# Are hazard ratios estimated from trials valid causal effect measures?

Jonathan Bartlett
www.thestatsgeek.com
www.missingdata.org.uk

Department of Mathematical Sciences
University of Bath, UK

7th May 2019

# Acknowledgement

Thanks to Rhian Daniel and Stijn Vansteelandt for helpful discussions.

## Motivation

## Causal inference using potential outcomes

## The hazards of hazard ratios

## Aalen et al 2015

## HR - a valid population level causal effect

## Conclusions

# Outline

# Motivation - Aalen et al 2015 [6]

Lifetime Data Anal (2015) 21:579–593
DOI 10.1007/s10985-015-9335-y

CrossMark

## Does Cox analysis of a randomized survival study yield a causal treatment effect?

Odd O. Aalen[1] · Richard J. Cook[2] · Kjetil Røysland[1]

# Aalen et al 2015

*"Despite the fact that treatment assignment is randomized, the hazard ratio is not a quantity which admits a causal interpretation in the case of unmodelled heterogeneity."*

*"This makes it unclear what the hazard ratio computed for a randomized survival study really means. Note, that this has nothing to do with the fit of the Cox model. The model may fit perfectly in the marginal case with $X$ as the only covariate, but the present problem remains."*

# Outline

# Causal inference using potential outcomes

We will first review causal inference using the potential or counterfactual outcomes framework.

We will draw heavily on Hernán and Robins' excellent (free so far) book, Causal Inference [1].

Consider some well defined population of invididuals.

Each individual will receive one of two interventions or treatments, coded 0 and 1.

We will measure an outcome $Y$, for illustration taken to be binary for the moment.

# Potential outcomes

| | $Y^0$ | $Y^1$ |
|---|---|---|
| Rheia | 0 | 1 |
| Kronos | 1 | 0 |
| Hades | 1 | 1 |
| Zeus | 1 | 0 |
| Athena | 0 | 0 |
| Aphrodite | 1 | 0 |
| Hermes | 0 | 1 |

- $Y_i^0$ and $Y_i^1$ are the **potential outcomes** for individual $i$ under treatment level 0 and 1.
- Somehow it is determined which treatment each individual will receive.
- We let $Z_i = 0$ or $Z_i = 1$ denote the treatment individual $i$ receives.
- The outcome we observe is $Y_i^0$ if $Z_i = 0$ and $Y_i^1$ if $Z_i = 1$ (consistency).

# Individual level causal effects

There is a causal effect for individual $i$ if $Y_i^0 \neq Y_i^1$.

An individual level causal effect can be quantified as some contrast of $Y_i^0$ with $Y_i^1$.

e.g. $Y_i^1 - Y_i^0$.

# Observable data

| | $Z$ | $Y^0$ | $Y^1$ |
|---|---|---|---|
| Rheia | 0 | 0 | . |
| Kronos | 1 | . | 0 |
| Hades | 1 | . | 1 |
| Zeus | 0 | 1 | . |
| Athena | 1 | . | 0 |
| Aphrodite | 1 | . | 0 |
| Hermes | 0 | 0 | . |

- For each individual we get to see only one of their potential outcomes.
- Which one depends on the value of $Z$.
- Hence the problem of causal inference can be viewed as a missing data problem.

# Population level causal effects

Without strong untestable assumptions we cannot identify/estimate individual level causal effects.

Under weaker assumptions we can estimate **population level** casual effects.

These are contrasts of a functional of the population distributions of $Y^0$ and $Y^1$.

e.g. $E(Y_i^1) - E(Y_i^0)$

They are sometimes referred to as **average causal effects**, because $E(Y_i^1) - E(Y_i^0) = E(Y_i^1 - Y_i^0)$.

I prefer the term **population level** effects, as other measures are not always averages of individual level effects.

# What can we estimate in practice?

Suppose that treatment $Z_i$ is randomly assigned, so $Z_i$ is independent of $(Y_i^0, Y_i^1)$ across the population.

