On the etiology of a (misspecified) medical risk prediction model

Thomas Alexander Gerds

Multi-state model

Multi-state models can be used to describe the life history of an individual:

- At all times each individual is in one of the states.
- The events are the transitions between the states.

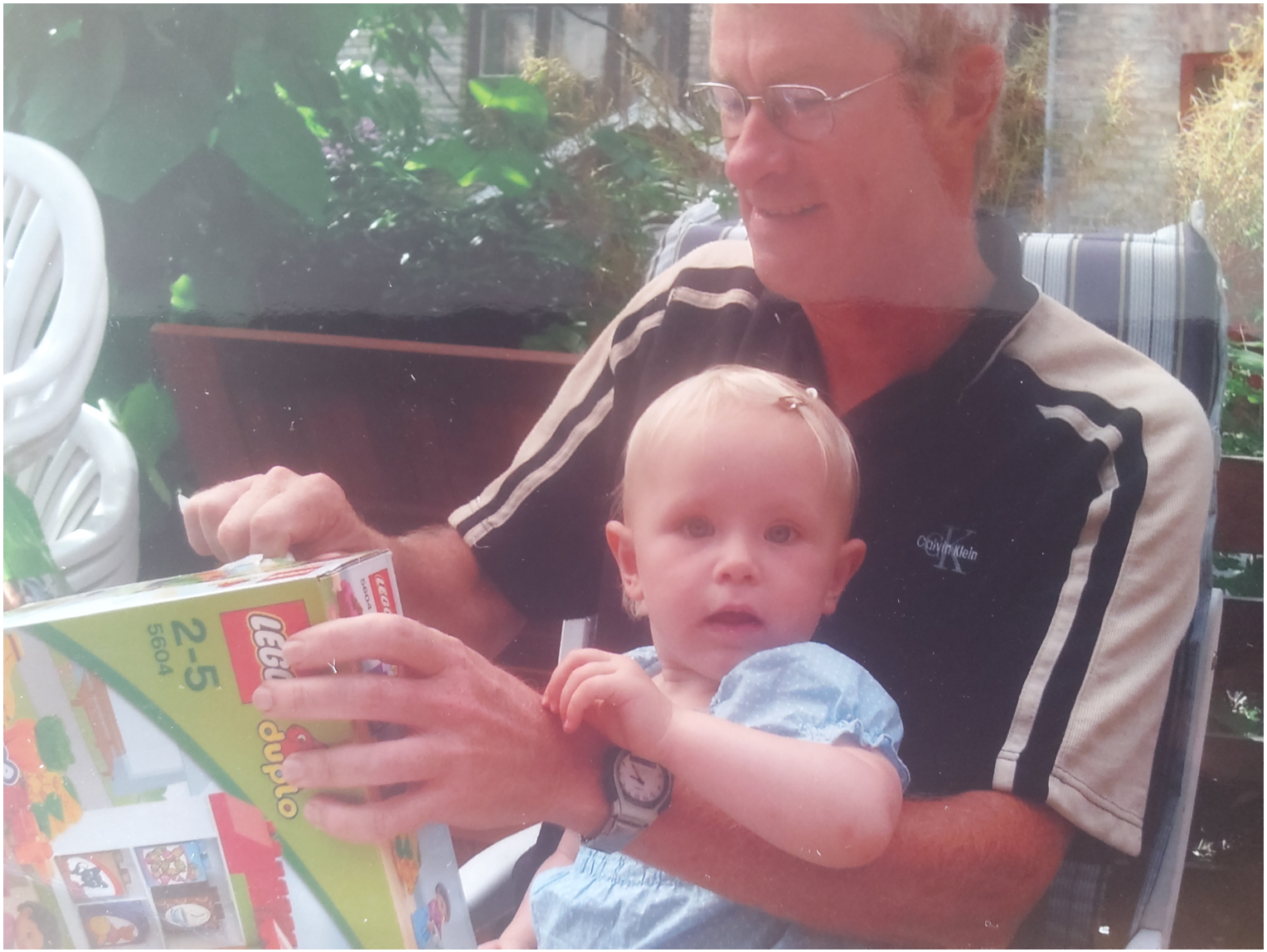
For example, PKA.

Tall basketball player

Married tall basketball player

Married tall basketball player with child

Married tall petanque player with children and grand child



Tall petanque player with children, grand children, a retired wife, and a sense for safety

RIGHT CENSORED IN

2022

Young number cruncher

Statistikerne studerer resultatet af hudkræft-undersøgelsen. Fra venstre: Per Kragh Andersen, Niels Keiding og Ole Olsen. **Foto: Bo Jarner**

Statistik mod hudkræft

Tal-eksperter hjælper forskere på mange områder.

Her er en liste med 200 mennesker, som er opereret for hudkræft. For hver af dem får du

kontrol efter en operation. Og de, der er næsten uden for fare, behøver ikke at komme til kon-

studenter på faget.

