How to interpret hazards and hazard ratios

Jonathan Bartlett
LSHTM
thestatsgeek.com

Graham Dunn Seminar
Current methods for time-to-event data
7th May 2025

Hazard and the Cox model

Interpreting changes in hazard and HRs

How to interpret hazards and hazard ratios

Conclusions

Acknowledgements

Thanks to Tim Morris (UCL) and Dominic Magirr (Novartis) for ongoing discussions and work on this topic.

Hazard and the Cox model

Hazard and the Cox model

We begin by briefly reviewing the definition of hazard and Cox's proportional hazards model.

Suppose we have a randomised trial with individuals randomised to control ($trt = 0$) or active treatment ($trt = 1$).

We measure time until experiences an event of interest.

T records the time to event for a given individual.

Almost always we don't observe the event time for all. These times are then right-censored.

The hazard function

The hazard function at time t, $h(t)$ is the instantaneous rate of failure *among survivors* at time t:

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

It is the failure rate at time t conditional on not yet failing (the $T \geq t$ bit).

Suppose at time t, $N(t)$ individuals are still at risk.

For small δ, we expect $N(t) \times \delta h(t)$ to fail in the next period of time δ.

The Cox model

Cox's famous model, if we put only TRT as covariate, assumes that

$$h(t|trt) = h_0(t)\exp(\beta trt)$$

$h_0(t)$ is an arbitrary baseline hazard function.

For any t, the hazard ratio (HR) comparing active to control is

$$\frac{h(t|trt=1)}{h(t|trt=0)} = \frac{h_0(t)\exp(\beta)}{h_0(t)\exp(0)} = \exp(\beta)$$

which is independent of t.

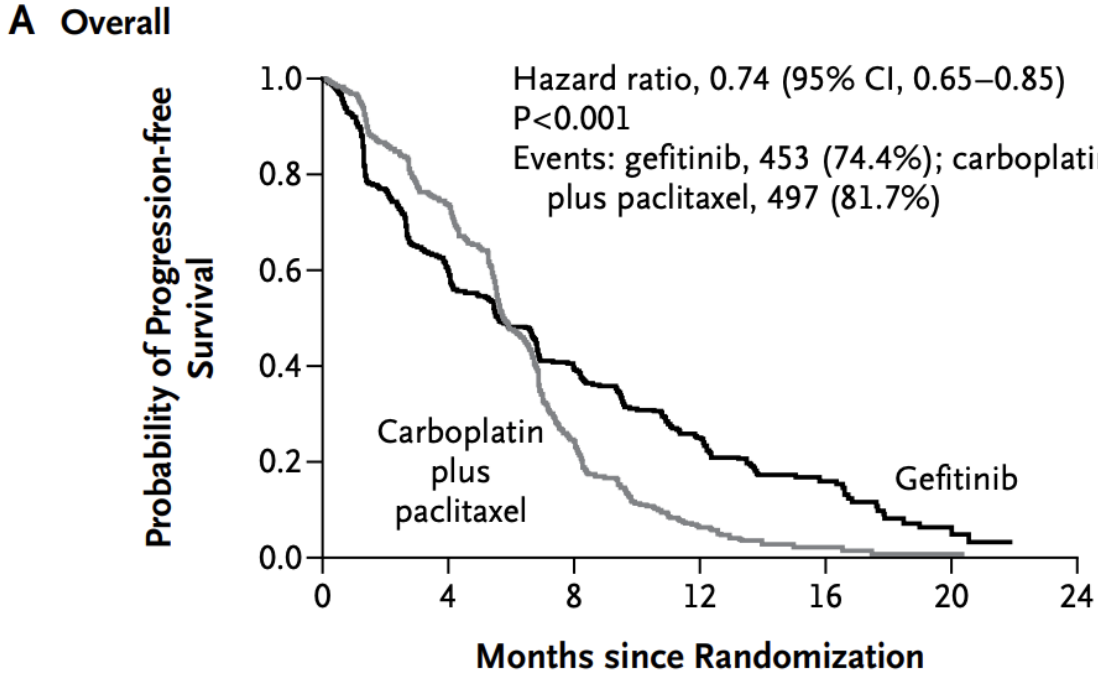
Hazards between groups are *proportional* over time (PH).

If this is violated, we say hazards are *non proportional* (NPH).

An example of NPH

The proportional hazards assumption is sometimes seen to be violated.

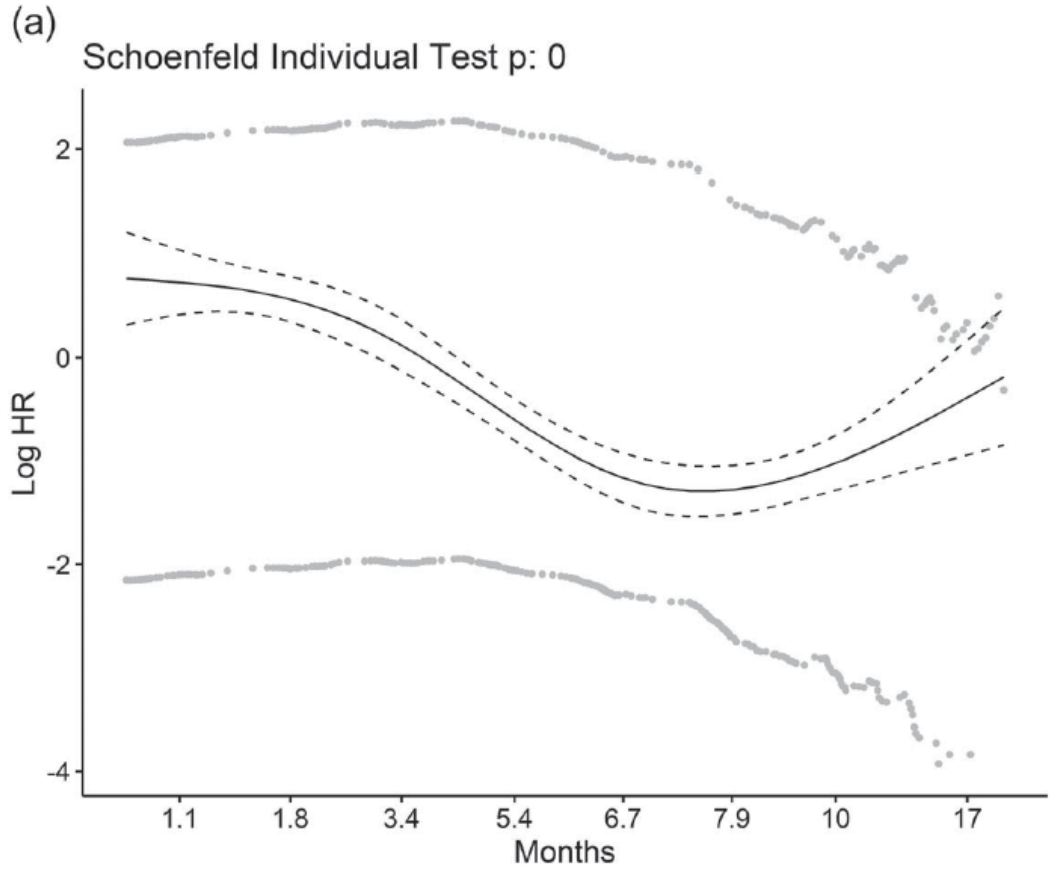
E.g. the following KM plot from the IPASS study (Mok et al. (2009)).