In missing data parlance, the data are **missing completely at random**.

Then $f(Y_i^1) = f(Y_i|Z_i = 1)$ and $f(Y_i^0) = f(Y_i|Z_i = 0)$.

In particular $E(Y_i^0) = E(Y_i|Z_i = 0)$ and $E(Y_i^1) = E(Y_i|Z_i = 1)$.

Hence we can estimate for example $E(Y_i^1) - E(Y_i^0)$ by $\hat{E}(Y_i|Z_i = 1) - \hat{E}(Y_i|Z_i = 0)$.

# Effect measures for binary outcomes

For binary outcome measures, the most common effect measure estimators used are:

- risk difference (RD): $\widehat{\text{RD}} = \bar{Y}^1 - \bar{Y}^0$
- risk ratio (RR): $\widehat{\text{RR}} = \bar{Y}^1 / \bar{Y}^0$
- odds ratio (OR): $\widehat{\text{OR}} = \frac{\bar{Y}^1}{1-\bar{Y}^1} / \frac{\bar{Y}^0}{1-\bar{Y}^0}$

where $\bar{Y}^0$ and $\bar{Y}^1$ denote the sample proportions of 1s in the two treatment groups.

# Stochastic vs. deterministic potential outcomes

Potential outcomes (POs) can be assumed to be deterministic or stochastic [1].

Deterministic POs: for individual $i$, $Y_i^0$ and $Y_i^1$ are fixed.

Stochastic POs: for individual $i$, $Y_i^0$ and $Y_i^1$ are draws from some probability distribution.

Quantum physics implies (apparently!) POs can't be truly deterministic.

Causal inference literature tends towards deterministic, often implicitly.

# Stochastic vs. deterministic binary potential outcomes

In the case of a binary outcome:

Stochastic POs: $Y_i^0 \sim Bernoulli(\pi_i^0)$, $Y_i^1 \sim Bernoulli(\pi_i^1)$.

'Purely' stochastic POs: $\pi_i^0 = \pi^0$, $\pi_i^1 = \pi^1$ for all $i$. This is implausible due to observed variation in risk between individuals.

'Partly' stochastic POs: $\pi_i^0 = g^0(\tilde{X}_i)$, $\pi_i^1 = g^1(\tilde{X}_i)$ for existent baseline variables $\tilde{X}_i$.

Deterministic POs: $\pi_i^0 = h^0(\tilde{X}_i) \in \{0, 1\}$, $\pi_i^1 = h^1(\tilde{X}_i) \in \{0, 1\}$.

# Causal effect measures - purely stochastic POs

Purely stochastic POs: $Y_i^0 \sim Bernoulli(\pi^0)$, $Y_i^1 \sim Bernoulli(\pi^1)$.

$Y_i^1 - Y_i^0$, is itself stochastic, so there is no fixed value to estimate.

$\bar{Y}^0$ and $\bar{Y}^1$ estimate $\pi^0$ and $\pi^1$, and $\widehat{\text{RD}}$ estimates $\pi^1 - \pi^0$.

$\pi^1 - \pi^0$ is the common individual level causal RD.

RR and OR can be interpreted as common individual level causal effects.

But we have said purely stochastic POs, with no variation in risk, are implausible!

# Causal effect measures - partly stochastic POs

'Partly' stochastic POs: $\pi_i^0 = g^0(\tilde{X}_i)$, $\pi_i^1 = g^1(\tilde{X}_i)$ for baseline (measured and unmeasured) variables $\tilde{X}_i$.

$Y_i^1 - Y_i^0$, is again stochastic.

$\pi_i^1 - \pi_i^0$ now varies across individuals, in general.

$\widehat{\text{RD}}$ estimates $E_i(\pi_i^1) - E_i(\pi_i^0) = E(\pi_i^1 - \pi_i^0)$.

This can be interpreted as a population level causal RD, or average individual level effect.

$\widehat{\text{RD}}$ can only be interpreted as **the** individual level RD if $\pi_i^1 - \pi_i^0$ does not vary across $i$.