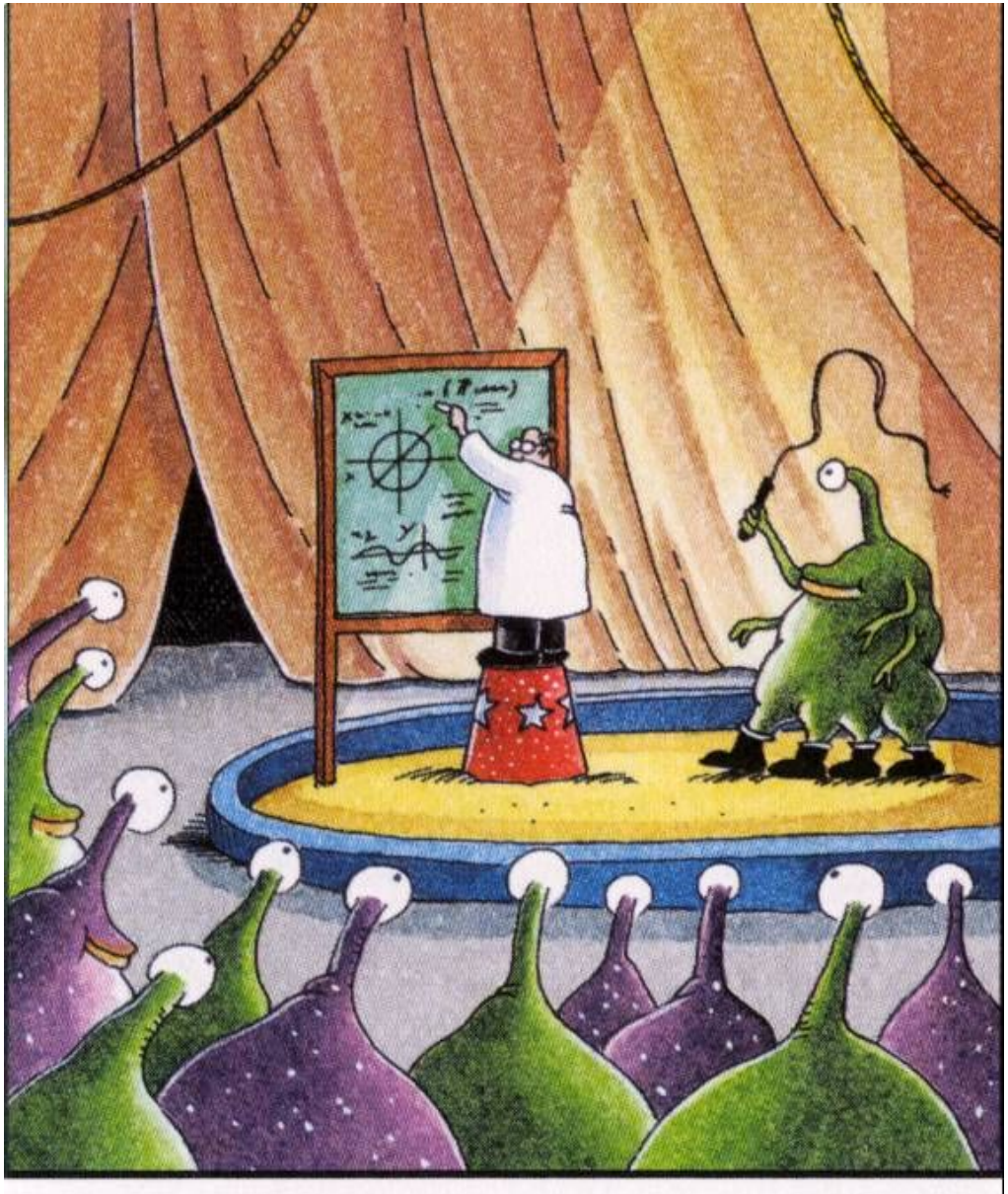
Selv blev han interesseret i statistik, fordi hans far som