HR over time in the IPASS study

It is not easy from survival curves to see what is happening to the HR.

The following PFS Schoenfeld residual plot from Lin et al. (2020) shows the estimated log HR over time.



HR over time in the IPASS study

One approach is to split time up into intervals and estimate separate, period specific HRs in each interval.

E.g. in IPASS the estimated HR in the first 6 months is 1.115 and from 6 months onwards is 0.343 (Lin et al. 2020).

Gefitinib seems to do worse/same as control initially, and then better.

Interpretation of hazards and HRs

The apparent interpretation when we see the HR varying over time is that the treatment effect is varying over time.

Indeed there is of course no reason why treatments' effects should be constant over time.

But could the picture be (even) more complicated?

Unfortunately, yes. . .

Interpreting changes in hazard and HRs

Interpreting changes in hazard and HRs

In recent years, the hazard ratio has come under some criticism, from the perspective of causal inference.

This began with the 'The Hazard of Hazard Ratios' paper (Hernán (2010)).

The source of the difficulty is due to the fact the hazard is a conditional (on survival) quantity and because of the ubiquitous presence of unmeasured/unknown individual level (frailty) factors which influence outcome.

We will look in detail at the very simplest type of frailty model at what is happening.

A two-arm clinical trial with frailty

Imagine we have a very large two-arm trial.

A yet to be discovered genetic risk factor exists which we call x.

50% of the population have x=0, and 50% have x=1.

Because treatment is randomly assigned, each treatment group will have (in expectation) 50% x=0 individuals and 50% x=1 individuals.

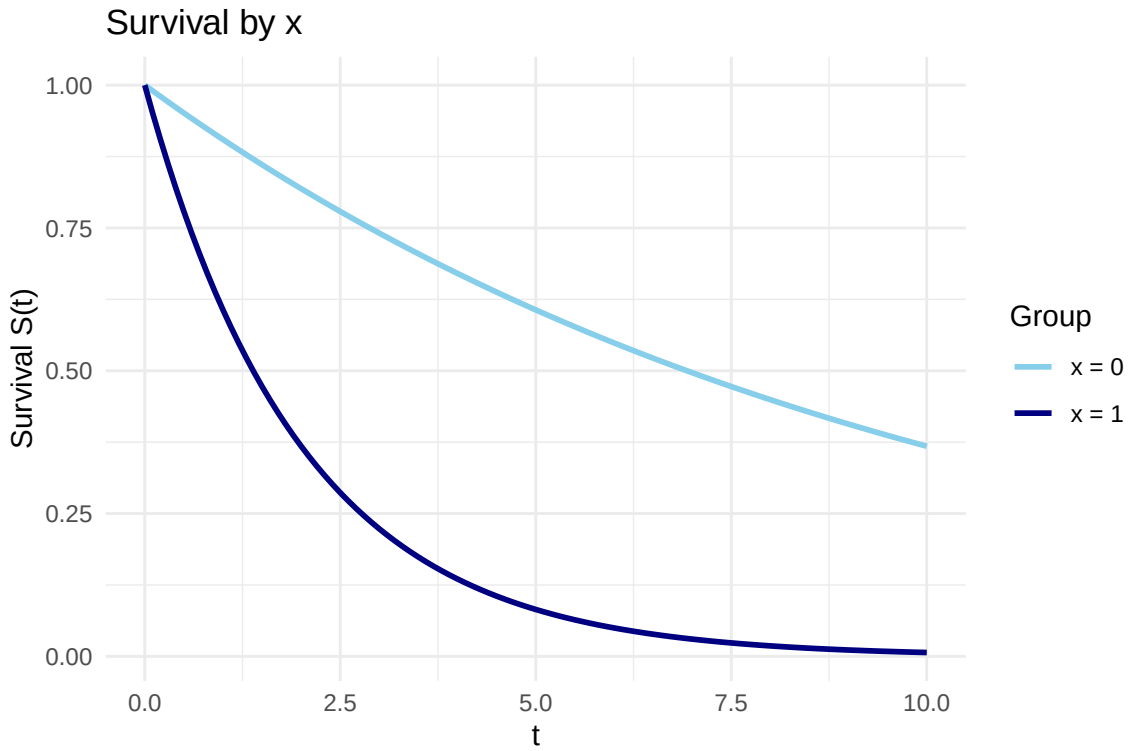
We will assume hazard of the event is constant over time given x and treatment.

In the control arm, we assume the following hazards:

Treatment	x	Hazard
control	0	0.1
control	1	0.5

Survival curves in the control arm, by frailty factor

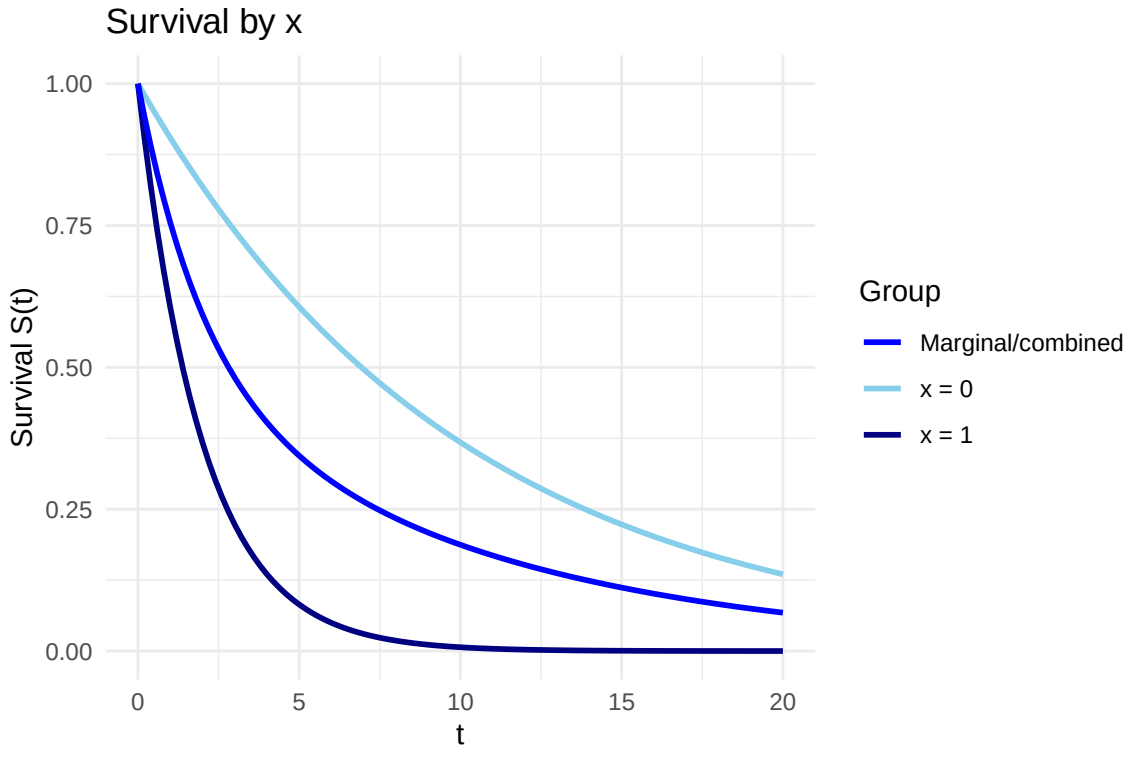
Under constant hazards, the event times are *exponentially* distributed, with the following survival functions.