# Causal effect measures - partly stochastic POs

$\widehat{\text{RR}}$ estimates $\frac{E_i(\pi_i^1)}{E_i(\pi_i^0)}$.

This is the population level RR.

Hernán and Robins [1] note that

$$\frac{E_i(\pi_i^1)}{E_i(\pi_i^0)} = E_i\left[\frac{\pi_i^0}{E_j(\pi_j^0)}\frac{\pi_i^1}{\pi_i^0}\right]$$

so that it can be viewed as a weighted average of the individual level risk ratios. Not clear though that this is useful though.

Again only if $\pi_i^1/\pi_i^0$ were identical across $i$ could $\widehat{\text{RR}}$ be interpreted as a common individual level RR. And if individual level RD is common, RR cannot be, and vice-versa.

# Causal effect measures - partly stochastic POs

$\widehat{\text{OR}}$ estimates population level OR: $\frac{E_i(\pi_i^1)}{1-E_i(\pi_i^1)} \Big/ \frac{E_i(\pi_i^0)}{1-E_i(\pi_i^0)}$

Due to non-collapsibility of the OR (Fine Point 4.3 of [1]), the population level OR does not equal the individual level OR even when the latter is identical across individuals.

If the individual level OR were identical across individuals, we need to condition adjust (correctly) on all prognostic factors $\tilde{X}_i$ to estimate it.

No reason to think that we will ever have all the prognostic variables measured, and if we did, that we correctly model their effects.

# Causal effect measures - deterministic POs

$Y_i^1 - Y_i^0$ is 0, 1, or -1, and is now fixed for each individual, although they are not identifiable.

$\widehat{\text{RD}}$ estimates population RD $E_i(Y_i^1) - E_i(Y_i^0)$, and similarly for $\widehat{\text{RR}}$ and $\widehat{\text{OR}}$.

There is no longer an individual level RD, RR, or OR, since there is no notion of randomness in the outcomes of an individual.

Probability models are still used for inference - randomness is due to random sampling of individuals from the population into the sample.

# Population vs. individual effects

Some advocate adjusting for covariates and interpreting the resulting estimate as an individual level effect (e.g. Harrell [4]).

But this relies on assuming:

- the chosen effect measure is common across individuals
- we correctly model covariate effects

Because of these issues, others choose instead to target population (marginal) effects (e.g. Steingrimsson *et al* [3]). An issue here though is whether samples are representatitive of target population.

Of course one could estimate sub-population effects, allowing for the possibility/fact(?!) that these will vary across sub-populations.

# Conclusions to take forward

Even in a randomised trial, to estimate individual level causal effects requires implausible and untestable assumptions.

Instead, we can more plausibly / reliably estimate population level effects.

With detemrinistic POs, these are some contrast of the population distributions $f(Y_i^0)$ and $f(Y_i^1)$.

With stochastic POs, these are contrasts of population distributions of parameters indexing stochastic PO distributions, e.g. $f(\pi_i^0)$ and $f(\pi_i^1)$.

In either case, they tell us about how the population distribution of outcomes would change if we gave treatment 1 to the population rather than treatment 0.

# Outline

# Time to event outcomes

Now suppose the outcome $T$ is a time to event outcome.

$T_i^0$ and $T_i^1$ are potential times to event under control and active treatments, for individual $i$.

# The hazard function and Cox model

The hazard function is

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

It is the instantaneous event rate at time $t$ among those individuals who have not yet 'failed'.

Cox's model with treatment group as covariate assumes that

$$\frac{\lambda(t|Z=1)}{\lambda(t|Z=0)} = \exp(\beta)$$

i.e. the hazard ratio is constant over time.