Seen with a well-known influencer

Team player

Professor

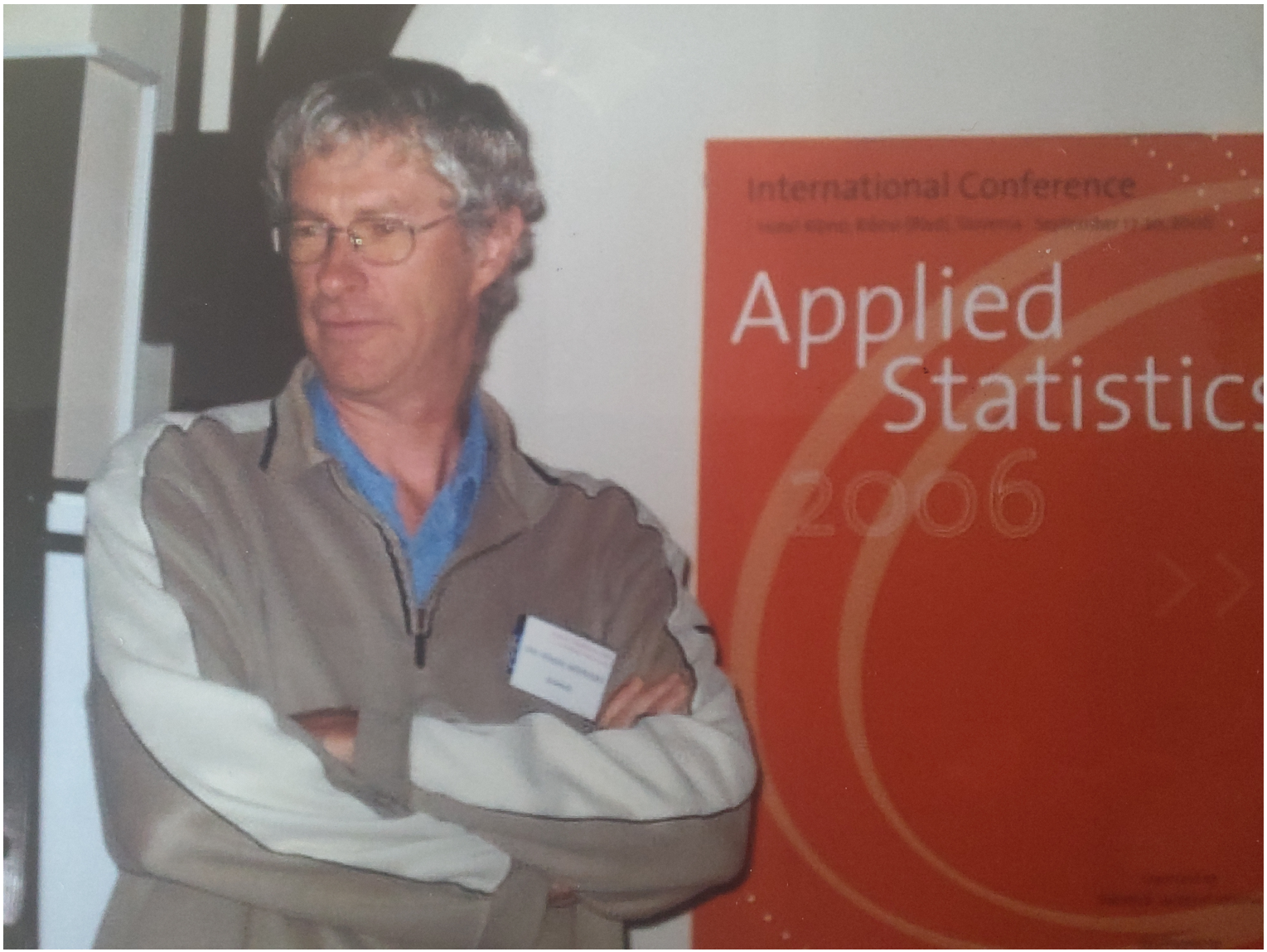
Abducted by an alien circus company, Professor Andersen is forced to write calculus equations in center ring.



Many publications

Boss of Biostat (in a limited period)

Medical statistician



RIGHT CENSORED AT

AGE 70

Research question

Daniel Mølager Christensen, Danish Heart Foundation:

Should people (who are not necessarily patients) take statins when they turn 70 years?

Methods: Danish Register Data, LTMLE (not Cox!)

