Fifty years with the Cox proportional hazards model

DSBS 30 years, April 2022

Per Kragh Andersen
Section of Biostatistics, University of Copenhagen

Paper in *Journal of the Indian Institute of Science*

Overview of talk

1. Past:
   - The paper

     Cox, D.R. (1972). Regression models and life-tables (with discussion). *J. Roy. Statist. Soc. B*, **34**, 187–220

     is one of the most cited statistical papers ever with >50000 citations
   - Influence on statistical and medical literature
2. Present:

   The model is being criticized for
   - being too restrictive and, yet, hard to interpret
   - being inferior for prediction compared to machine learning
   - lacking a causal interpretation
3. Future?

David Roxbee Cox 1924-2022

(recent obituary by Vern Farewell in *Pharmaceutical Statistics*)

The Cox (1972) paper

- T failure time, $\lambda(t)$ *hazard*
- $\mathbf{z} = (z_1, \ldots, z_p)$ "further measured (explanatory) variables", hazard

$$\lambda(t; \mathbf{z}) = \exp(\mathbf{z}\boldsymbol{\beta})\lambda_0(t)$$

  ($\mathbf{z}$ may depend on time; $\exp(\mathbf{z}\boldsymbol{\beta})$ may be $h(\mathbf{z}, \boldsymbol{\beta})$)

- A "conditional" likelihood for $\boldsymbol{\beta}$ when $\lambda_0(t)$ is arbitrary:

$$\prod_{i=1}^{k} \frac{\exp(\mathbf{z}_{(i)}\boldsymbol{\beta})}{\sum_{\ell \in \mathcal{R}(t_{(i)})} \exp(\mathbf{z}_{(\ell)}\boldsymbol{\beta})}$$

  $t_{(1)} < \ldots < t_{(k)}$ distinct observed failure times

- Ad hoc estimator for survival function corresponding to $\lambda_0(t)$
- Example: Freireich et al. data with a test for proportional hazards using a time-dependent variable $z \cdot (t - 10)$, and alternative models with $\lambda_0(t)$ piecewise constant or of a Weibull form
- "Physical interpretation of model": "model intended as ... convenient, flexible and yet entirely empirical", discussion of alternative AFT model
- Several discussion contributions have also been widely cited (e.g., Downton, Breslow, Kalbfleisch & Prentice)

Short-term influence on statistical literature

The paper contains several new ideas that it took the statistical community at least a decade to comprehend:

- The unusual likelihood construction was further discussed in *Biometrika* papers by Kalbfleisch & Prentice (1973), Crowley (1974), Cox (1975) leading to the new concept "partial likelihood". (Also: Johansen, 1983, profile likelihood; Jacobsen, 1984, MLE in topologically extended model *Int. Statist. Rev.*)
- The unusual "semi-parametric" model formulation led to efficiency considerations compared to parametric models (Efron, 1977, *JASA*; Oakes, 1977, *Biometrika*)

  Also, how to estimate the baseline hazard and, thereby, survival probabilities $S(t \mid z)$ for given covariates (Breslow, 1972, 1974; Link, 1979; Kalbfleisch & Prentice, 1980)
- The combination of an unusual likelihood, the semi-parametric nature of the model and censoring led to many alternative approaches to studying asymptotic properties of estimators from the model (Tsiatis, 1981; Andersen & Gill, 1982, Bailey, 1983, *Ann. Statist.*; Næs, 1982, *Scand. J. Statist.*)

The 'modern' approach to inference in the Cox model

The data for subject i are represented by a *counting process* $N_i(t) = I(T_i \wedge C_i \leq t, T_i \leq C_i)$ and an *at risk indicator* $Y_i(t) = I(T_i \wedge C_i \geq t)$. The counting process has the decomposition:

$$N_i(t) = \int_0^t \lambda_i(s; \boldsymbol{z}_i) ds + M_i(t)$$

where $M_i(t)$ is a *martingale* and $\lambda_i(s; \mathbf{z}_i) = Y_i(s) \exp(\mathbf{z}_i \boldsymbol{\beta}) \lambda_0(s)$.

NB: independent censoring!

