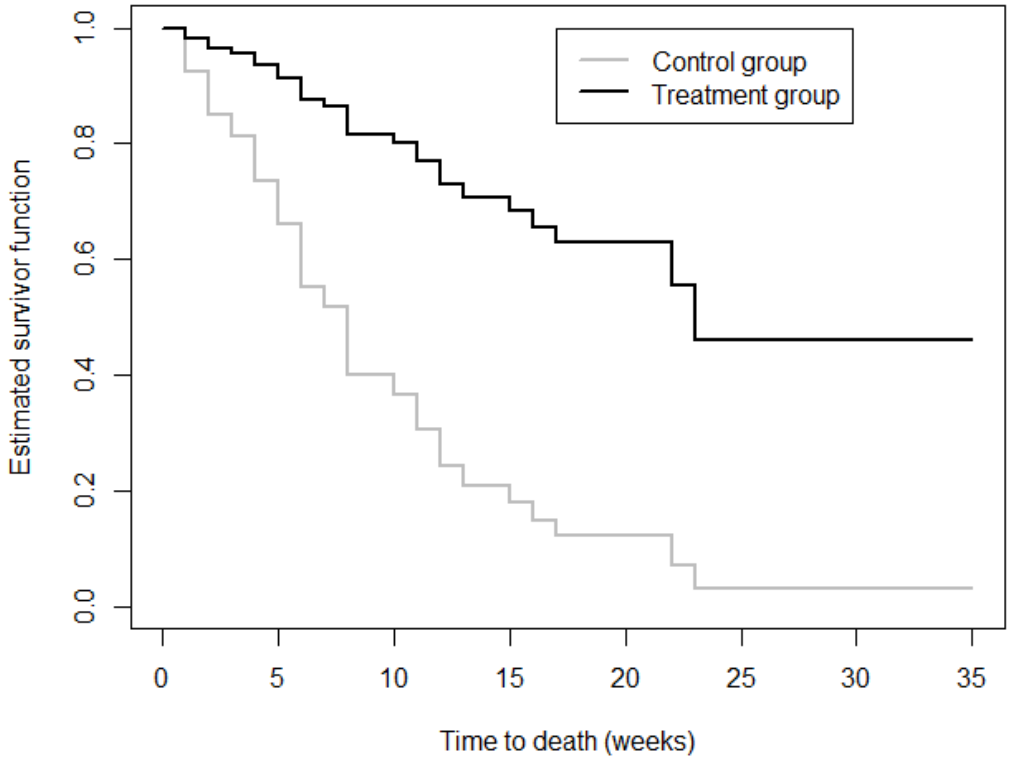
$$\begin{aligned}\Pr(T \leq t^* | X = 1) - \Pr(T \leq t^* | X = 0) &= (1 - S(t^* | X = 1)) - (1 - S(t^* | X = 0)) \\ &= S(t^* | X = 0) - S(t^* | X = 1)\end{aligned}$$

Risk ratio at time t^*

$$\frac{\Pr(T \leq t^* | X = 1)}{\Pr(T \leq t^* | X = 0)} = \frac{1 - S(t^* | X = 1)}{1 - S(t^* | X = 0)}$$

The risk difference and risk ratio can be estimated following a Cox regression by making use of the estimated survival probabilities.