At later times, only those with $x = 0$ are still alive.

Survival curves in the control arm

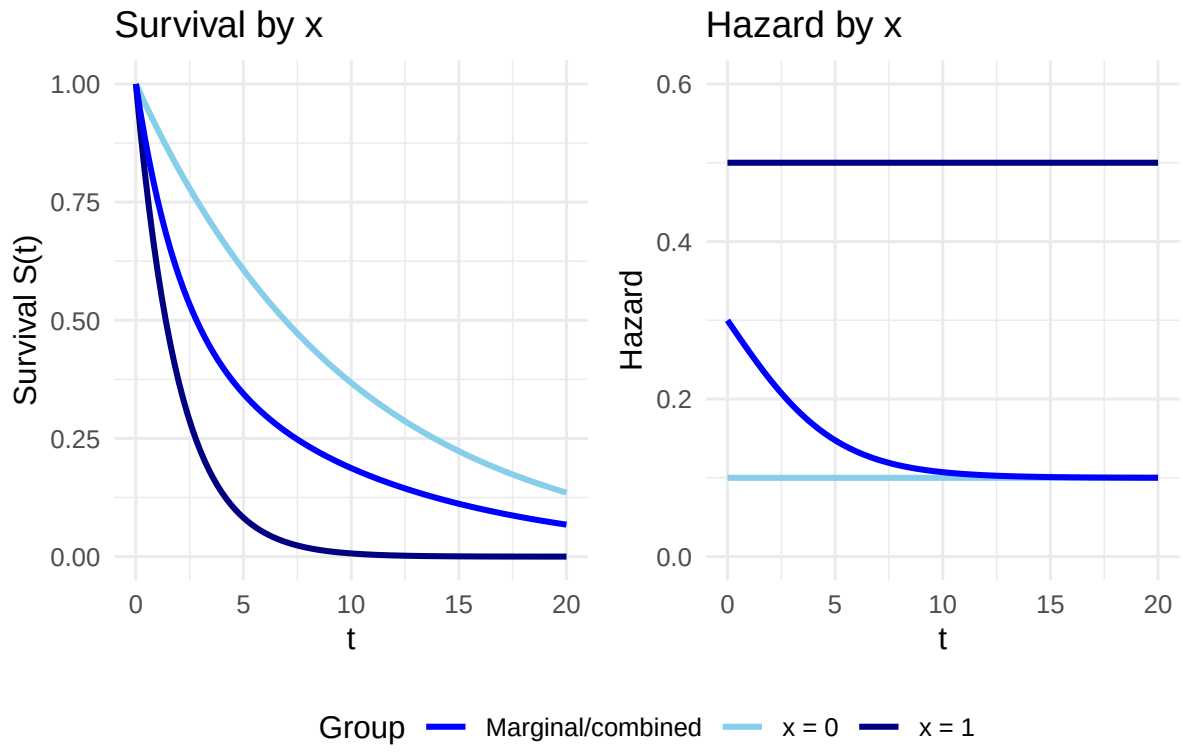
We now add the overall control arm survival to the plot.



Marginal/overall survival is average (due to 50:50 x dist.) of two frailty-specific survival curves.

Given x has not yet been discovered, this is the survival curve we can actually estimate.

Hazards in the control arm



Initially the marginal/overall hazard is the average of the two group specific hazards.

But as time goes on, the marginal hazard converges to that of the x=0 group, since the hazard is the failure rate *among the survivors*.

Changes in hazard

If the frailty variable is unmeasured, all we can estimate is this marginal/overall hazard.

The marginal hazard decreases over time.

But the individual-level hazards are not changing over time - the change occurs because the risk set becomes dominated by x=0 individuals.

Adding the treatment arm

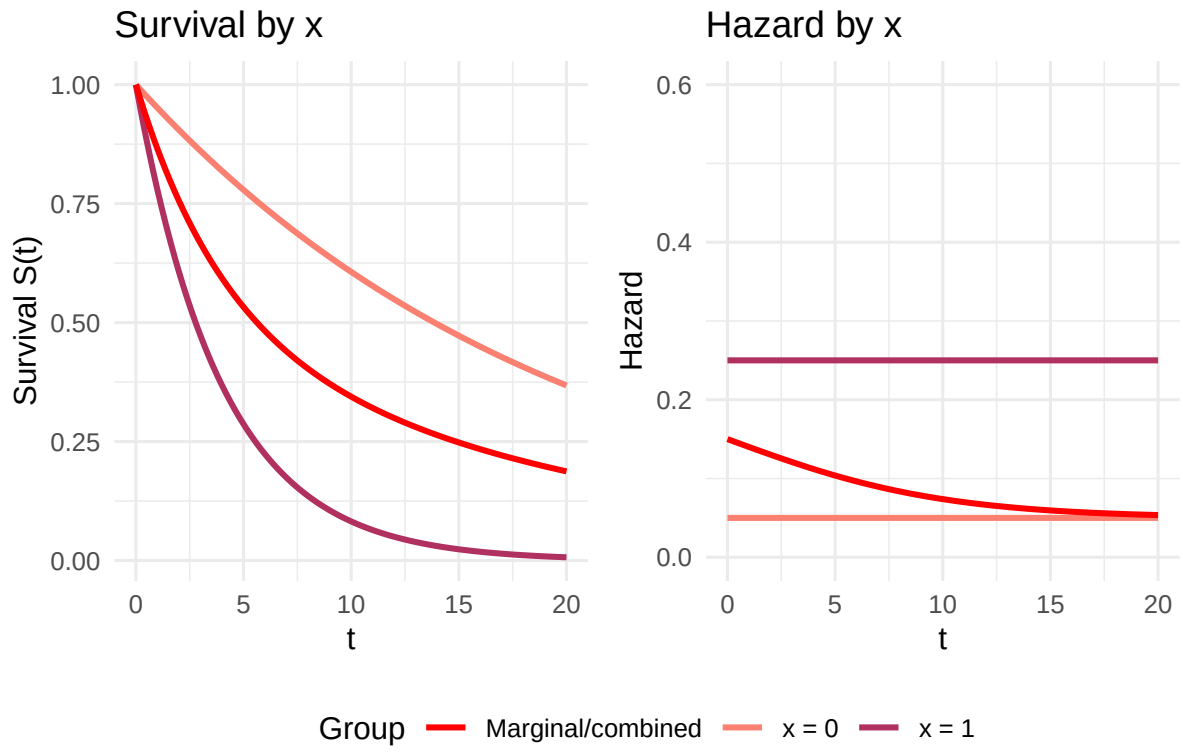
Now we add in a (active) treatment arm.