The *likelihood* is given by Jacod's formula

$$L = \prod_{i=1}^n \exp\big(-\int_0^\tau Y_i(t) \lambda_i(t; \mathbf{z}_i) dt\big) \prod_t \big(Y_i(t) \lambda_i(t; \mathbf{z}_i)\big)^{dN_i(t)}$$

and profiling out $\lambda_0(t)$ leads to Cox's partial likelihood:

$$PL(\boldsymbol{\beta}) = \prod_i \prod_t \Big(\frac{Y_i(t) \exp(\mathbf{z}_i \boldsymbol{\beta})}{\sum_j Y_j(t) \exp(\mathbf{z}_j \boldsymbol{\beta})}\Big)^{dN_i(t)}.$$

The 'modern' approach to inference in the Cox model (ctd.)

The score

$$\mathbf{U}_\tau(\boldsymbol{\beta}) = \frac{\partial}{\partial \boldsymbol{\beta}} \log(PL(\boldsymbol{\beta}))$$

$$= \sum_i \int_0^\tau Y_i(t)\big(\mathbf{z}_i - \frac{\sum_j Y_j(t)\mathbf{z}_j \exp(\mathbf{z}_j\boldsymbol{\beta})}{\sum_j Y_j(t)\exp(\mathbf{z}_j\boldsymbol{\beta})}\big) dN_i(t)$$

is a martingale when evaluated at the true parameter $\boldsymbol{\beta}_0$.

Martingale CLT may be used for the score and standard Taylor expansions give asymptotic normality of $\widehat{\boldsymbol{\beta}}$. Also properties of the Breslow estimator

$$\widehat{\Lambda}_0(t) = \int_0^t \frac{\sum_i dN_i(s)}{\sum_i Y_i(s)\exp(\mathbf{z}_i\widehat{\boldsymbol{\beta}})}$$

and the plug-in estimator for the survival function $S(t \mid \mathbf{z})$ may be derived.

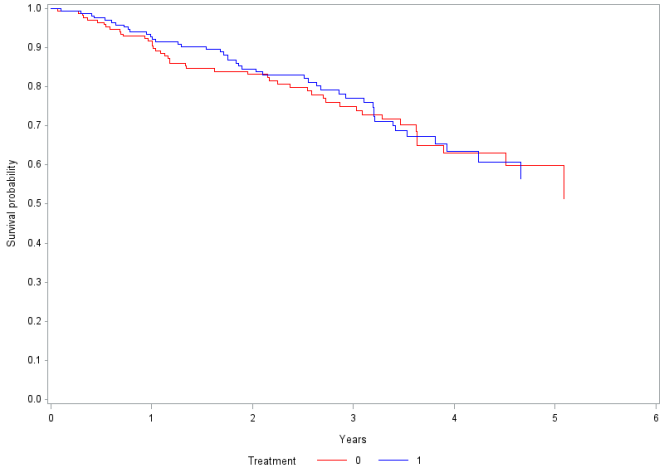
Influence on medical journals

- The majority of the >50000 citations are from medical journals and "the Cox model" has become the standard tool for regression analysis of failure time data: you have to give an argument for doing something else!
- The concurrent initiations of cancer clinical trials and development of flexible software strengthened this trend
- Many of these applications are not very carefully conducted and little attention is paid to aspects like
  - goodness of fit of the model (proportional hazards, log-linearity, ...)
  - validity of asymptotic results (numbers of subjects, failures, parameters)
- Does it matter?

Example: The PBC3 trial in liver cirrhosis

- Randomized trial in primary biliary cirrhosis 1983-88 (Lombard et al., 1993, *Gastroenterol.*), 6 European centers
- Patients randomized to CyA ($n = 176$) or placebo ($n = 173$)
- Composite end-point of either death or liver transplantation (CyA: 44, placebo: 46)

Kaplan-Meier plot (NB: $>$ 60000 citations)



Example: The PBC3 trial in liver cirrhosis (ctd.)

Cox model including only treatment gives $\widehat{\beta}_1 = -0.059$ with an estimated standard deviation of 0.211, leading to an estimated hazard ratio of $\exp(-0.059) = 0.94$ with 95% confidence limits from 0.62 to 1.43.

