



Fifty Years with the Cox Proportional Hazards Regression Model

Per Kragh Andersen*

Abstract | The 1972 paper introducing the Cox proportional hazards regression model is one of the most widely cited statistical articles. In the present article, we give an account of the model, with a detailed description of its properties, and discuss the marked influence that the model has had on both statistical and medical research. We will also review points of criticism that have been raised against the model.

Keywords: Covariates, Hazard function, Partial Likelihood, Regression model, Survival analysis

1 Introduction

Many biostatistical problems deal with the way in which aspects of the distribution of a certain *outcome variable*, Y is related to *covariates* (or *explanatory variables*), $\mathbf{z} = (z_1, \dots, z_p)$ (e.g., Andersen and Skovgaard⁵). Standard statistical models in this area include the *linear regression* model $Y = \boldsymbol{\beta}^T \mathbf{z} + \varepsilon$ for a quantitative Y where the *error terms* $\varepsilon_1, \dots, \varepsilon_n$ are i.i.d. mean-zero random variables, some times assumed to follow a normal distribution. A less restrictive model assumes Y_1, \dots, Y_n to be conditionally independent given $\mathbf{z}_1, \dots, \mathbf{z}_n$ with conditional expectations $E(Y_i | \mathbf{z}_i) = \boldsymbol{\beta}^T \mathbf{z}_i$ that are linear in the covariates but with no further distributional assumptions. In either case, the β_j -coefficients have the interpretation of *mean differences* for Y for a one-unit difference in the associated covariate z_j . For a binary Y , a standard regression model is the *logistic* model $\log(P(Y = 1 | \mathbf{z}) / (1 - P(Y = 1 | \mathbf{z}))) = \beta_0 + \boldsymbol{\beta}^T \mathbf{z}$ where the β_j -coefficients ($j = 1, \dots, p$) are *log(odds ratios)* for the event $\{Y = 1\}$ for a one-unit difference in the associated covariate z_j .

In *survival analysis*, the outcome is the survival time, say T , measured *from* a well-defined *time origin*, such as time of disease diagnosis, time of treatment initiation, or time of randomization, *to* the occurrence of an event of interest, often death from any cause. Thus, in survival analysis, the outcome variable is quantitative and one may wonder if standard linear regression models, as just described, would be applicable? Similarly, if interest focuses on the survival status (dead

or alive) at a certain time point, τ then one may wonder whether a logistic regression model for the binary outcome $Y = I(T \leq \tau)$ (where $I(A)$ is the indicator for the event A) would be applicable? In both cases, the answer is ‘not directly’ because, in survival analysis, one has to face the problem of *incomplete observation* of T caused by *right-censoring*. Thus, practical restrictions in data collection will typically have the consequence that, for some subjects i , only the information that the survival time T_i exceeds a certain observed value, say C_i , is available. This would be the case if, at the time of analysis, subject i is still alive and it is not known when that subject will ultimately die. Alternatively, subject i may drop out of the study (at time C_i) before its planned termination. In both cases, the observation of T_i is said to be right-censored at C_i and the presence of censoring has the consequence that standard methods such as linear or logistic regression analysis are not directly applicable and alternative methods of inference are needed.

Before discussing such alternatives, it should be mentioned that a natural question that arises is whether the available, incompletely observed survival times allow valid inference for the distribution of T . This is the assumption of *independent censoring* meaning that the knowledge at time t that a subject is still alive and uncensored should carry no information on the survival time T over and above the fact that $T > t$. This is a crucial assumption for any survival analysis technique, an assumption that,

¹ Section of Biostatistics, University of Copenhagen, Øster Farimagsgade 5, 1353 Copenhagen K, Denmark.
*pka@biostat.ku.dk

unfortunately, is untestable based on the available incomplete data. Thus, in any study involving survival analysis, it should be considered whether the mechanisms causing censoring can be assumed to be ‘independent’.

Returning to the problem of doing regression analysis for survival data, an obvious approach would be to try to adapt the linear regression model, e.g. by letting $Y = \log(T)$ and assuming $Y = \beta^T z + \varepsilon$ combined with a suitable parametric choice for the distribution of the error terms. Possible such choices would be a normal distribution or an extreme value distribution, the latter corresponding to an assumption of T following a Weibull or, more specially, an exponential distribution. However, since little information on the shape of the right-hand tail of the distribution of T is available due to censoring, one is typically reluctant to impose such parametric assumptions. This was the situation when the paper by Cox¹³ was conceived. In this article, we will describe the *Cox proportional hazards regression model* with a focus on how inference is done in the model and on the statistical properties of the resulting estimators. We will also give an account of the influence that the model has had in the field of statistics over the past 50 years and discuss some points of criticism of the model that have been put forward. An example from liver cirrhosis is used as illustration.