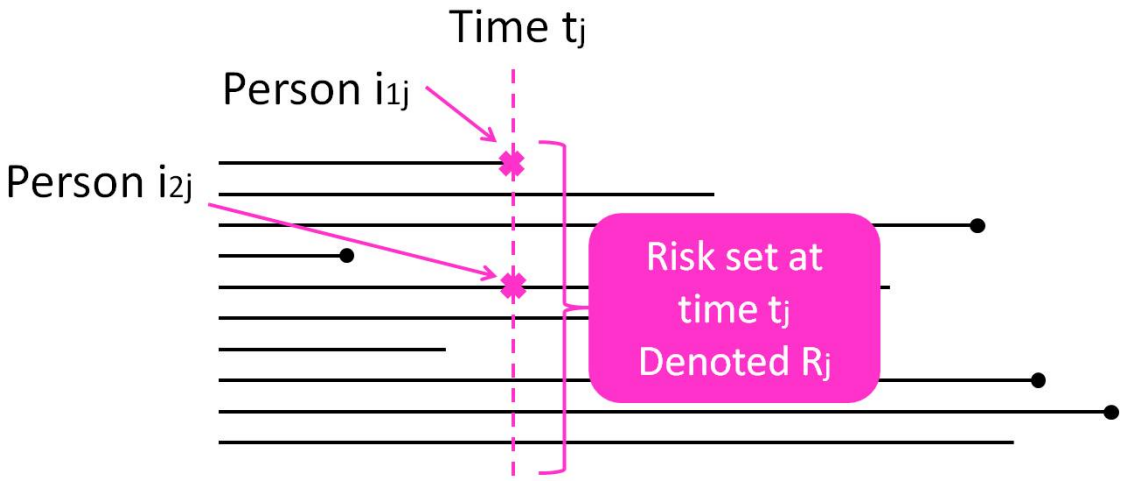
Example: Leukaemia patient data



- Risk difference at time 5: 0.25
- Risk difference at time 10: 0.43

Handling tied survival times

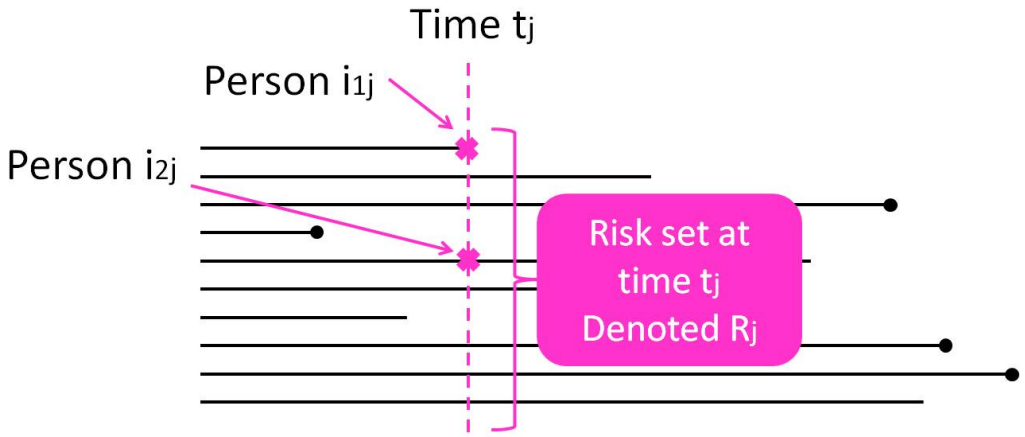
- In some studies, there will be tied survival times, that is, some individuals will have the event recorded at the same time.
- This can be incorporated into the partial likelihood analysis used for the Cox proportional hazards model.
- Suppose that at a particular time t_j there are m individuals who have the event.



Tied survival times: conditional probability

Question

Given that the set of individuals R_j have survived up to time t_j without having the event of being censored, what is the probability that it was the set of individuals $i_1, i_2, \ldots, i_m$ with vectors of exposures $x_{i1}, x_{i2}, \ldots, x_{im}$ who had the event at time t_j when it might have been any other set of size m formed from the individuals in the risk set R_j?



Tied survival times: conditional probability

Question

Given that the set of individuals R_j have survived up to time t_j without having the event of being censored, what is the probability that it was the set of individuals $i_1, i_2, \ldots, i_m$ with vectors of explanatory variables $x_{i1}, x_{i2}, \ldots, x_{im}$ who had the event at time t_j when it might have been any other set of size m formed from the individuals in the risk set R_j?

$$\frac{h_0(t_j)e^{\beta x_{i1}} \times h_0(t_j)e^{\beta x_{i2}} \times \ldots \times h_0(t_j)e^{\beta x_{im}}}{\sum_{L \in R_{mj}} \prod_{k \in L} h_0(t_j)e^{\beta x_k}}$$

$$= \frac{\exp\left(\beta x_{i1} + \beta x_{i2} + \cdots + \beta x_{im}\right)}{\sum_{L \in R_{mj}} \prod_{k \in L} e^{\beta x_k}}$$

The partial likelihood is the product of these terms over all event times

Partial likelihood for tied survival times

$$L_P = \prod_j \frac{\exp\left(\beta x_{i1_j} + \beta x_{i2_j} + \cdots + \beta x_{im_j}\right)}{\sum_{L \in R_{mj}} \prod_{k \in L} e^{\beta x_k}}$$

This becomes computationally difficult to fit if there are a few ties

An approximation

$$L_{P*} = \prod_j \frac{\exp\left(\beta x_{i1_j} + \beta x_{i2_j} + \cdots + \beta x_{im_j}\right)}{\left\{\sum_{k \in R_j} \exp \beta x_k\right\}^{m_j}}$$

This is called Breslow's approximation

Assessing the assumptions of the Cox proportional hazards model

What are we assuming?