Randomization was not perfect and adjustment for biochemical variables $z_2 =$ Se-Albumin and $z_3 = \log_2$(Se-Bilirubin) gives $\widehat{\beta}_1 = -0.574$ (SD=0.224), leading to an estimated hazard ratio of 0.56 with 95% confidence limits from 0.36 to 0.87.

Survival curves for given covariates may now be predicted and a single set of curves for the two treatment groups may be obtained using the g-formula:

$$\widehat{S}_j(t) = \frac{1}{n} \sum_i \widehat{S}(t \mid z_1 = j, z_{2i}, z_{3i}), \quad j = 0, 1.$$

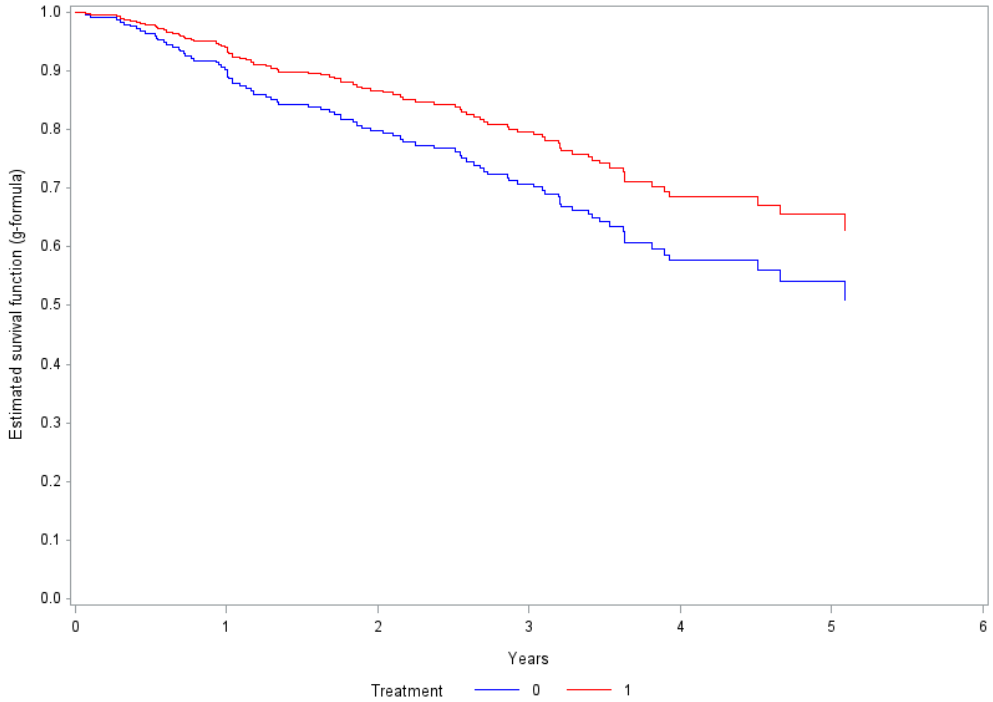


Figure 1: Estimated survival curves in the two treatment groups in the PBC3 trial based on the g-formula. The estimated risk difference at 2 years is 0.867-0.799=0.068, and it has an estimated SD of 0.027 based on 200 bootstrap replications.

Long term influence on statistical literature - 1

The Cox paper deals with survival data and was innovative by modeling the *hazard function*

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P(T \leq t + \Delta t \mid T > t)}{\Delta t} = \frac{f(t)}{S(t)},$$

so there has, obviously, been an influence on the literature on survival and event history analysis:

- Multi-state models, including competing risks and recurrent events: transition intensities (e.g., cause-specific hazards) (Prentice et al., *Biometrics*, 1978; Kalbfleisch & Prentice, Wiley, 1980, 2002; Andersen, Borgan, Gill & Keiding, Springer, 1993):

$$\lambda_{hj}(t) = \lim_{\Delta t \to 0} \frac{P(X(t + \Delta t) = j \mid X(t) = h)}{\Delta t}.$$