2 The Cox Model

We let T be the survival time measured from the chosen time origin. For the moment we will assume that T is a *proper* random variable, i.e., the event of interest is all-cause death or another event which will eventually happen to all subjects under study. In later sections, we will also study events which do not necessarily occur for all subjects, such as death from a specific cause. Let $S(t) = P(T > t)$ be the survival distribution function which we assume to be absolutely continuous with density f . By the assumption for T , we will have that $S(t) \rightarrow 0$ as $t \rightarrow \infty$. The *hazard function* for T is then

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

and has the interpretation as the instantaneous risk per time unit of failing just after time t given survival till t , i.e. $\lambda(t) \approx P(T \leq t + \Delta t \mid T > t) / \Delta t$ when $\Delta t > 0$

is small. It follows that $S(t)$ can be recovered from the hazard via the relation

$$S(t) = \exp\left(-\int_0^t \lambda(u) du\right),$$

so, the hazard provides a way of characterizing the distribution of T . This has for long been a well-established fact, but it was an important novel contribution of Cox¹³ to use the hazard function as the basis for modeling the distribution of T in relation to covariates $z = (z_1, \dots, z_p)$. This was done by letting

$$\lambda(t; z) = \lambda_0(t) \exp(\beta^T z), \tag{1}$$

where $\lambda_0(t) \geq 0$ is an unspecified function of time t . Note that, by modeling the hazard function that conditions on the past information at any time t , and by letting $\lambda_0(t)$ be completely unspecified, one avoids making strict assumptions about the right-hand tail of the distribution. Apart from modeling survival data via the hazard function, the Cox model was innovative by being *semi-parametric*. Thus, the combination of a linear predictor $\beta^T z$ including a finite number, p of regression coefficients β with an unspecified (‘non-parametric’) baseline hazard $\lambda_0(t)$ raised the question of how to do inference in the suggested model. The original argument¹³ was that if $\lambda_0(t)$ is arbitrary, then it could be zero between the observed event times $t_{(1)}, \dots, t_{(k)}$ (assumed distinct) which, therefore, must carry all information on β . This argument led to conditioning on the observed event times and multiply the conditional probabilities

$$\frac{\exp(\beta^T z_{(i)})}{\sum_{j \in R(t_{(i)})} \exp(\beta^T z_{(j)})}$$

that the ‘right’ subject fails at $t_{(i)}$ given a failure at that time. Here, $R(t)$, the *risk set* at time t , contains all subjects still alive and uncensored at that time and $z_{(j)}$ is the covariate vector for the subject with observation time $t_{(j)}$. The resulting estimating equations for β are then obtained in connection with maximizing

$$PL(\beta) = \prod_{i=1}^k \frac{\exp(\beta^T z_{(i)})}{\sum_{j \in R(t_{(i)})} \exp(\beta^T z_{(j)})}.$$

This unusual likelihood construction led to a detailed discussion in the literature. Thus, Kalbfleisch and Prentice²⁸ showed that, under some conditions, $PL(\beta)$ is the marginal likelihood of

the rank vector for the observed failure times while other approaches were discussed by, e.g. Bailey ⁶, Jacobsen ²⁶, Johansen²⁷, Kalbfleisch and Prentice ²⁹. To shed further light on the properties of *PL*, Cox ¹⁴ introduced the concept of a ‘partial likelihood’ and showed that *PL* is, indeed, an example of this. Similarly, estimation of the (cumulative) baseline hazard $\Lambda_0(t) = \int_0^t \lambda_0(u)du$ as well as asymptotic properties of the resulting estimators were not obvious and a further discussion took place in the literature. Thus, results on asymptotic normality of β were presented by, e.g. Andersen and Gill⁴ and Tsiatis⁴⁶. The latter approach using *counting processes* will be summarized in the next section. Finally, discussions of the *efficiency* of the model appeared ^{16,40}, raising the question about how much statistical precision was lost by making no assumptions concerning the shape of the survival time distribution (unspecified baseline hazard) compared to assuming, e.g. a Weibull shape.