Assumptions

1. The proportional hazards assumption: that the explanatory variables act on survival in such a way that the hazard ratio is constant over time
   - i.e. that the model $h(t|x) = h_0(t)e^{\beta x}$ is correct.
2. The assumption that we have correctly specified the form for how the explanatory variables act on the hazard
   - e.g. $h(t|x) = h_0(t)e^{\beta x}$ versus $h(t|x) = h_0(t)e^{\beta_1 x + \beta_2 x^2}$.

As for any other regression model we should check these assumptions

Other assumptions:

- Censoring is uninformative for the event of interest
- Individuals are independent

Assessing the proportional hazards assumption

The proportional hazards assumption

1. The proportional hazards assumption: that the explanatory variables act on survival in such a way that the hazard ratio is constant over time.

- For the remainder of this lecture we focus on this assumption
  - with a focus on simple methods of assessment
- In Lecture 5 you will cover some additional methods for investigating this assumption
- ...and a method for assessing whether you have chosen the right functional form for continuous explanatory variables

Assessing the proportional hazards assumption

Recall: Assessing whether a Weibull model is appropriate

Under a Weibull model we found

$$\log\{-\log S(t|x)\} = \log\lambda + \kappa\log t + \beta x$$

What kind of plot can be used to assess whether a Weibull model is appropriate?

Extending this idea to the Cox proportional hazards model

The survivor function:

$$S(t|x) = \exp\left\{-H_0(t)e^{\beta x}\right\}$$

Take logs:

$$-\log S(t|x) = H_0(t)e^{\beta x}$$

Take logs again:

$$\log\left\{-\log S(t|x)\right\} = \log H_0(t) + \beta x$$

What sort of plot would be useful?

Assessing the Cox proportional hazards assumption

Result

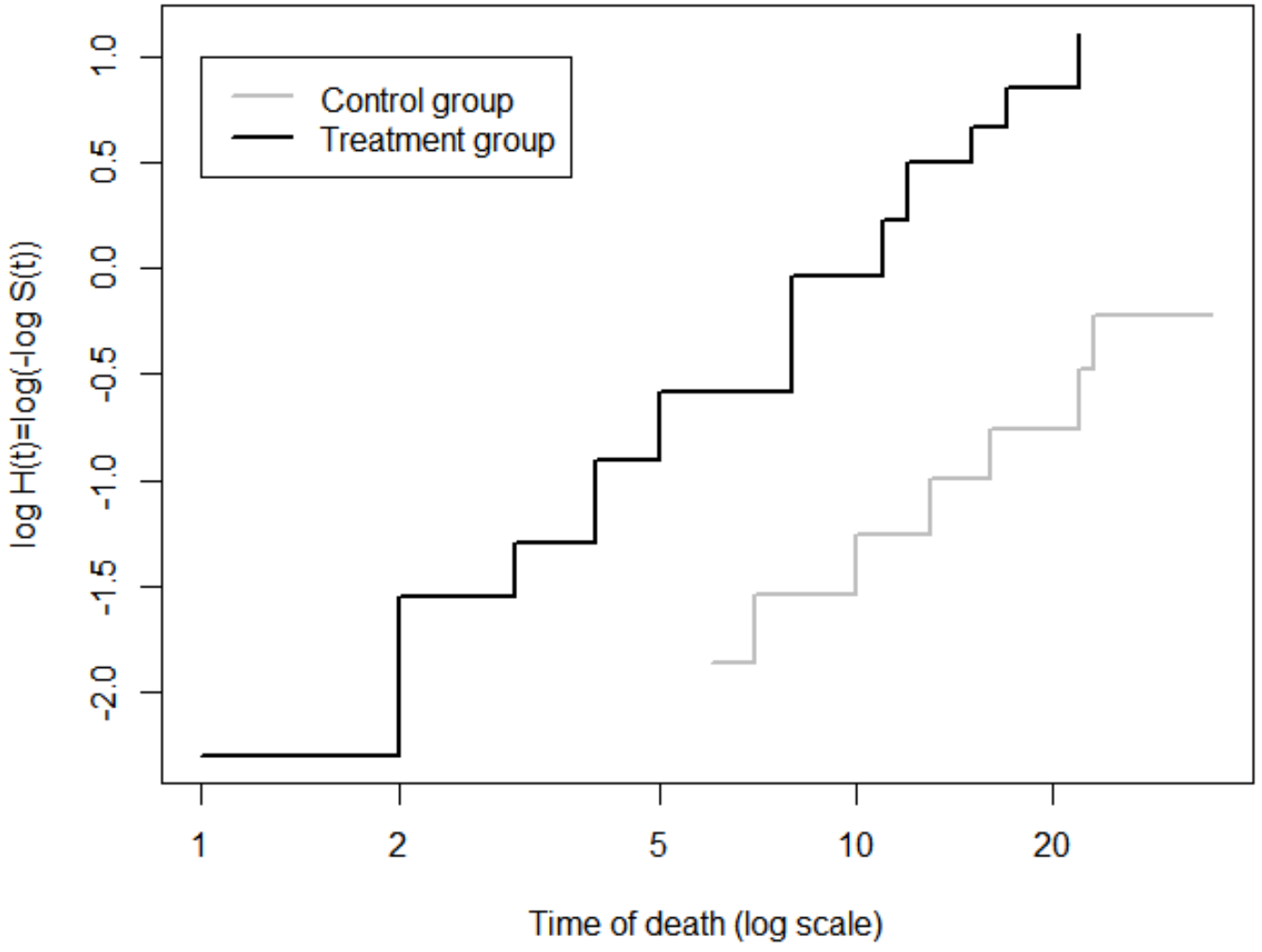
$$\log\{-\log S(t|x)\} = \log H_0(t) + \beta x$$