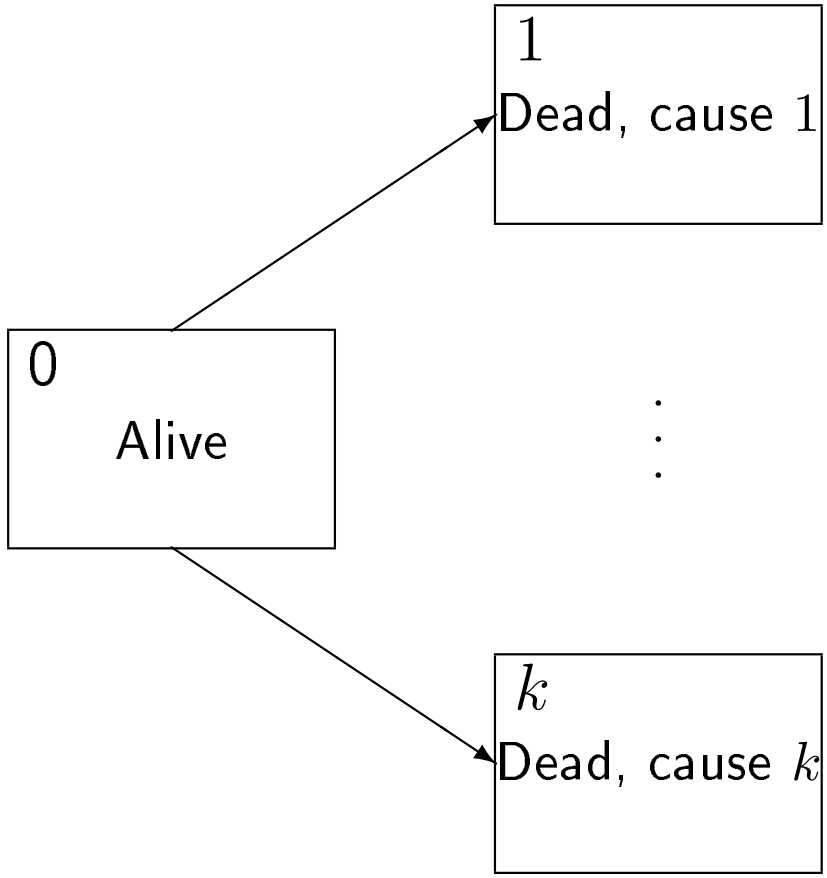


Figure 2: The competing risks model with k causes of death.