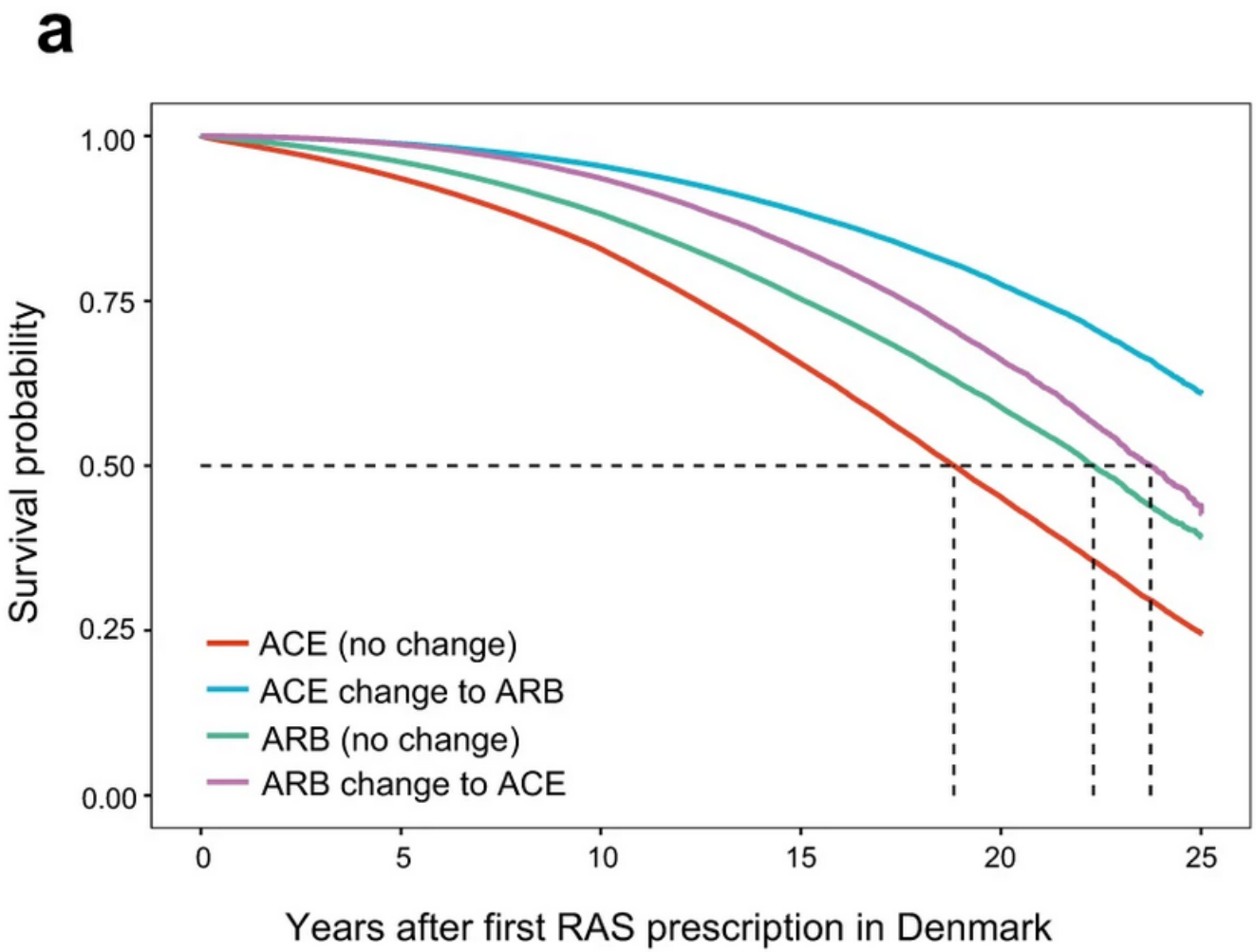
Analysis of Danish register data

- Time zero: age 70
- Exposure: statins vs no statins
- Outcome: Cardiovascular disease or all-cause death
- Exclusion/selection: people who have had the outcome

Problems:

- Don't know why people take statins
- Treatment dropin: Some patients start taking statins during the study period!

Research from Denmark (2021)



Let's ask a researcher from Denmark

Already in 1998

Lipoprotein Changes and Reduction in the Incidence of Major Coronary Heart Disease Events in the Scandinavian Simvastatin Survival Study (4S)

Terje R. Pedersen, MD; Anders G. Olsson, MD; Ole Færgeman, MD; John Kjekshus, MD; Hans Wedel, PhD; Kåre Berg, MD; Lars Wilhelmsen, MD; Torben Haghfelt, MD; Gudmundur Thorgeirsson, MD; Kalevi Pyörälä, MD; Tatu Miettinen, MD; Bjørn Christophersen, MD; Jonathan A. Tobert, MD, PhD; Thomas A. Musliner, MD; Thomas J. Cook, MS; for The Scandinavian Simvastatin Survival Study Group*

Background—The Scandinavian Simvastatin Survival Study (4S) randomized 4444 patients with coronary heart disease (CHD) and serum cholesterol 5.5 to 8.0 mmol/L (213 to 310 mg/dL) with triglycerides ≤2.5 mmol/L (220 mg/dL) to simvastatin 20 to 40 mg or placebo once daily. Over the median follow-up period of 5.4 years, one or more major coronary events (MCEs) occurred in 622 (28%) of the 2223 patients in the placebo group and 431 (19%) of the 2221 patients in the simvastatin group (34% risk reduction, $P<.00001$). Simvastatin produced substantial changes in several lipoprotein components, which we have attempted to relate to the beneficial effects observed.

Methods and Results—The Cox proportional hazards model was used to assess the relationship between lipid values (baseline, year 1, and percent change from baseline at year 1) and MCEs. The reduction in MCEs within the simvastatin group was highly correlated with on-treatment levels and changes from baseline in total and LDL cholesterol, apolipoprotein B, and less so with HDL cholesterol, but there was no clear relationship with triglycerides. We estimate that each additional 1% reduction in LDL cholesterol reduces MCE risk by 1.7% (95% CI, 1.0% to 2.4%; $P<.00001$).

Conclusions—These analyses suggest that the beneficial effect of simvastatin in individual patients in 4S was determined mainly by the magnitude of the change in LDL cholesterol, and they are consistent with current guidelines that emphasize aggressive reduction of this lipid in CHD patients. (***Circulation*. 1998;97:1453-1460.**)

Key Words: coronary disease ■ lipoproteins ■ cholesterol ■ simvastatin

Statistical methods: ... avoiding the problem of "using the future to predict the future."