- Consider a binary variable x
- Make non-parametric plots of

$$\log\{-\log S(t|0)\} \text{ and } \log\{-\log S(t|1)\}$$

  against time (or some other function of time, such as log time)
- The curves for the $X = 0$ and $X = 1$ groups should appear to be approximately parallel over time if the proportional hazards assumption is met.
- If the two curves cross then this is clear evidence against the proportional hazards assumption

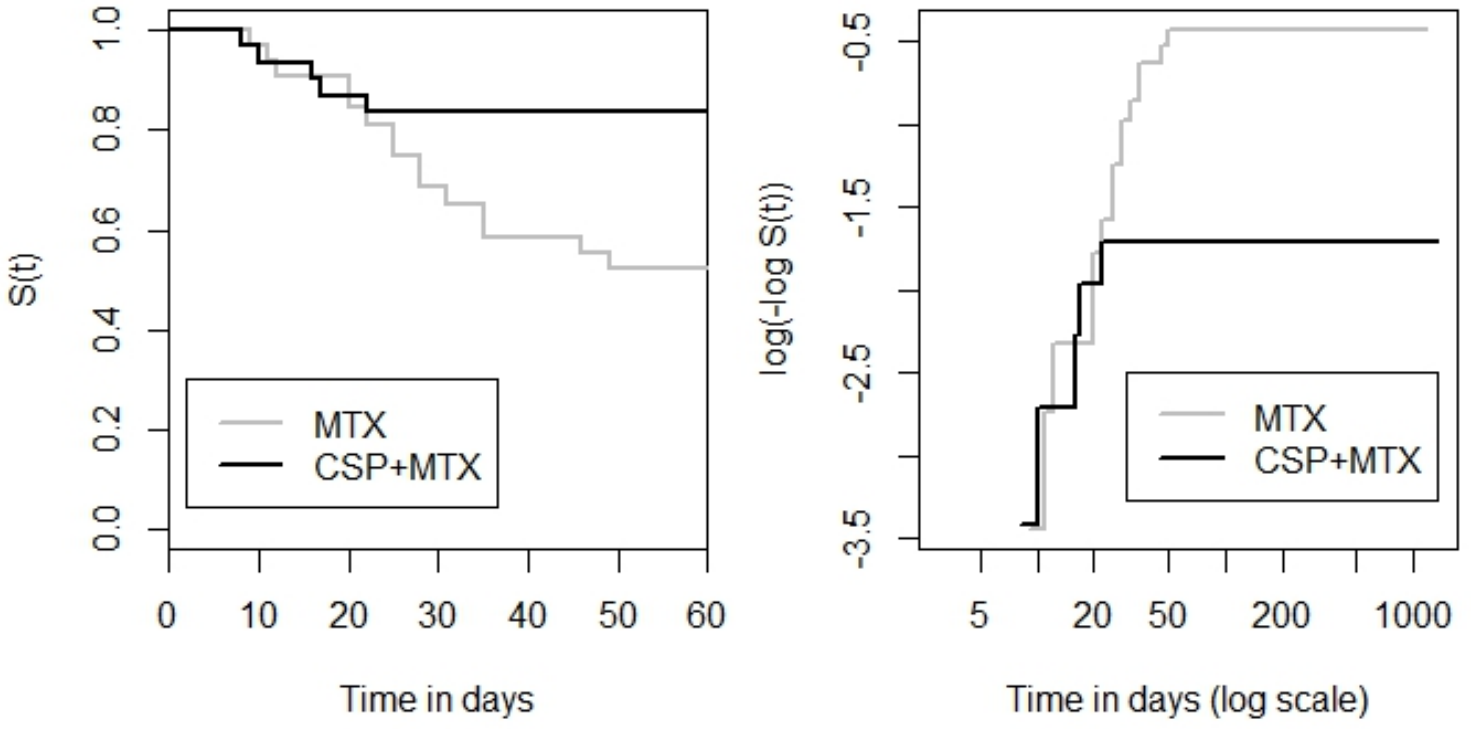
Example: Leukaemia patient data: assessing the proportional hazards assumption