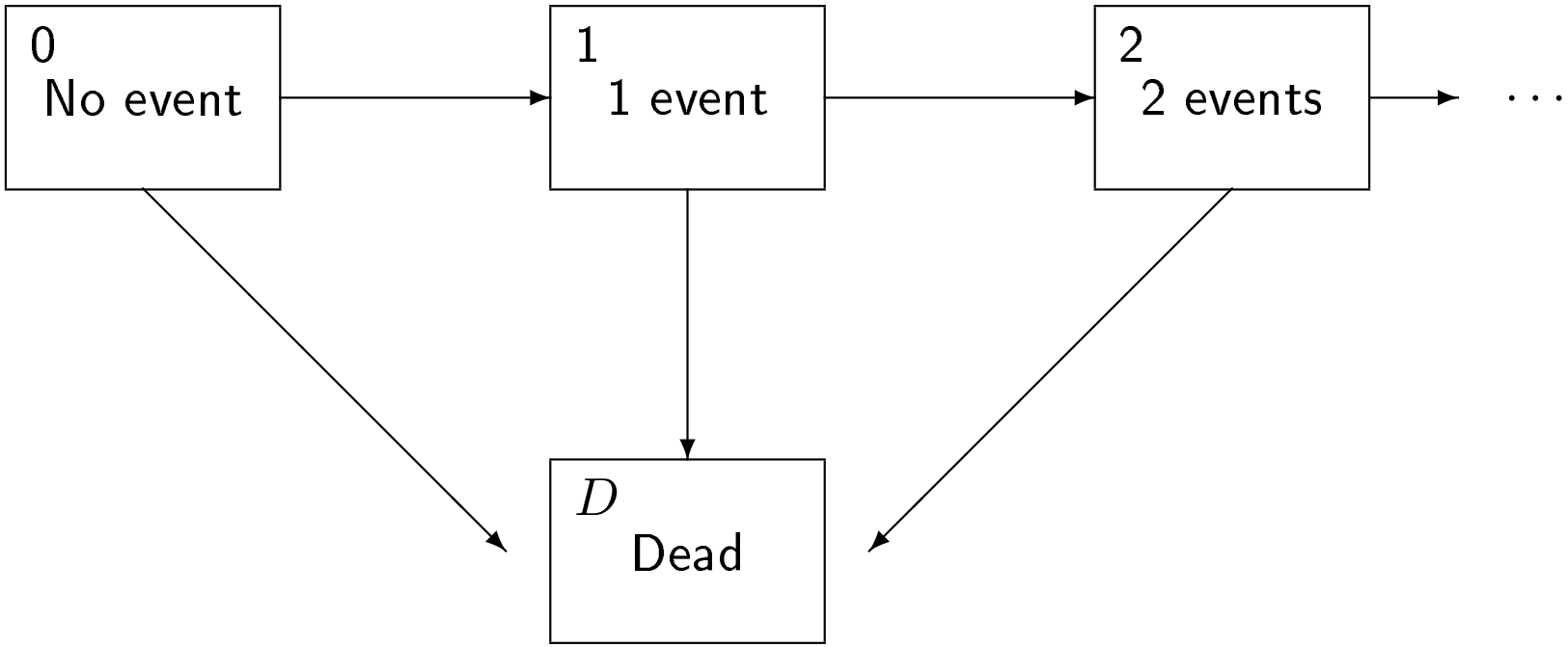


Figure 3: A multi-state model for recurrent events with a terminal event and no 'gaps' between at-risk periods.

- Fine-Gray model for a competing risks sub-distribution hazard, i.e. for the improper random variable, $T_h = \inf_{t>0}(X(t) = h)$:

  $\lim_{\Delta t \to 0} \frac{P(T_h \leq t + \Delta t | T_h > t)}{\Delta t}$,

  leading to a model for the cumulative incidence (Fine & Gray, *JASA*, 1999):

  $$-\log(1 - P(X(t) = h \mid \mathbf{z})) = \Lambda_0(t) \exp(\mathbf{z}\boldsymbol{\beta}).$$

- Models for the mean numbers of recurrent events in $[0, t]$ (without or with competing risks, e.g., Lin et al., 2000, *JRSS (B)*; Ghosh & Lin, *Stat. Sinica*, 2002; Cook & Lawless, Springer, 2007):

  $$\mu(t \mid \mathbf{z}) = \mu_0(t) \exp(\mathbf{z}\boldsymbol{\beta}).$$

- Model for (censored) medical costs (e.g., Lin, *Biometrics*, 2000):

  $$\mu(t \mid \mathbf{z}) = \mu_0(t) \exp(\mathbf{z}\boldsymbol{\beta}).$$

Long term influence on statistical literature - 1 (ctd.)

- Covariate measurement errors (e.g., Carroll, Ruppert, Stefanski & Crainiceanu, Chapman and Hall, 2006) (though additive hazard models include measurement errors more neatly)
- Random effects ("frailty") models (e.g., Duchateau & Janssen, Springer, 2007) (though additive hazard models include random effects more neatly):

$$\lambda(t \mid \mathbf{z}; W = w) = w \cdot \lambda_0(t) \exp(\mathbf{z}\boldsymbol{\beta})$$

- In general, the non-parametric baseline hazard, $\lambda_0(t)$ has led to de-emphasazing parametric inference in survival analysis,

though there are exceptions:

Martinussen & Scheike and Aalen, Borgan & Gjessing (Springer, 2006, 2008) give additive hazard models considerable coverage,

Lawless (Wiley, 2002) discusses parametric models

If you wish to be taken seriously in survival analysis then you should 'do it' for the Cox model!

Long term influence on statistical literature - 2

Google gives $\approx$ 100,000,000 hits for "partial likelihood"

The idea of having a full likelihood from which informative factors are selected in an intelligent way has gained widespread popularity

- Wong (1986), *Ann. Statist.*: General asymptotic theory (consistency, asymptotic normality, asymptotic efficiency) for maximum partial likelihood estimators
- Slud (1992), *Scand. J. Statist.*: Partial likelihood for continuous-time stochastic processes

  Gill (1992), *Scand. J. Statist.*: Marginal partial likelihood (e.g. frailty models)
- Borgan, Goldstein & Langholz (1995), *Ann. Statist.*: Methods for the analysis of sampled cohort data in the Cox proportional hazards model.