Why not Cox? Results of a recent meta analysis

Role of Statins in the primary prevention of cardiovascular disease and mortality in the population with mean cholesterol in the near-optimal to borderline high range ...

Statin therapy was associated with a decreased risk of composite cardiovascular outcome (RR = 0.71, 95% CI: 0.62 to 0.82) [1]

- Immediate reaction: How very impressive! We should add statins to drinking water.
- Second thought: Wait a minute: what scale is used to quantify the effect? [2]

[1] Singh et al. (2020) Adv Prev Med.

[2] Probably the same scale on which corona vaccine efficacy is quantified

Why not Cox or Andersen-Gill regression?

The Annals of Statistics
1982, Vol. 10, No. 4, 1100–1120

COX'S REGRESSION MODEL FOR COUNTING PROCESSES: A LARGE SAMPLE STUDY

BY P. K. ANDERSEN AND R. D. GILL

Statistical Research Unit, Copenhagen; and Mathematical Centre, Amsterdam

The Cox regression model for censored survival data specifies that covariates have a proportional effect on the hazard function of the life-time distribution of an individual. In this paper we discuss how this model can be extended to a model where covariate processes have a proportional effect on the intensity process of a multivariate counting process. This permits a statistical regression analysis of the intensity of a recurrent event allowing for complicated censoring patterns and time dependent covariates. Furthermore, this formulation gives rise to proofs with very simple structure using martingale techniques for the asymptotic properties of the estimators from such a model. Finally an example of a statistical analysis is included.

1. Introduction. The Cox-model for censored survival data (Cox, 1972) specifies the hazard rate or intensity of failure $\lambda(t) = \lim_{h \downarrow 0} \mathscr{P}[T \leq t + h \mid T > t]$ for the survival time T of an individual with covariate vector z which may depend on the time t to have the form

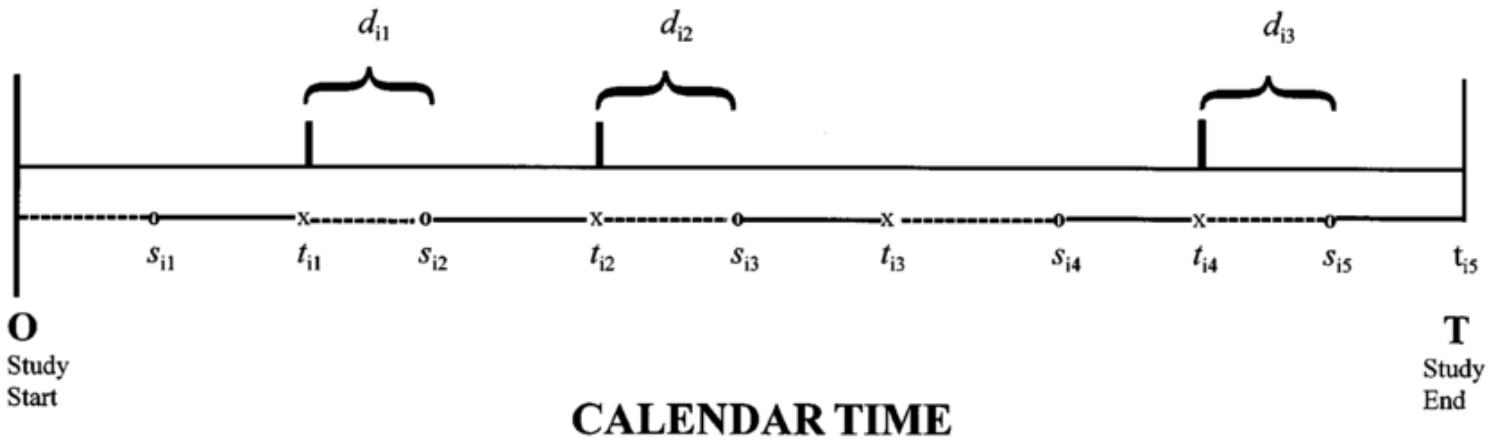
$$\lambda(t; z) = \lambda_0(t)\exp\{\beta_0' z(t)\}, \quad t \geq 0. \tag{1.1}$$

Here β_0 is a p-vector of unknown regression coefficients and $\lambda_0(t)$, the underlying hazard, is an unknown and unspecified nonnegative function. The statistical problem is the one of estimating β_0 and the function λ_0 on the basis of, say, n possibly right censored survival times $T_1, \cdots, T_n$ and the corresponding covariate vectors $z_1, \cdots, z_n$, where z_i is observed on $[0, T_i]$.

Cox (1972) suggested that inference on β_0 be based on the function

$$L(\beta) = \prod_{i=1}^{n} \left\{ \frac{e^{\beta' z_i(T_i)}}{\cdots} \right\}^{\delta_i} \tag{1.2}$$

The Andersen-Gill model



Schematic with notation and at-risk periods for the Andersen-Gill model. The vertical bars indicate morbid episode onset, and the d braces indicate duration of episode. The dashed portions of the bottom line are periods when the person is not under observation or at risk of experiencing the event; the solid portions are at-risk periods

I don't know if this is such a wise thing to do, Torben.