Example: Acute graft versus host disease (AGVHD) data

- Randomized trial involving 64 patients with severe aplastic anaemia who had received bone marrow donated by a family member
- Two treatment groups
  - one group received methotrexate alone (MTX)
  - the other group received methotrexate plus cyclosporine (CSP+MTX)
- The patients were followed up until diagnosis of a life-threatening stage of acute graft versus host disease (AGVHD), death or censoring (this was the time of the last contact with the patient, which was after several years of follow-up in some cases)

Example: Acute graft versus host disease (AGVHD) data



What do you think?

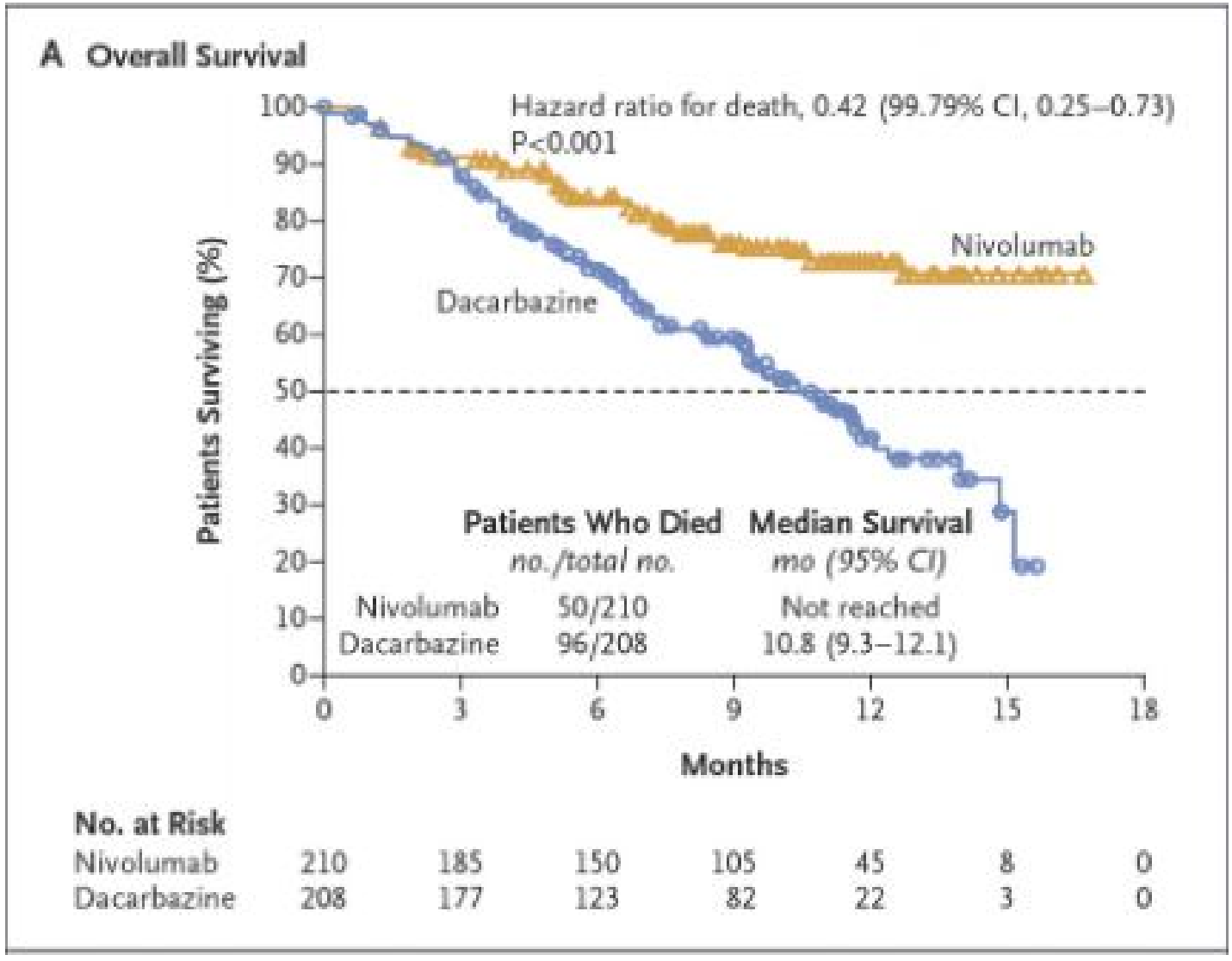
The NEW ENGLAND JOURNAL of MEDICINE

ORIGINAL ARTICLE

Nivolumab in Previously Untreated Melanoma without *BRAF* Mutation

Caroline Robert, M.D., Ph.D., Georgina V. Long, M.D., Ph.D., Benjamin Brady, M.D., Caroline Dutriaux, M.D., Michele Maio, M.D., Laurent Mortier, M.D., Jessica C. Hassel, M.D., Piotr