  Partial likelihood for $\boldsymbol{\beta}$ in the intensity process $\lambda_{(i,\mathbf{r})}(t) = Y_i(t)\lambda_0(t)\exp(\mathbf{z}_i(t)\boldsymbol{\beta}\pi_t(\mathbf{r} \mid i)$ for the counting process $N_{(i,\mathbf{r})}(t)$:

$$\prod_{u\in[0,\tau]}\prod_{\mathbf{r}\in\mathcal{P}}\prod_{i\in\mathbf{r}}\left(\frac{Y_i(u)\exp(\mathbf{z}_i(u)\boldsymbol{\beta})\pi_u(\mathbf{r}|i)}{\sum_{l\in\mathbf{r}}Y_l(u)\exp(\mathbf{z}_l(u)\boldsymbol{\beta})\pi_u(\mathbf{r}|l)}\right)^{\mathrm{d}N_{(i,\mathbf{r})}(u)}$$

Long term influence on statistical literature - 3

Google gives $\approx$ 46,700,000 hits for "semi-parametric"

- Begun, Hall, Huang & Wellner (1983), *Ann. Statist.*: Information and asymptotic efficiency in parametric-nonparametric models
- van der Vaart & Wellner (1996), Springer: *Weak Convergence and Empirical Processes*
- Bickel, Klaassen, Ritov & Wellner (1998), Springer: *Efficient and Adaptive Estimation for Semiparametric Models*
- Tsiatis (2006), Springer: *Semiparametric Theory and Missing Data*

Conclusions concerning the past

- This is one of the most influential statistical papers ever
- Its influence is widespread:
  - Both statistical and medical literature
  - Not only survival analysis, but also other branches of mathematical and applied statistics
- The paper has set standards for survival and event history analysis

Present (critique) 1a

One of the beauties of the Cox model is that it provides a *one-number summary* of the treatment effect in a randomized trial (or the exposure effect in an epidemiological cohort study).

This has been criticized for being too simple:

- van Houwelingen & Putter (2012, CRC/Chapman & Hall, *Dynamic Prediction in Survival Analysis*) have a full chapter about 'Mechanisms explaining violation of the Cox model'
- Stensrud & Hernan (2020, *JAMA* letter 'Why test for proportional hazards?'):

  'In practice, a constant hazard ratio does not occur for most medical applications'
- Flexible models relaxing the proportional hazards assumption have been developed

Present (critique) 1b

The hazard ratio is not a relative risk:

In the case of no competing risks:

$$RR = \frac{1 - S(t \mid z = 1)}{1 - S(t \mid z = 0)} = \frac{1 - \exp(-\Lambda_0(t) \exp(\beta))}{1 - \exp(-\Lambda_0(t))}$$

and only in a 'low risk' situation (i.e., $\exp(-\Lambda(t)) \approx 1 - \Lambda(t)$) is $RR \approx \exp(\beta)$.

In the presence of competing risks, this is 'even worse'.

Present (critique) 1c

Non-collapsibility of the hazard ratio:

Let z_1 and z_2 be *independent* and consider two Cox models:

$$\lambda_0(t)\exp(\beta_1 z_1)$$

and

$$\widetilde{\lambda}_0(t)\exp(\widetilde{\beta_1} z_2 + \widetilde{\beta}_2 z_2).$$

If $\widetilde{\beta}_2 \neq 0$, i.e. when z_2 is associated with survival, then $\beta_1 \neq \widetilde{\beta}_1$.

This means that whenever an estimate from a Cox model is quoted, it should be emphasized which other variables were included in the model.

NB: Same situation for logistic regression.