Treatment	x	Hazard
control	0	0.10
control	1	0.50
active	0	0.05
active	1	0.25

The HR comparing to active to control, conditional on x=0 is 0.5, and for x=1 is 0.5.

These (conditional on x) HRs comparing treatment groups do not change over time.

Survival and hazards in the active arm



Similar to the control arm, marginal/overall hazard converges to the x=0 hazard.

But *crucially*, this convergence happens slower because of the lower hazard in the active arm.

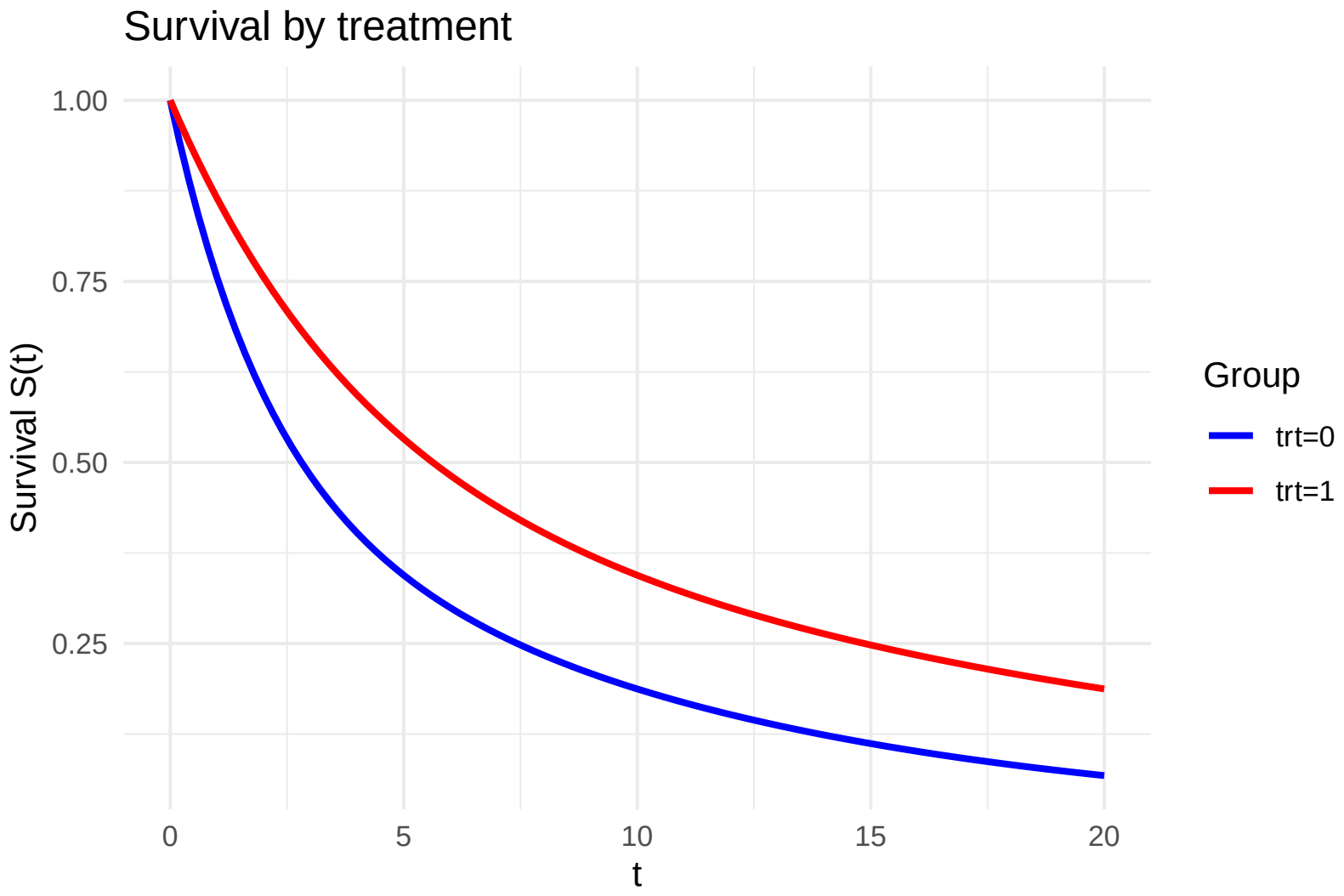
What we get to see

Recall the frailty factor x is unmeasured.