In 2020

Lifetime Data Analysis
https://doi.org/10.1007/s10985-020-09501-5

Check for updates

Subtleties in the interpretation of hazard contrasts

Torben Martinussen[1] · **Stijn Vansteelandt**[2,3] · **Per Kragh Andersen**[1]

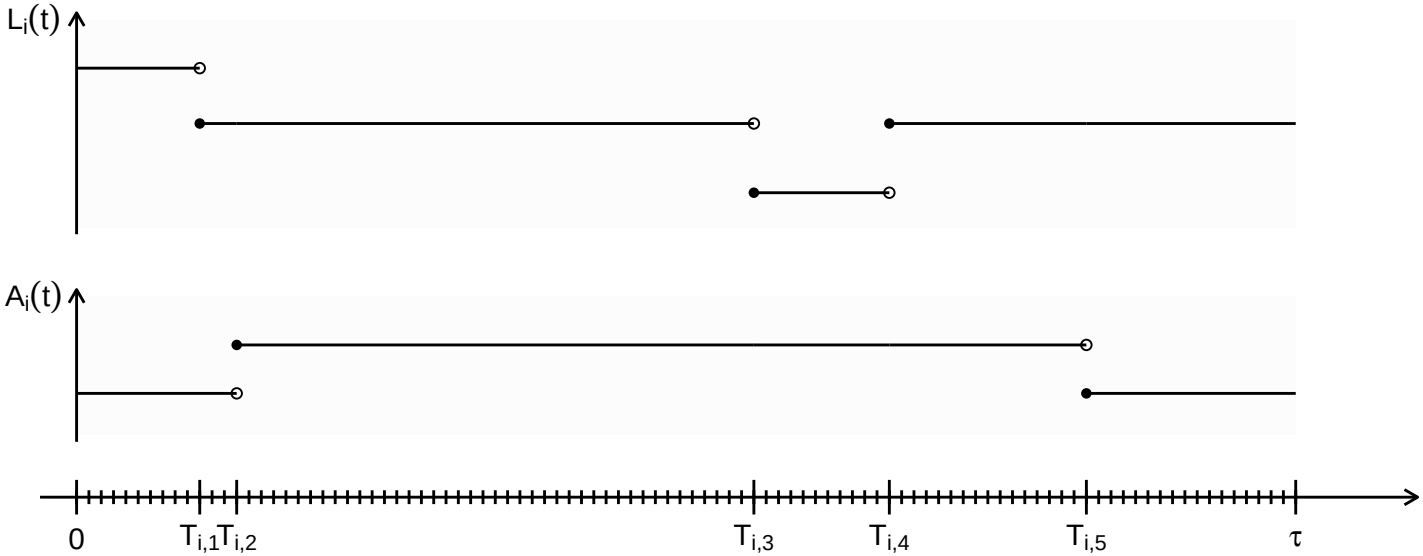
Received: 4 April 2019 / Accepted: 23 June 2020

Abstract

The hazard ratio is one of the most commonly reported measures of treatment effect in randomised trials, yet the source of much misinterpretation. This point was made clear by Hernán (Epidemiology (Cambridge, Mass) 21(1):13–15, 2010) in a commentary, which emphasised that the hazard ratio contrasts populations of treated and untreated individuals who survived a given period of time, populations that will typically fail to be comparable—even in a randomised trial—as a result of different pressures or intensities acting on different populations. The commentary has been very influential, but also a source of surprise and confusion. In this note, we aim to provide more insight into the subtle interpretation of hazard ratios and differences, by investigating in particular what can be learned about a treatment effect from the hazard ratio becoming 1 (or the hazard difference 0) after a certain period of time. We further define a hazard ratio that has a causal interpretation and study its relationship to the Cox hazard ratio, and we also define a causal hazard difference. These quantities are of theoretical interest only, however, since they rely on assumptions that cannot be empirically evaluated. Throughout, we will focus on the analysis of randomised experiments.