Present (critique) 2

Machine learning outperforms the Cox model for prediction:

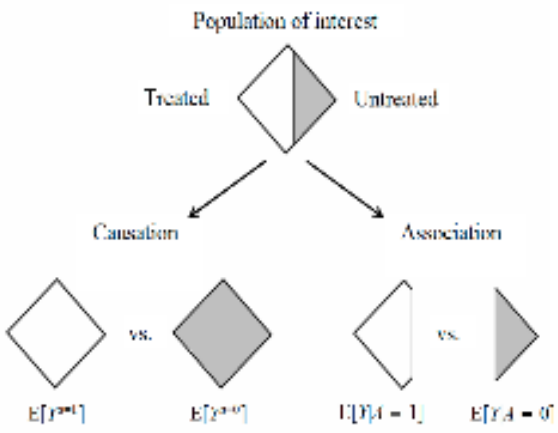
- Kim et al. (2019, *JMIR Medical Informatics*): compared a deep learning algorithm with versions of the Cox model
- Leger et al. (2017, *Nature Sci. Reports*): compared a number of machine learning methods with versions of the Cox model
- Beaulac et al. (2020, *arXiv*): similarly
- See also Fang et al. (2017, *JRSS (B)*), Hao et al. (2021, *Mathematics*), and probably many more ...

Present (critique) 3

The hazard ratio does not have a causal interpretation:

- Hernan (2010, *Epidemiology*: The hazards of hazard ratios)
- Aalen, Cook, Røysland (2015, *LIDA*)
- Martinussen, Vansteelandt, Andersen (2020, *LIDA*, Subtleties in the interpretation of hazard contrasts)

T^1, T^0 potential outcomes under treatment and control:



Present (critique) 3 (ctd.)

Marginal structural Cox model: $\lambda(t,a) = \lambda_0(t)\exp(\beta a)$

$$\exp(\beta) = \frac{\log P(T^1 > t)}{\log P(T^0 > t)}$$

is a causal contrast, however the interpretation

$$\exp(\beta) = \frac{\lim_{\Delta t \to 0} P(t \leq T^1 < t + \Delta t \mid T^1 > t)}{\lim_{\Delta t \to 0} P(t \leq T^0 < t + \Delta t \mid T^0 > t)}$$

is not causal.

Thus, if $\lambda(t,1)/\lambda(t,0) < 1$ for $t < \tau$ and $\lambda(t,1)/\lambda(t,0) = 1$ for $t > \tau$ then saying that 'treatment only works for $t < \tau$' is not justified.

Future?

Robert Storm Petersen (1882-1949)

Quotations (?):

"Statistics is like a street-lamp - not very enlightening but convenient to lean on"

"Don't make predictions, particularly not about the future"

Future?

In view of this criticism we may ask the question:

What is the future of the Cox model?

I certainly think that it does have a future – but its role may change from being *the* default method to being one component of larger systems.

Machine learning

- Penalized regression (Gui & Li, 2006, *Bioinformatics*): Penalized Cox model
- Deep learning (e.g., Katzman et al., 2018, *BMC Med. Res. Meth.*): '... A Cox PH deep neural network'
- Kvamme et al. (2019, *J. Machine Learning*): extended the Cox model to a neural network
- – and, actually, Kattan (2003, *J. Urology*) found that machine learning algorithms not always outperform the Cox model

Causal inference etc.

- The Cox model is useful in connection with using the g-formula:

$$\widehat{P}(T^a > t) = \frac{1}{n}\sum_i \widehat{P}(T > t \mid \mathbf{z}_i, a)$$

- The Cox model may be useful for estimating a parameter like the *restricted mean survival time*

$$E(T \wedge \tau) = \int_0^\tau P(T > t)dt$$

- The Cox model will often be part of ensemble methods, e.g., the TMLE

Conclusions

- It *is*, indeed, a simple method that provides a one-number summary of survival curves
- It is so well-established in the medical world that it is likely to be still used
- Even in machine learning, the Cox model is a bench-mark against which other methods are compared
- The model is a much used tool in causal inference
- .

Conclusions

- It *is*, indeed, a simple method that provides a one-number summary of survival curves
- It is so well-established in the medical world that it is likely to be still used
- Even in machine learning, the Cox model is a bench-mark against which other methods are compared
- The model is a much used tool in causal inference
- The t-test discussed in Gossett's paper from 1908 has proven still to be useful after > 100 years ...