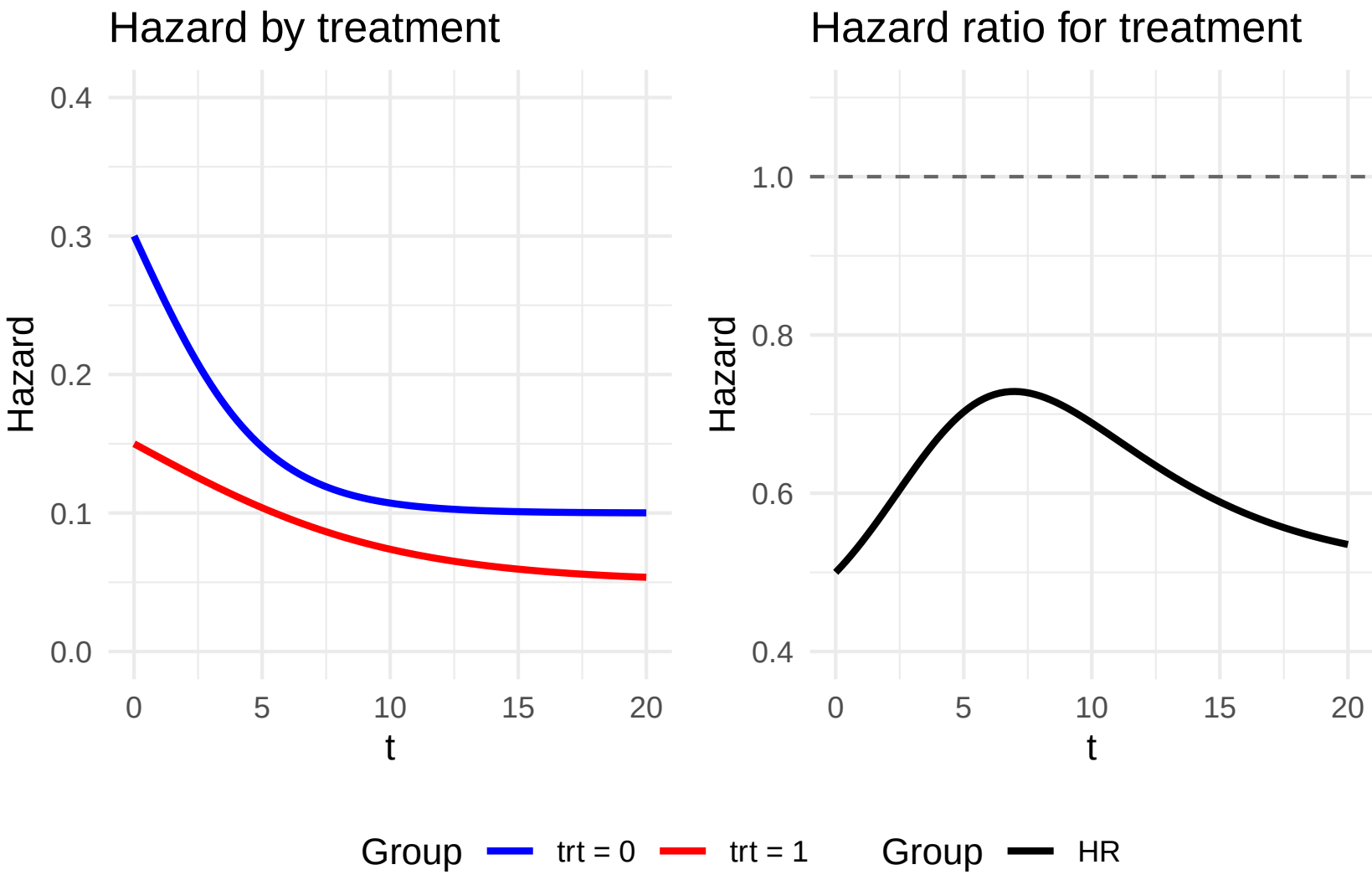
All we get to see are marginal survival or hazard in each arm.

So let's now compare the marginal survival hazards between arms...

Survival marginal to frailty by arm



Hazards marginal to frailty by arm



What's going on?

Initially the (marginal to x) HR is close to 0.5.

As time moves on, there are relatively more high risk (x=1) individuals still at risk in the active arm compared to the control arm.

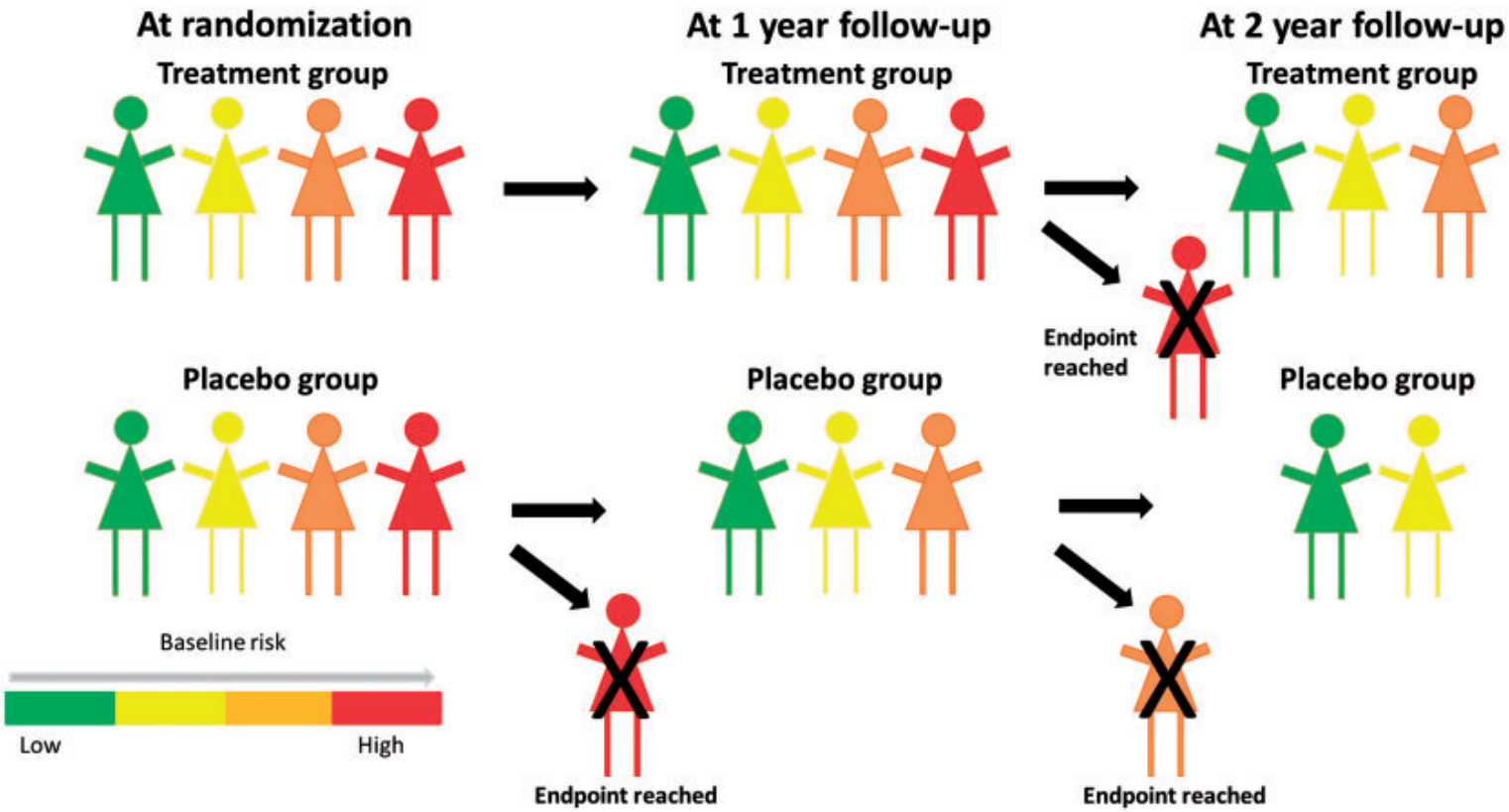
This starts to pull the HR up towards one.