Scenario

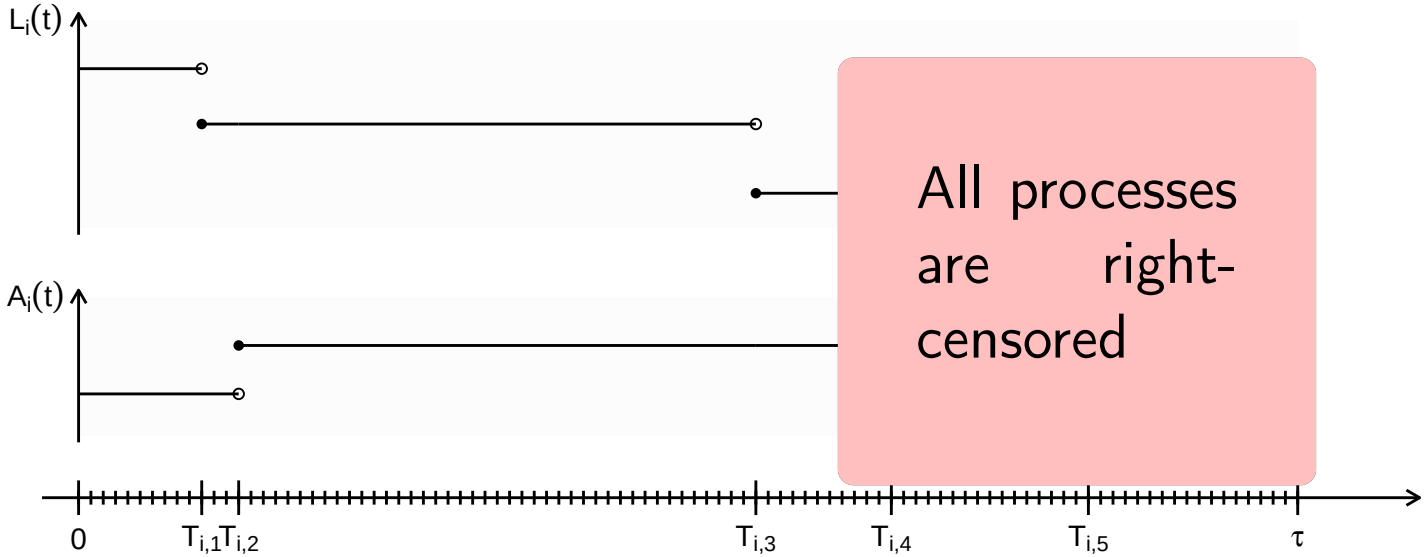
Consider a calendar time interval $[0, \tau)$.



- Treatment process history: $\bar{A}(t) = \{A(s) : 0 \leq s \leq t\}$
- Covariate process history: $\bar{L}(t) = \{L(s) : 0 \leq s \leq t\}$

Our scenario

Consider a calendar time interval $[0, \tau)$.



- Treatment process history: $\bar{A}(t) = \{A(s) : 0 \leq s \leq \min(t, T)\}$
- Covariate process history: $\bar{L}(t) = \{L(s) : 0 \leq s \leq \min(t, T)\}$

Truncated at time T where outcome or death happens.

The Cox regression model with time-dependent covariates

The outcome-specific hazard rate at time t depends on the treatment and covariate history:

$$\lambda(t|\bar{A}(t), \bar{L}(t)) = \lambda_0(t) \exp\{\beta \bar{A}(t) + \gamma \bar{L}(t)\}$$

The hazard ratio e^{β} can be estimated via maximum partial likelihood, Poisson regression or conditional logistic regression (nested case control design).

The Cox regression model with time-dependent covariates

The outcome-specific hazard rate at time t depends on the treatment and covariate history:

$$\lambda(t|\bar{A}(t), \bar{L}(t)) = \lambda_0(t) \exp\{\beta \bar{A}(t) + \gamma \bar{L}(t)\}$$

The hazard ratio e^{β} can be estimated via maximum partial likelihood, Poisson regression or conditional logistic regression (nested case control design).

Can anyone ask a clinically meaningful question such that e^{β} is the answer?

Instead of Cox: perhaps a personalized predicted risk?

Am J Epidemiol. 2021 Oct; 190(10): 2000–2014.
Published online 2021 Feb 17. doi: 10.1093/aje/kwab031
PMCID: PMC8485151
PMID: 33595074

Am J Epidemiol

Prediction of Cardiovascular Disease Risk Accounting for Future Initiation of Statin Treatment

Zhe Xu, Matthew Arnold, David Stevens, Stephen Kaptoge, Lisa Pennells, Michael J Sweeting, Jessica Barrett, Emanuele Di Angelantonio, and Angela M Wood

We derived age- and sex-specific prediction models including conventional risk factors and a time-dependent effect of statin initiation constrained to 25% risk reduction (from trial results).

Stick to this world?

Invited Commentary[3]: Treatment Drop-in – Making the Case for Causal Prediction

As soon as one entertains the need for "treatment naive-risk", one is targeting estimands that require causal reasoning to estimate well, given that they are hypothetical or counterfactual predictions.

We note that hypothetical prediction aims at answering "what if" questions about the future, while counterfactual prediction requires contemplating states contrary to what has truly happened, and this difference can be important.

[3]Sperrin et al. Am J Epidemiol. 2021 Oct; 190(10): 2015–2018.

From the statistical methods section [4]

In the first stage, to better utilize repeat risk factors and allow for incomplete data, error-free risk-factor values for

- SBP
- total cholesterol, HDL cholesterol
- smoking status

were estimated as best linear unbiased predictors from landmark age- and sex-specific multivariate mixed-effects linear regression models . . .