# The hazard ratio

The hazard ratio at time $t$ comparing active to control treatments is

$$HR(t) = \frac{\lim_{\Delta t \to 0} P(t \leq T \leq t + \Delta t | T \geq t, Z = 1)/\Delta t}{\lim_{\Delta t \to 0} P(t \leq T \leq t + \Delta t | T \geq t, Z = 0)/\Delta t}$$

This can be rewritten [5] in terms of POs as

$$HR(t) = \frac{\lim_{\Delta t \to 0} P(t \leq T^1 \leq t + \Delta t | T^1 \geq t)/\Delta t}{\lim_{\Delta t \to 0} P(t \leq T^0 \leq t + \Delta t | T^0 \geq t)/\Delta t}$$

# Hazard ratio

$$HR(t) = \frac{\lim_{\Delta t \to 0} P(t \leq T^1 \leq t + \Delta t | T^1 \geq t)/\Delta t}{\lim_{\Delta t \to 0} P(t \leq T^0 \leq t + \Delta t | T^0 \geq t)/\Delta t}$$

So, $HR(t)$ is comparing short term risk in those who would survive to $t$ under active ($T^1 \geq t$) to those who would survive to $t$ under control ($T^0 \geq t$).

Randomisation ensures baseline covariates $X$ are balanced (in distribution) between treatment groups.

But not necessarily that there is balance between the two groups of survivors at times $t > 0$.

# Herná 2010 - the hazards of hazard ratios

- Because of this issue, in 2010 Hernán argued that the HR has a built in 'selection bias'.
- e.g. the Women's Health Initiative randomised 16,000 women to hormone therapy or placebo, and followed them up for coronary heart disease CHD.
- So, is hormone therapy protective after 5 years?

| Time (years) | HR |
|---|---|
| 0-1 | 1.81 |
| 1-2 | 1.34 |
| 2-3 | 1.27 |
| 3-4 | 1.25 |
| 4-5 | 1.45 |
| 5- | 0.7 |

# The hazards of hazard ratios

As argued by Hernán [2], it is entirely possible that hormone therapy could increase risk for CHD for some women.

These women will tend to experience CHD earlier in the hormone therapy group.

At later times, the women still event free (at risk) in the two groups then differ in terms of their risk for CHD.

The HR at years 5+ of 0.7 could therefore purely reflect selection effects, rather than it meaning the individual level effect of treatment switches direction over time.

# Outline

# Aalen et al 's critique - part 1

Aalen *et al* gives a number of perspectives for why the HR is not a valid causal effect, even if $HR(t)$ is constant over time.

The first describes Hernán's point in more detail - at times $t > 0$ the $HR(t)$ is not making a fair comparison, due to selection effects.