Eventually though, most of the high risk (x=1) individuals in the active arm fail as well.

Then the two risk sets (almost) only consist of x=0 individuals, and the HR converges to the x=0 specific HR.

The explanation in pictures

Stensrud et al. (2019) explains the issue very nicely in pictures:



The explanation in a DAG

O. O. Aalen, Cook, and Røysland (2015) shows how the difference in risk sets / lack of comparability can be seen using DAGs.

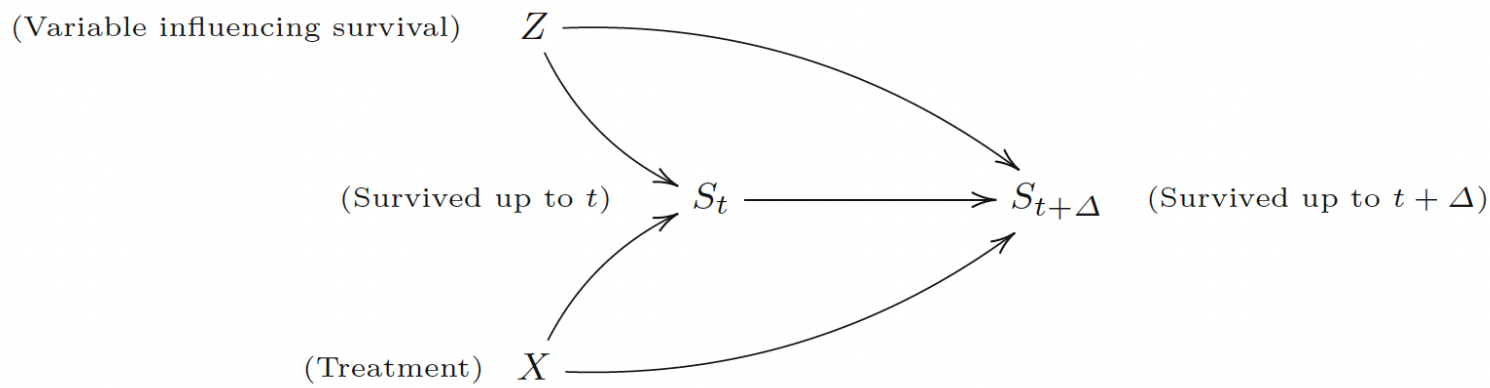


Fig. 2 A directed acyclic graph of X, Z and S_t and $S_{t+\Delta}$

Conditioning on survival to time t amounts to 'conditioning on a collider'.

Frailty variable (here labelled Z) and treatment (here labelled X) become associated, conditional on survival to some time t.

Other scenarios

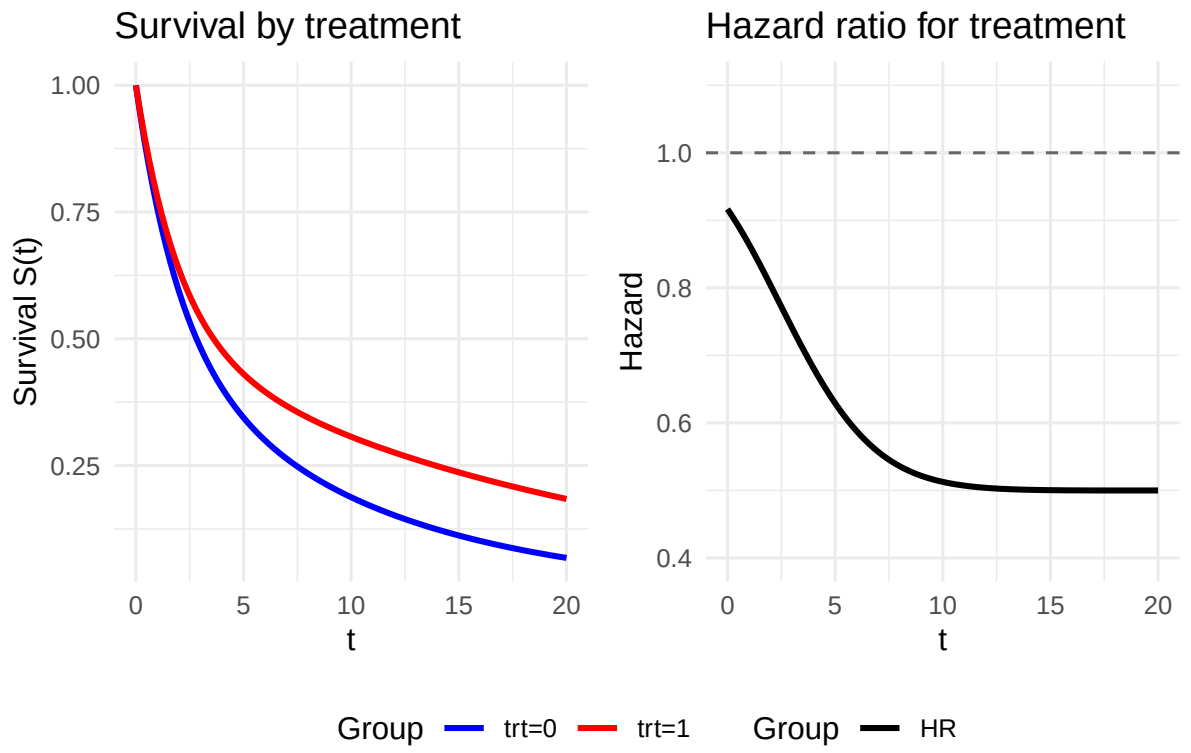
We have shown the survival and hazard functions for one particular setup.

More variations are possible, even within the simple model we have used where the frailty variable is binary and hazards are constant conditional on the frailty variable and treatment.

To illustrate, we show two further scenarios in which the hazard for $x = 1$ individuals under active is varied.

Scenario 2

Treatment	x	Hazard
control	0	0.10
control	1	0.50
active	0	0.05
active	1	0.50



Scenario 2 comments

Marginally, at first active treatment has small benefit, and this benefit increases over time.