3 The ‘Modern’ Approach to Inference in the Cox Model

The observed survival data may be represented as *counting processes*, as follows (e.g., Andersen et al.², Andersen and Gill ⁴). For independent subjects $i = 1, \dots, n$, we observe $(\tilde{T}_i, D_i, \mathbf{z}_i)$, where $\tilde{T}_i = T_i \wedge C_i$ is the minimum of the true survival time T_i and a time of *right-censoring*, C_i and $D_i = I(\tilde{T}_i = T_i)$ is the indicator of observing the true survival time T_i for that subject. The counting process observed for subject i is then

$$N_i(t) = I(\tilde{T}_i \leq t, D_i = 1)$$

and under *independent censoring* (e.g., Andersen et al.², Kalbfleisch and Prentice ²⁹), $N_i(t)$, by the Doob-Meyer decomposition (e.g., Andersen et al.²), can be written as

$$N_i(t) = \int_0^t Y_i(u)\lambda_i(u)du + M_i(t). \tag{2}$$

In (2), $\lambda_i(t)$ is the hazard function for the distribution of T_i , $Y_i(t) = I(\tilde{T}_i \geq t)$ is the *at-risk* indicator and $M_i(t)$ is a *martingale*. The integrand $Y_i(u)\lambda_i(u)$ is the *intensity process* for $N_i(t)$. The counting process approach to survival analysis has a number of advantages. First, via (2), a number of mathematical results on martingales may be applied when studying (large sample) properties of the inference methods. We will exemplify this shortly. Second, via Jacod’s formula (e.g., Andersen et al.²) it is possible to establish a likelihood on which inference may be based leading, in

fact, to the Cox partial likelihood discussed in the previous section. Third, this formulation directly extends to situations where modeling the hazard of an event which will not necessarily occur to all subjects with probability 1 is considered.

According to Jacod’s formula, the conditional likelihood given \mathbf{z} based on observing $((N_i(t), Y_i(t), \mathbf{z}_i), i = 1, \dots, n)$ for $0 < t \leq \tau$ is

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

where $dN_i(t) = N_i(t) - N_i(t-)$ is the jump in N_i at time t (0 or 1). (In fact, the full conditional likelihood given \mathbf{z} also contains factors arising from observing some censoring times, however, under the assumption of *non-informative censoring*, i.e. the censoring times carry no information on the hazard model parameters, L is proportional to the full likelihood; e.g. Andersen et al.², Kalbfleisch and Prentice²⁹)

If T_i follows the Cox model (1), then L is a function of the baseline hazard $\lambda_0(\cdot)$ and of the regression coefficients β . Differentiating, formally, $\log(L)$ with respect to a single $\lambda_0(t)$ and solving the resulting score equation for given β lead to

$$\widehat{\lambda_0(t)}dt = \frac{\sum_i dN_i(t)}{\sum_i Y_i(t) \exp(\beta^T \mathbf{z}_i)}, \tag{3}$$

and inserting (3) into L yields the *profile likelihood*

$$PL(\beta) \times \exp\left(-\int_0^\tau \sum_i dN_i(t)\right) \prod_t \left(\sum_i dN_i(t)\right)^{\sum_i dN_i(t)}$$

where the first factor

$$PL(\beta) = \prod_i \prod_t \left(\frac{Y_i(t) \exp(\beta^T \mathbf{z}_i)}{\sum_j Y_j(t) \exp(\beta^T \mathbf{z}_j)}\right)^{dN_i(t)} \tag{4}$$

is *Cox’s partial likelihood* and the second factor does not depend on the β -parameters. To estimate β , $PL(\beta)$ is maximized by computing the *Cox score*

$$\begin{aligned} \mathbf{u}_\tau(\beta) &= \frac{\partial}{\partial \beta} \log(PL(\beta)) \\ &= \sum_i \int_0^\tau Y_i(t) \left(\mathbf{z}_i - \frac{\sum_j Y_j(t)\mathbf{z}_j \exp(\beta^T \mathbf{z}_j)}{\sum_j Y_j(t) \exp(\beta^T \mathbf{z}_j)}\right) dN_i(t) \end{aligned} \tag{5}$$

and solving the resulting score equations. This leads to the Cox maximum partial likelihood estimator $\hat{\beta}$ and inserting this into (3) yields the Breslow estimator ⁽⁹⁾ of the cumulative baseline hazard $\Lambda_0(t)$:

$$\hat{\Lambda}_0(t) = \int_0^t \frac{\sum_i dN_i(u)}{\sum_i Y_i(u) \exp(\hat{\beta}^\top z_i)}. \tag{6}$$