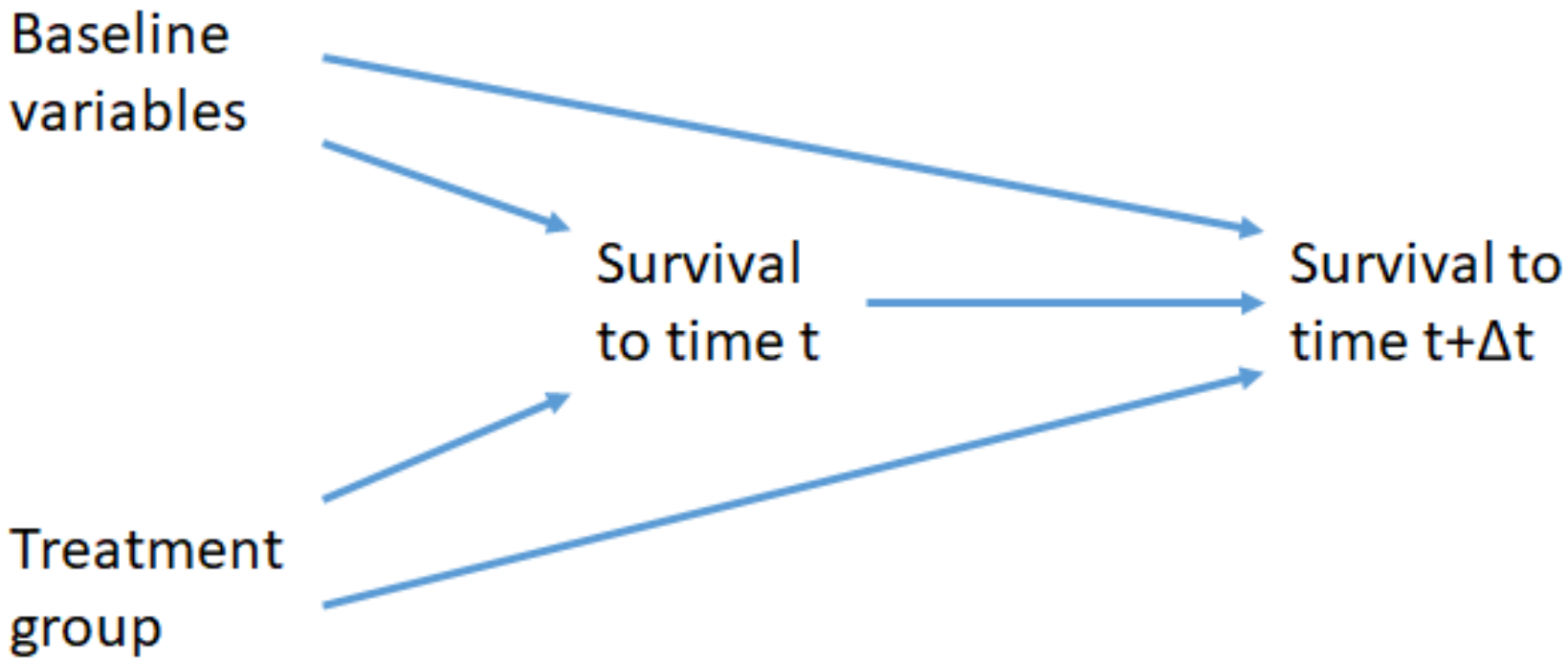
Aalen *et al* show that survivors in two treatment groups are balanced w.r.t. baseline variables, i.e. $\tilde{X} \perp\!\!\!\perp Z | T > t$ only if

$$\lambda(t|\tilde{X}, Z) = a(t, \tilde{X}) + b(t, Z)$$

for functions $a(.,.)$ and $b(.,.)$.

# Aalen et al 's critique - part 2

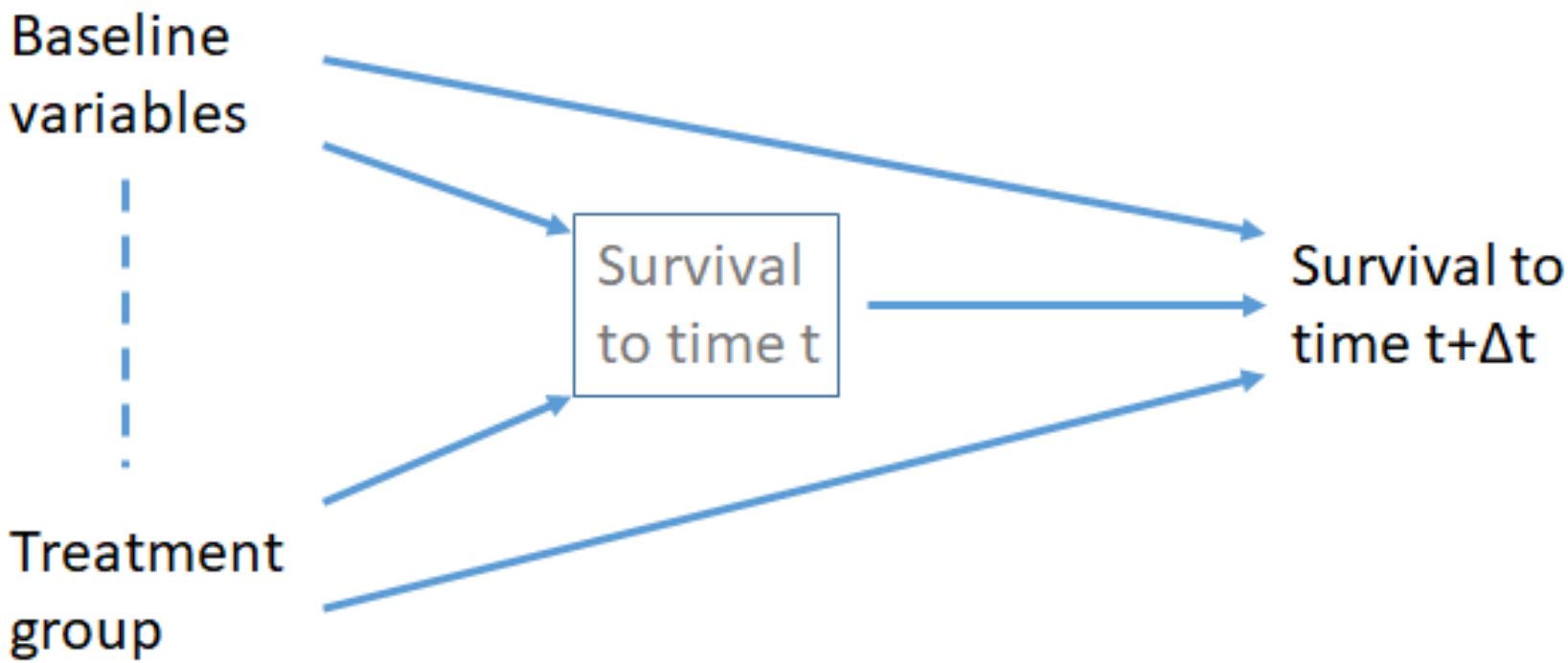
This issue can also be viewed via **direct acyclic graphs** (DAGs).