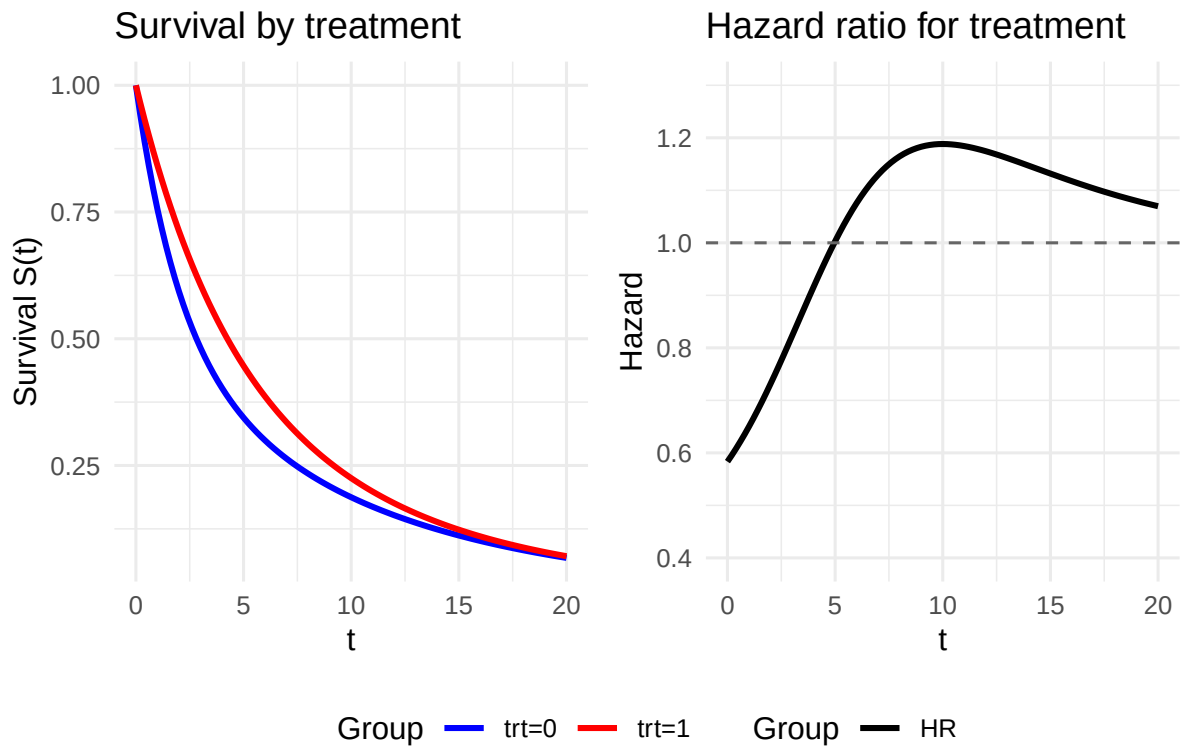
This might be described as a delayed benefit of active treatment.

Yet conditional on frailty variable x, the HR for $x = 0$ individuals is 0.5 at all times, and for $x = 1$ individuals is 1 at all times.

That is, for some individuals active treatment has no effect on hazard, while for others it reduces it.

Scenario 3

Treatment	x	Hazard
control	0	0.10
control	1	0.50
active	0	0.10
active	1	0.25



Scenario 3 comments

Marginally, at first active treatment has benefit, but later the HR switches to being > 1, apparently suggesting active treatment is harmful.

Yet conditional on frailty variable x, the HR for $x = 0$ individuals is 1 at all times, and for $x = 1$ individuals is 0.5 at all times.

That is, for some individuals active treatment has no effect on hazard, while for others it reduces it (at all times).

Interim conclusions

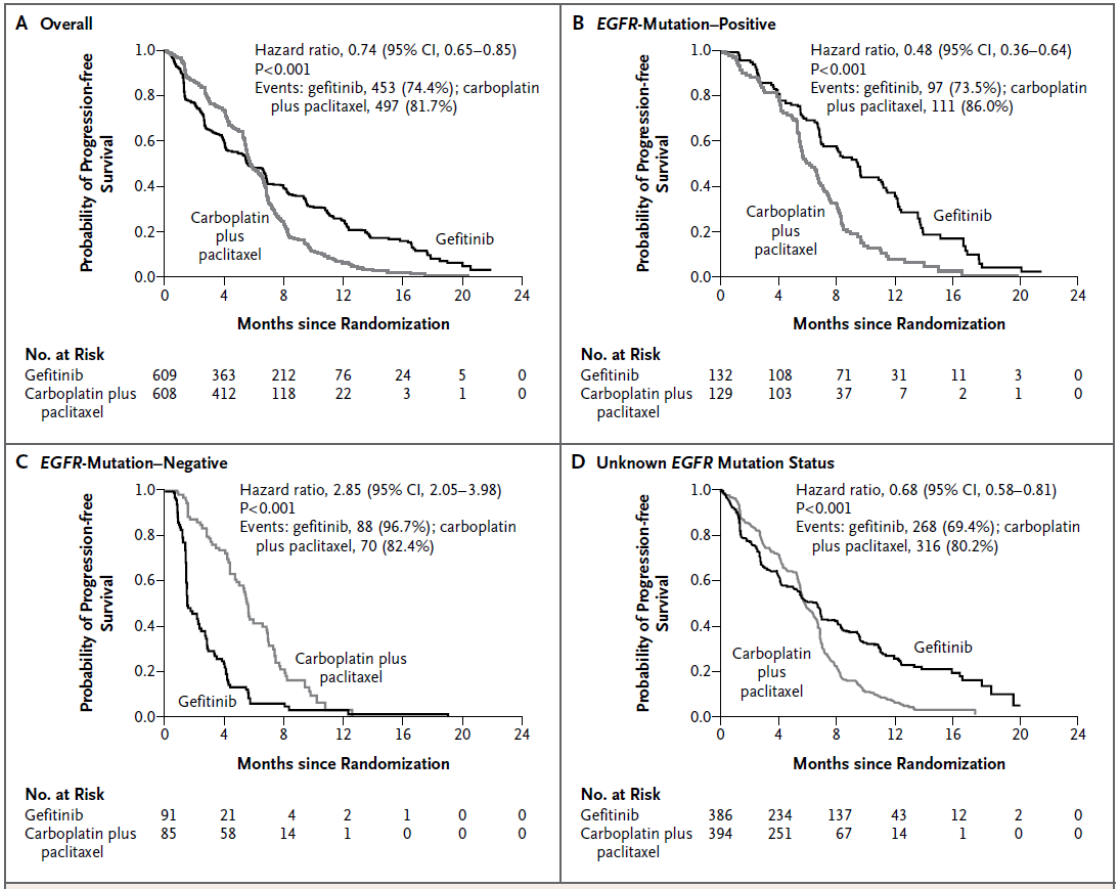
- Frailty can cause observable hazards and HRs to do some interesting things.
- The three scenarios assumed HRs that were constant over time conditional on x.
- A myriad of alternative scenarios can be constructed, including where HR conditional on frailty varies over time.
- In fact, the marginal HR can be constant even though conditional HR is time-varying (Stensrud et al. (2019)).
- This means that even if marginally hazards are proportional between treatment groups, conditional/individual HRs could be time-varying!