It turns out that large-sample inference for $\hat{\beta}$ may be based on standard likelihood results for $PL(\beta)$. A crucial step is here to note that, evaluated at the true regression parameter, β_0 and considered as a process in t when based on the data from $[0, t]$, (5) is a martingale ^{2, 4}. Thus, by (2), $N_i(t) = \int_0^t Y_i(u) \lambda_0(u) \exp(\beta_0^\top z_i) du + M_i(t)$ and

$$\begin{aligned} U_t(\beta_0) &= \sum_i \int_0^t Y_i(u) (z_i \\ &\quad - \frac{\sum_j Y_j(u) z_j \exp(\beta_0^\top z_j)}{\sum_j Y_j(u) \exp(\beta_0^\top z_j)}) dN_i(u) \\ &= \sum_i \int_0^t Y_i(u) (z_i \\ &\quad - \frac{\sum_j Y_j(u) z_j \exp(\beta_0^\top z_j)}{\sum_j Y_j(u) \exp(\beta_0^\top z_j)}) dM_i(u). \end{aligned}$$

Thereby, martingale central limit theorems (e.g., Andersen et al.²) may be applied to the score and, via fairly standard Taylor expansion arguments, asymptotic normality for $\hat{\beta}$ may be obtained. Further, model-based standard deviations of $\hat{\beta}$ may be obtained from the second derivative of $\log(PL(\beta))$ and, thereby, the resulting Wald tests (as well as score- and likelihood ratio tests) are also valid. Martingale results may also be applied when studying large-sample properties of the Breslow estimator $\hat{\Lambda}_0(t)$ and the plug-in estimator

$$\hat{S}(t | z_0) = \exp\left(-\hat{\Lambda}_0(t) \exp(\hat{\beta}^\top z_0)\right) \tag{7}$$

of the estimated survival function for given covariates z_0 .

This derivation only builds on properties of counting processes showing that a Cox-type model may also be studied for the intensity process of a counting process that counts events such as cause-specific deaths, or recurrent events such as repeated hospitalisations (though the conditions needed for applying the martingale central limit theorem will vary among these different

situations). Thereby, Cox-type models have much wider applicability, going beyond the simple survival situation, as we will further discuss in Sect. 5.

4 An Example

Lombard et al.³⁷ reported on a randomized trial, PBC3, conducted at six European hospitals between 1983 and 1988. In brief, 349 patients with primary biliary cirrhosis (PBC) were randomized to the active drug CyA ($n = 176$) or to placebo ($n = 173$). Patients were followed until death from any cause or to liver transplantation, with 44 patients experiencing this composite end-point in the CyA group and 46 in the placebo group. Four patients dropped out of the trial before its planned termination on 1 January 1989 and the remaining (255) patients were alive without a liver transplantation on that date and these patients were all censored. The data are available from the web pages of Andersen and Skovgaard⁵.

Figure 1 shows the Kaplan and Meier³⁰ curves by treatment and it is seen that they are quite close, a fact that is sustained by fitting a Cox model including only the treatment indicator $z_1 = I(\text{CyA treatment})$. In this model, the estimated regression coefficient is $\hat{\beta}_1 = -0.059$ 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.

Even though PBC3 was a randomized study, it turned out that the CyA group had slightly less favorable average values of the important biochemical prognostic variables serum albumin (37.1 g/L for CyA, 39.3 g/L for placebo) and serum bilirubin (48.6 $\mu\text{mol/L}$ for CyA, 42.3 $\mu\text{mol/L}$ for placebo). Thus, the estimated treatment effect after adjustment for these two variables (bilirubin after a log-transformation) differs somewhat from the unadjusted coefficient cited above, see Table 1. The treatment groups now seem to differ significantly (95% confidence interval for the treatment hazard ratio ranges from 0.36 to 0.87). The log-transformation of serum bilirubin was determined after careful inspection of plots of cumulative martingale residuals (e.g., Lin et al.³⁶).