Rutkowski, M.D., Ph.D., Catriona McNeil, M.D., Ph.D., Ewa Kalinka-Warzocha, M.D., Ph.D., Kerry J. Savage, M.D., Micaela M. Hernberg, M.D., Ph.D., Celeste Lebbé, M.D., Ph.D., Julie Charles, M.D., Ph.D., Catalin Mihalcioiu, M.D., Vanna Chiarion-Sileni, M.D., Cornelia Mauch, M.D., Ph.D., Francesco Cognetti, M.D., Ana Arance, M.D., Ph.D., Henrik Schmidt, M.D., D.M.Sc., Dirk Schadendorf, M.D., Helen Gogas, M.D., Lotta Lundgren-Eriksson, M.D., Christine Horak, Ph.D., Brian Sharkey, Ph.D., Ian M. Waxman, M.D., Victoria Atkinson, M.D., and Paolo A. Ascierto, M.D.

N Engl J Med 2015; 372:320-330 | January 22, 2015 | DOI: 10.1056/NEJMoa1412082


A Overall Survival
Hazard ratio for death, 0.42 (99.79% CI, 0.25–0.73)
P<0.001
Nivolumab
Dacarbazine
Patients Surviving (%)
100
90
80
70
60
50
40
30
20
10
0
Patients Who Died
no./total no.
Median Survival
mo (95% CI)
Nivolumab 50/210 Not reached
Dacarbazine 96/208 10.8 (9.3–12.1)
0 3 6 9 12 15 18
Months
No. at Risk
Nivolumab 210 185 150 105 45 8 0
Dacarbazine 208 177 123 82 22 3 0