[4] Wood et al. (2021) Am J Epidemiol. 190(10)

From the statistical methods section [5]

In the second stage, 10-year statin-naive CVD risk was modeled using landmark age- and sex-specific Weibull models, with time since landmark age as the time scale and with the following risk factors:

- the most recently observed diabetes status
- hypertension treatment status
- estimated error-free risk-factor values for SBP, ...

and a time-dependent effect of statin initiation constrained to a 25% risk reduction as reported from published meta-analyses of trials.

[5] Wood et al. (2021) Am J Epidemiol. 190(10)

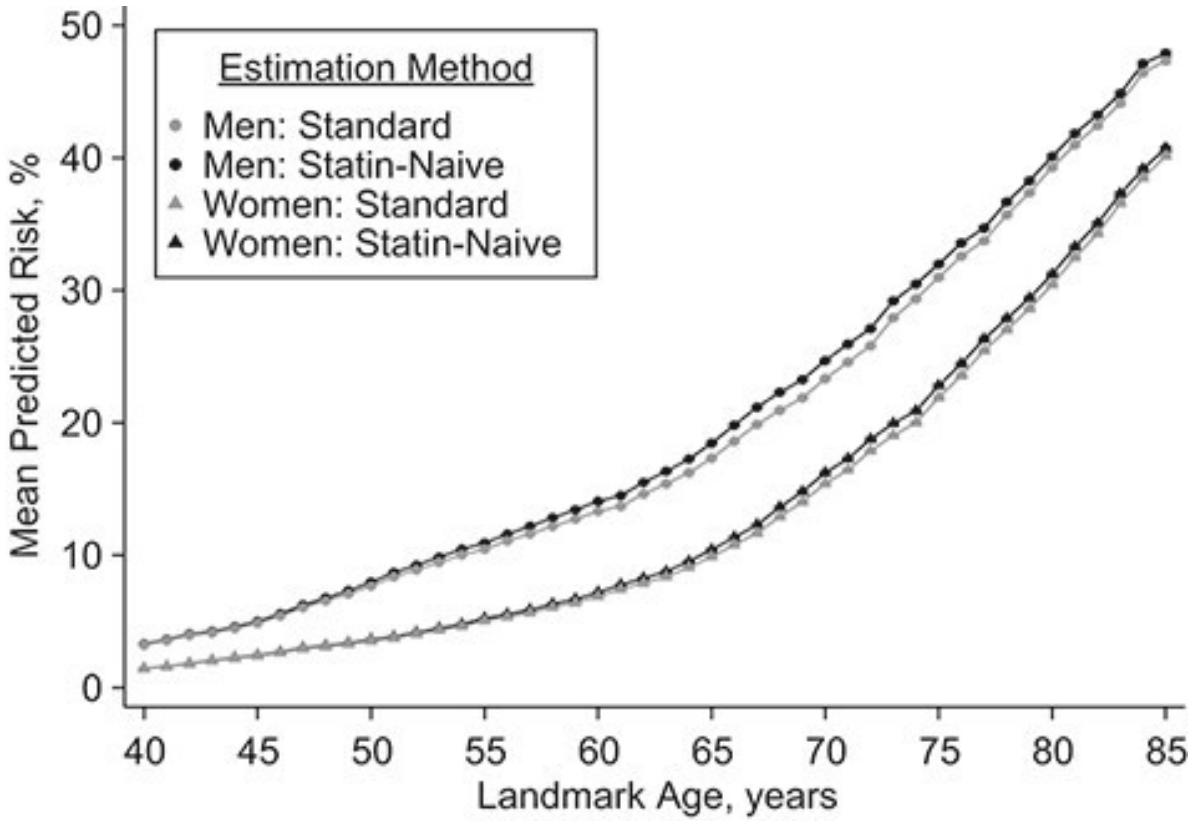
From the statistical methods section [6]

Reclassification measures, including the net reclassification improvement (NRI), with both continuous NRI and categorical NRI using the predicted 10-year risk cutoff at ... together with the integrated discrimination index (IDI), were used to compare the statin-naive and the standard CVD risks at ages 40, 50, 60, and 70 years.

But: Hilden (2014, Epidemiology): On NRI, IDI and "Good-Looking" statistics with nothing underneath

[6] Wood et al. (2021) Am J Epidemiol. 190(10)

Wood et al. (2021)



In the validation data, the means of 10-year statin-naive risks were slightly higher than standard CVD risks, especially among 60- to 70-year-olds.

Suggestive conclusions

- Don't put statins into drinking water.
- No one is like Per, not at the landmark age, and not 3 years 2 months and 4 days later, but not in a different way as at the landmark age.
- Final question:

Suggestive conclusions

- Don't put statins into drinking water.
- No one is like Per, not at the landmark age, and not 3 years 2 months and 4 days later, but not in a different way as at the landmark age.
- Final question:

Mirror, mirror on the wall, who is the most efficient statistician of them all?

Thank you, Per.