# Aalen et al 's critique - part 2

The hazard ratio conditions on survival to $t$.

We are conditioning on a **collider**, and open up a path between baseline variables and treatment group.



# Aalen et al 's critique - part 3

Aalen *et al* then consider an analysis based on the notion of an **individual level hazard** function.

This corresponds to our notion of stochastic POs, with individual $i$ having hazards $\lambda_i^0(t)$ and $\lambda_i^1(t)$ under control and active treatments.

Suppose:

$$\begin{aligned} \lambda_i^0(t) &= \lambda_0(t) \\ \lambda_i^1(t) &= \lambda_0(t)\exp(\beta) \end{aligned}$$

Then the HR $\exp(\beta)$ represents the (common) individual level effect of treatment.

It could be estimated by fitting a Cox model to the observed data.

# Aalen et al 's critique - part 3 continued

But the preceding model is completely implausible - individual hazard will depend on individual baseline variables $\tilde{X}$.

Suppose:

$$\begin{aligned} \lambda_i^0(t) &= g(\tilde{X}_i, t) \\ \lambda_i^1(t) &= g(\tilde{X}_i, t)\exp(\beta) \end{aligned}$$

for baseline variables $\tilde{X}$ and function $g(.,.)$.

Then $\exp(\beta)$ represents the (common) individual level effect of treatment

# Aalen et al 's critique - part 3 continued

In practice you can never hope to measure all the components of $\tilde{X}$.

Like logistic regression, the Cox model is not **collapsible**.

If you marginalise over $X$, you lose proportional hazards (in general), and the resulting HR coefficient for $Z$ you estimate is not equal to $\exp(\beta)$.

Hence you can never hope to estimate the assumed common individual level effect $\exp(\beta)$ from the trial.

This is very reasonable, and is in agreement with our conclusions about hoping to estimate (assumed common) individual level effects for binary outcomes.

# Outline

# $HR(t)$ is a valid causal contrast

We can express the hazard as

$$\lambda(t) = \frac{f(t)}{S(t)}$$

where $f(t)$ is the density function and $S(t) = \int_t^\infty f(u)du$ is the survival function.