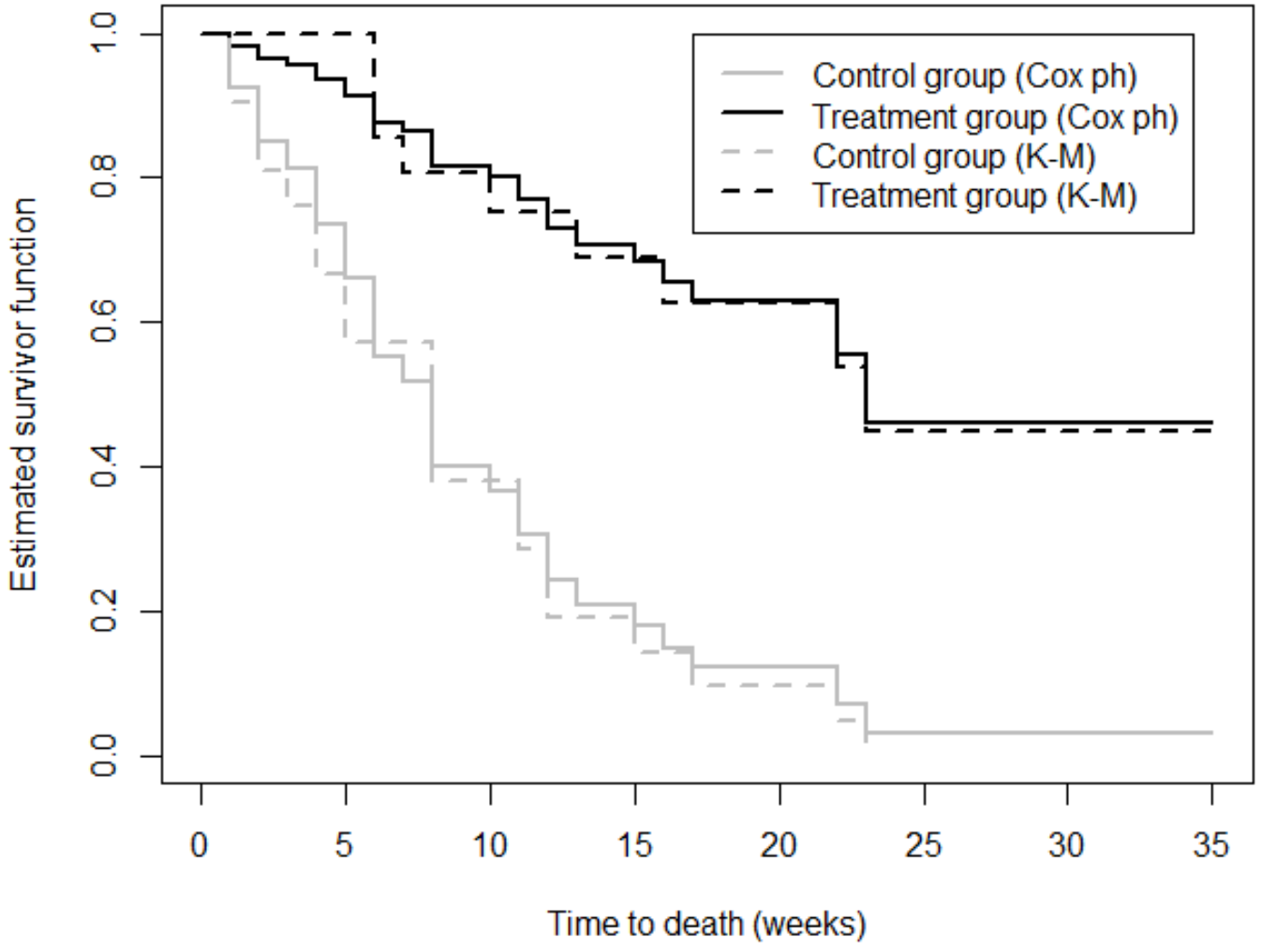
Making comparisons with non-parametric estimates

As a check of whether the proportional hazards assumption is appropriate...