Graphical illustration of the results from the multiple Cox regression model can be done by estimating survival curves for treated and control patients with some given value of the biochemical adjustment variables, see Eq. (7). However, since that would provide one set of curves for each chosen covariate pattern, it would be of interest

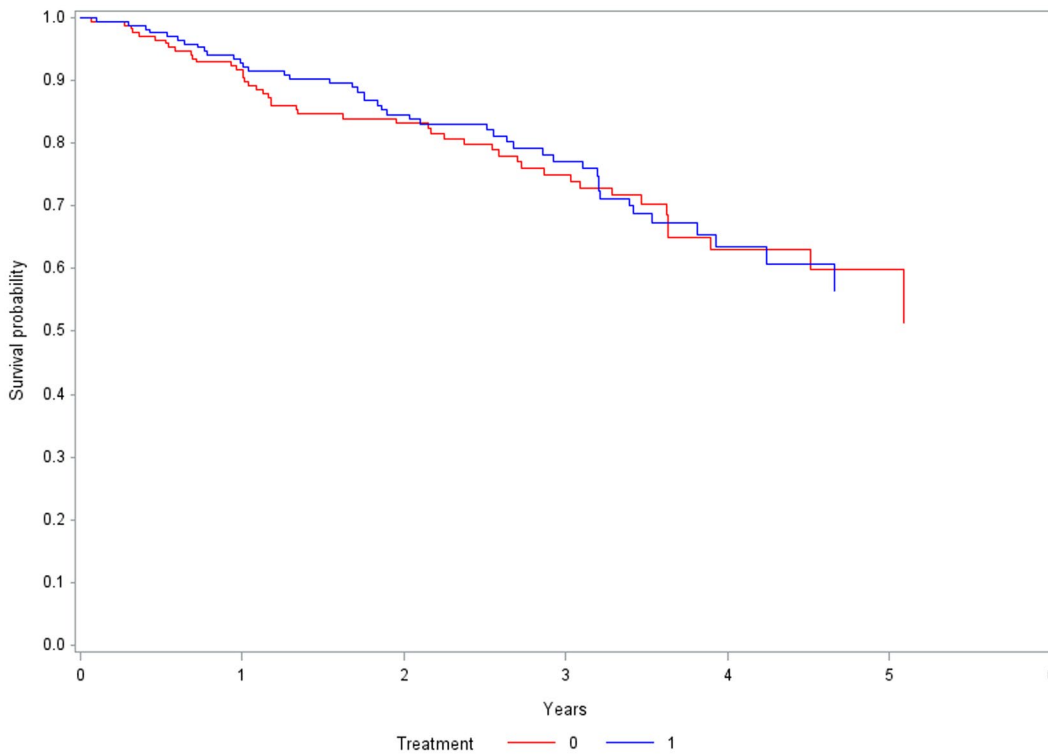


Figure 1: Kaplan–Meier estimates by treatment in the PBC3 trial.

Table 1: Estimated coefficients (and SD) from a Cox model for the PBC3 data with linear effects of albumin and $\log_2(\text{bilirubin})$.

Covariate		$\hat{\beta}$	SD
Treatment	CyA vs placebo	− 0.574	0.224
Albumin	per 1 g/L	− 0.091	0.022
$\log_2(\text{bilirubin})$	per doubling	0.665	0.074

to summarize the multiple regression model graphically by a single set of curves. This may be done using the g -formula (e.g., Hernan and Robins²⁵), as follows. Two predictions are performed for each subject, i , one setting treatment to CyA and one setting treatment to placebo and in both predictions keeping the observed values (z_{2i}, z_{3i}) for albumin and bilirubin. The predictions for each value of treatment are then averaged over $i = 1, \dots, n$:

$$\hat{S}_j(t) = \frac{1}{n} \sum_i \hat{S}(t \mid z_1 = j, z_{2i}, z_{3i}), \quad j = 0, 1. \tag{8}$$

The g -formula results in the curves shown in Fig. 2. Note that, if randomization in the PBC3

study had been more successful, then these curves would resemble the Kaplan–Meier estimates in Fig. 1. However, using the curves obtained based on the g -formula, it is possible to visualize the treatment effect on the probability scale after covariate-adjustment.

5 Influence of the Cox Paper

The short-term influence of the paper was touched upon in Sect. 2. Thus, as mentioned there, it triggered a discussion dealing with the kind of likelihood on which inference should be based, with how to non-parametrically estimate the baseline hazard, with large sample properties of the estimators, and with their asymptotic efficiency.