Rare outcomes

- The issues arise because the risk sets are (in general) no longer similar between the treatment arms.
- An important exception are settings where the event of interest is rare.
- In this case, the risk sets at later times are not so different to at baseline, and thus similarity of risk sets is (roughly speaking) maintained.

Back to IPASS

Reproduced from Mok et al. (2009).



Here a factor (EGFR mutation) was measured in some patients, and could explain some of the heterogeneity in outcomes.

Causal criticisms of HR

Concerns about the effects of frailty have led some researchers, particularly those from causal inference, to question the use of HRs (especially time-dependent / period specific ones) as valid measures of treatment effect. (Hernán 2010; O. O. Aalen, Cook, and Røysland 2015; Martinussen, Vansteelandt, and Andersen 2020; Stensrud et al. 2019)

For example, Hernán (2010) stated 'the HR has a built-in selection bias'.

Rather than HRs, alternatives can be used that do not suffer in the same way:

- present the survival curves
- restricted mean survival
- risk ratios for survival/event at chosen 'landmark' time

But...

I think it nevertheless remains important to understand how hazards and HRs should be interpreted.

The hazard is a critical component of the machinery of survival analysis.

And the Cox model and HRs will likely not soon be usurped in my view, for various reasons.

How to interpret hazards and hazard ratios

How NOT to interpret hazards and hazard ratios

HRs should *not* be interpreted as (common) individual level effects of treatment on one's hazard.

Such a conclusion would only be justified if one could plausible argue that no frailty factors exist.

Even if we adjust for some important covariates, there is no reason to think there do not exist further factors which explain heterogeneity between individuals.

Individual level effects

Such a conclusion is not so surprising, not least given that for continuous or binary outcomes our effect measures are not valid individual level effects either, unless we make strong assumptions.

E.g. mean blood pressure is 10mmHg lower on active compared to control, but we wouldn't suggest active treatment lowers every individual's blood pressure by the same (10mmHg) amount.

E.g. active treatment group has 5% risk of death, and control group has 10% risk. Does that mean we believe active treatment reduces *every* patient's risk by half?

How to interpret hazards

How then should we interpret hazards and HRs?

Go back to the definition.

For the hazard in one group, for small δ, we expect $N(t) \times \delta h(t)$ individuals to fail in the next period of time δ.

That is, of those still at risk, the proportion who will have the event in the next δ units of time is $\delta h(t)$.

How to interpret hazard ratios

Now consider an RCT, with marginal hazards $h_1(t)$ and $h_0(t)$ in the two arms.

What is $HR(t) = \frac{h_1(t)}{h_0(t)}$?

If we assign the population active treatment, of those still at risk at time t, a proportion $\delta h_1(t)$ will fail in the next period δ of time.

If we assign the population control treatment, of those still at risk at time t, a proportion $\delta h_0(t)$ will fail in the next period δ of time.

So $HR(t) = \frac{\delta h_1(t)}{\delta h_0(t)} = \frac{h_1(t)}{h_0(t)}$ is the ratio of the proportions of those still at risk we expect to fail in the next period δ of time.

Those still at risk at time t in these two scenarios are *not* in general the same individuals.

Is $HR(t)$ 'causal'?

There has been some debate as to whether HRs are 'causal'.

No clear consensus on what it means for a measure to be causal.

But it is causal in the sense that $HR(t) = \frac{h_1(t)}{h_0(t)}$ measures how the proportion failing in the next small δ amount of time changes if we assign active treatment compared to if we assign control.

It tells us how something would change if we intervene (in this case on treatment for the population).

Conclusions

Conclusions 1

- Interpreting changes in hazards and HRs over time is tricky.
- In general they could be due to a combination of selection effects and time-varying treatment effects, and we cannot disentangle these two.
- If hazards are proportional marginally, I think the (constant) HR remains a useful effect measure, with a valid population level interpretation.

Conclusions 2

- If (marginal) HR varies over time, time-varying/specific HRs may be useful descriptively, but they are easily misinterpreted if one forgets about frailty (Bartlett et al. 2020).
- The survival curves give, in some sense, the fullest picture.
- If a summary measure is required (e.g. RMST, ratio of proportion surviving to time τ), discussion needed as to which is most relevant for the intended purpose.

Other points

- I have talked about trials, but the same issues & conclusions hold for observational studies where we have adjusted for confounders.
- The issues discussed are less of a concern when the outcome is rare.
- Stensrud et al. (2019) is a very nice overview of the topic.
- Chapter 6 of O. Aalen, Borgan, and Gjessing (2008) is also recommended.

References I

Aalen, O O, R J Cook, and K Røysland. 2015. "Does Cox analysis of a randomized survival study yield a causal treatment effect?" *Lifetime Data Analysis* 21 (4): 579–93.

Aalen, Odd, Ornulf Borgan, and Hakon Gjessing. 2008. *Survival and Event History Analysis: A Process Point of View*. Springer Science & Business Media.

Bartlett, Jonathan W, Tim P Morris, Mats J Stensrud, Rhian M Daniel, Stijn K Vansteelandt, and Carl-Fredrik Burman. 2020. "The Hazards of Period Specific and Weighted Hazard Ratios." *Statistics in Biopharmaceutical Research* 12 (4): 518.

Hernán, Miguel A. 2010. "The Hazards of Hazard Ratios." *Epidemiology (Cambridge, Mass.)* 21 (1): 13.

References II

Lin, Ray S, Ji Lin, Satrajit Roychoudhury, Keaven M Anderson, Tianle Hu, Bo Huang, Larry F Leon, et al. 2020. "Alternative Analysis Methods for Time to Event Endpoints Under Nonproportional Hazards: A Comparative Analysis." *Statistics in Biopharmaceutical Research* 12 (2): 187–98.

Martinussen, Torben, Stijn Vansteelandt, and Per Kragh Andersen. 2020. "Subtleties in the Interpretation of Hazard Contrasts." *Lifetime Data Analysis* 26 (4): 833–55.

Mok, Tony S, Yi-Long Wu, Sumitra Thongprasert, Chih-Hsin Yang, Da-Tong Chu, Nagahiro Saijo, Patrapim Sunpaweravong, et al. 2009. "Gefitinib or Carboplatin–Paclitaxel in Pulmonary Adenocarcinoma." *New England Journal of Medicine* 361 (10): 947–57.

Stensrud, Mats J, John M Aalen, Odd O Aalen, and Morten Valberg. 2019. "Limitations of Hazard Ratios in Clinical Trials." *European Heart Journal* 40 (17): 1378–83.

References III