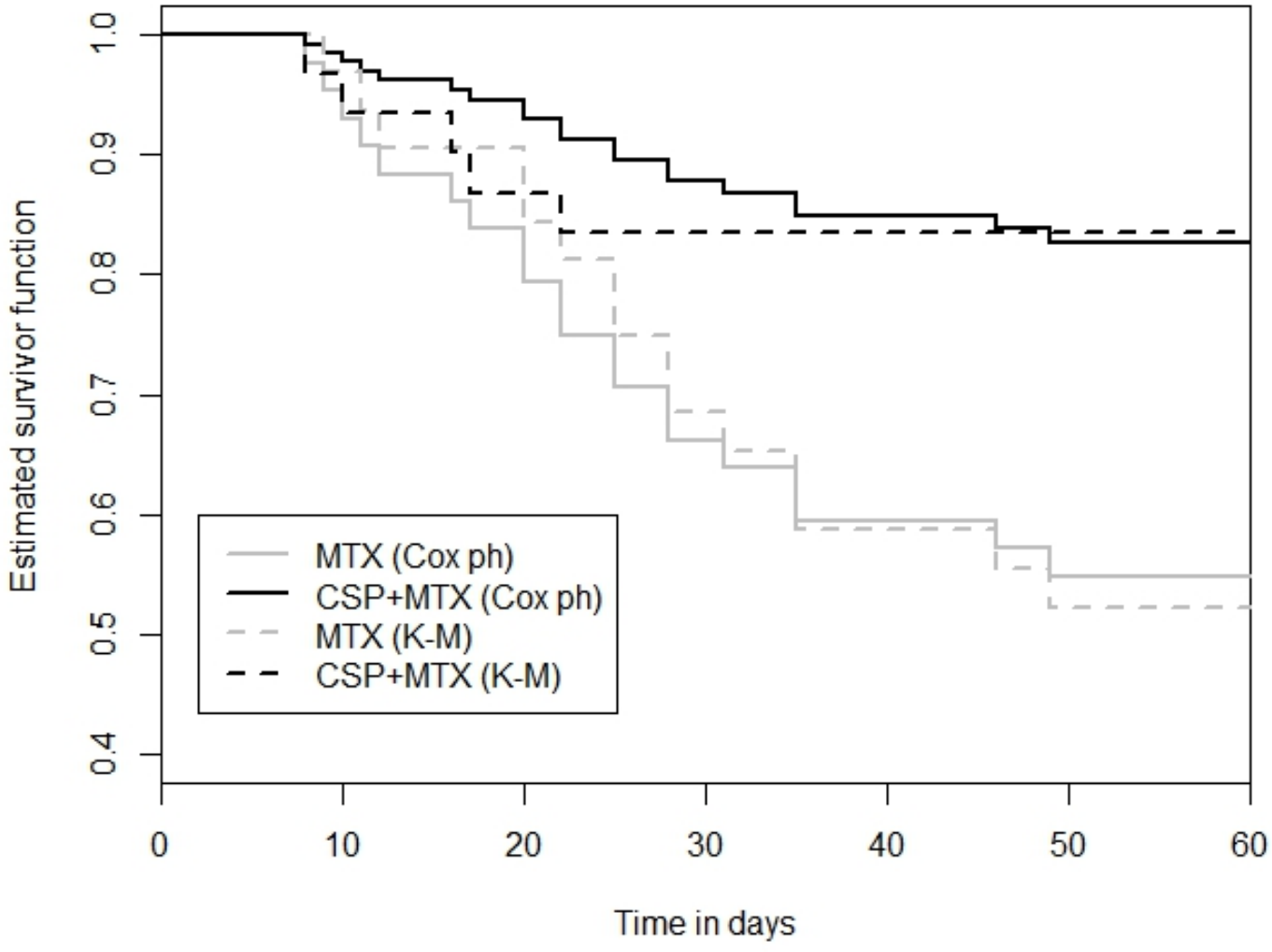
We can compare the survivor curves estimates from the fitted Cox proportional hazards model with the Kaplan-Meier estimates

- Systematic differences between the two plots suggest that the proportional hazards assumption is not met
- This only works for binary or categorical variables obviously

Example: Leukaemia patients data: assessing the proportional hazards assumption



Example: Acute graft versus host disease (AGVHD) data



Extensions

- Assessment of the proportional hazards assumption using non-parametric plots is very useful
- ...though it depends on the explanatory variables being categorical
- Things also get tricky when we have several explanatory variables to consider simultaneously
  - in observational studies this is nearly always the case
- In the next lecture you will learn about alternative methods which can be used more generally.

To be semi-parametric or fully-parametric?

- We have focused in this lecture on the Cox proportional hazards model.
- An alternative is to use a fully-parametric model, e.g. exponential or Weibull model.

What do you think are the advantages and disadvantages of the two approaches?

To be semi-parametric or fully-parametric?

- The Cox model uses fewer assumptions...but a fully parametric model may often be perfectly appropriate:
- If an exponential or Weibull model is OK then a Cox PH model is also OK
- The two approaches will give very similar hazard ratio estimates
- The fully-parametric approach may give slightly more precise estimates but the gain in precision will usually be small.
- ... but this is traded off against the concern that the baseline hazard may have been mis-specified

Flexible parametric survival models

Royston P, Parmar MKB. Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. Statistics in Medicine 2002; 21: 2175-2197.

- Royston and Parmar introduced a new class of parametric survival models, referred to as 'flexible parametric survival models'
- The baseline hazard is modelled smoothly using splines:

$$h(t|x) = h_0(t)e^{\beta x}$$

- These combine some advantages of the Cox model and some of fully parametric models and have started to become popular