Let $f^0(t)$ and $f^1(t)$ denote the densities of the potential failure times $T^0$ and $T^1$, and $S^0(t)$ and $S^1(t)$ the corresponding survival functions. Then

$$HR(t) = \frac{f^0(t)/S^0(t)}{f^1(t)/S^1(t)}$$

Thus $HR(t)$ is a contrast of a function of the two population densities $f^0(t)$ and $f^1(t)$, and is a valid population level causal effect.

# Interpreting $HR(t)$

$HR(t)$ **is** a population level causal effect.

$HR(t)$ is the ratio of instantaneous event rates in survivors at time $t$ if we assign the population to level 1 vs. level 0 of the treatment.

This doesn't rely on any assumption of proportional hazards.

# Interpreting $HR(t)$

Stochastic POs: due to the aforementioned selection/confounding issues $HR(t)$ is **not** an individual level effect at time $t$ (except under strong untestable assumptions).

Therefore, changes in $HR(t)$ **should not** be interpreted as representing solely changes in individual level treatment effect over time.

Deterministic POs: $HR(t)$ as defined here is not the HR in the subpopulation which would survive to $t$ under both treatment 0 and treatment 1.

# What if $HR(t)$ is constant over time?

If $HR(t)$ appears constant over time, can it be interpreted as meaning the individual level treatment effect is constant over time?

**No.**

Changes in $HR(t)$ can be some mixture of selection effects and time-varying individual level effects.

However, $HR(t) = \exp(\beta)$ a constant implies $S^1(t) = S^0(t)^{\exp(\beta)}$, and so

$$\exp(\beta) = \frac{\log\{S^1(t)\}}{\log\{S^0(t)\}}$$

But this interpretation is not nice nor easy to communicate.

# Is HR a useful causal effect measure?

$HR(t)$ is a valid causal effect measure, but is it answering a useful question?

$\Rightarrow$ For individuals, the answer seems no. For policy makers at the population level, maybe.

If marginally hazards are (approximately) proportional, is HR useful?

$\Rightarrow$ For individuals and policy makers, maybe. But even here, important to note HR is not a risk ratio, as is sometimes implied [7].

Other measures, e.g. risk differences or ratios for a landmark time, or differences/ratios of restricted mean survival time, may be preferable.

# Outline

# Conclusions

- Standard RCTs can estimate population level effects, but not individual level causal effects.
- $HR(t)$ **is a valid population level** causal effect, but its interpretation is subtle.
- $HR(t)$ is not an individual level causal effect, except under strong, implausible, and untestable assumptions.
- Changes in $HR(t)$ should not be interpreted simply as changes in individual level treatment effect over time.
- Even when $HR(t)$ is constant, alternatives to Cox's model may be preferable for quantifying causal effects.
- These slides, plus some discussion with Aalen at `www.thestatsgeek.com`

# References


[1] M Hernán and J Robins.
*Causal Inference*.
Boca Raton: Chapman & Hall/CRC, 2019.

[2] Miguel A Hernán.
The hazards of hazard ratios.
*Epidemiology (Cambridge, Mass.)*, 21(1):13, 2010.

[3] Steingrimsson JA, Hanley DF, and Rosenblum M.
Improving precision by adjusting for prognostic baseline variables in randomized trials with binary outcomes, without regression model assumptions.
*Contemporary Clinical Trials*, 54:18–24, 2017.

[4] Harrell FE Jr.
Biostatistics for Biomedical Research.
`http://hbiostat.org/doc/bbr.pdf`.
Accessed: 2018-08-24.

[5] Torben Martinussen, Stijn Vansteelandt, and Per Kragh Andersen.
Subtleties in the interpretation of hazard ratios.
*arXiv preprint arXiv:1810.09192*, 2018.

[6] Aalen OO, Cook RJ, and Røysland K.
Does Cox analysis of a randomized survival study yield a causal treatment effect?
*Lifetime Data Analysis*, 21(4):579–593, 2015.

[7] Rinku Sutradhar and Peter C Austin.
Relative rates not relative risks: addressing a widespread misinterpretation of hazard ratios.
*Annals of Epidemiology*, 28(1):54–57, 2018.