But also on a long-term basis, the paper has been very influential, both within the field of survival analysis and more generally in theoretical statistics. In survival analysis, the multiplicative semi-parametric structure (1) soon became the primary choice when setting up regression models for various parameters. In Sect. 2, we described the model as a model for the hazard function $\lambda(t) = f(t)/S(t)$ of a proper random variable, T with density f and survival distribution function

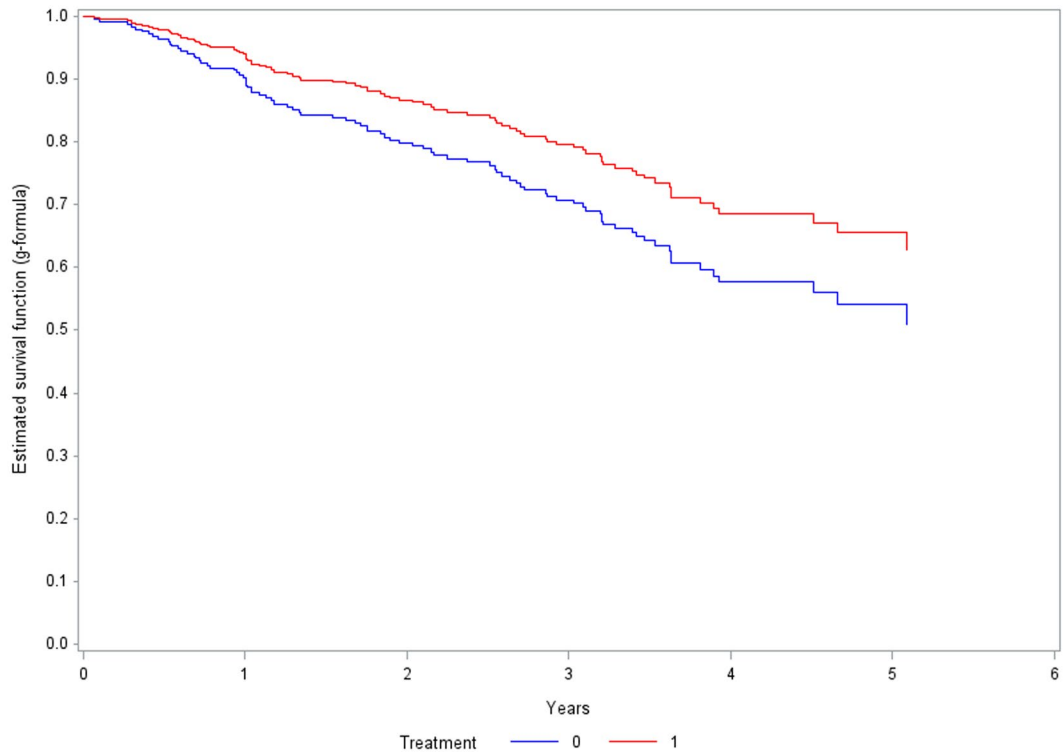


Figure 2: Estimated survival curve in the two treatment groups in the PBC3 trial based on the g -formula.

S. However, a similar model was proposed for the *cause-specific hazard* in *competing risks*⁴¹. Here, a number of distinct causes of death $D \in \{1, \dots, k\}$ are envisaged and the cause j -specific hazard

$$\lambda_j(t) = \lim_{\Delta t \rightarrow 0} \frac{P(T \leq t + \Delta t, D = j \mid T > t)}{\Delta t}$$

is assumed to have the form

$$\lambda_j(t; \mathbf{z}) = \lambda_{j0}(t) \exp(\boldsymbol{\beta}_j^T \mathbf{z}).$$

The competing risks model is a simple example of a multi-state model with one initial transient state ‘0: alive’ and k absorbing states $j = 1, \dots, k$ representing death from the different causes, and the cause-specific hazard, $\lambda_j(t)$ is then the $0 \rightarrow j$ transition hazard. Andersen and Borgan¹ studied general multi-state processes, say $X(t)$, with transition probabilities $P_{hj}(s, t) = P(X(t) = j \mid X(s) = h), s < t$ (where h, j are states) and modeled the *transition intensities*

$$\lambda_{hj}(t) = \lim_{\Delta t \rightarrow 0} \frac{P_{hj}(t, t + \Delta t)}{\Delta t}, \quad j \neq h$$

in the form (1). In both cases, inference and asymptotic properties follow the lines of Sect. 3.

In the competing risks model, the time of transition into state j is the improper random variable $T_j = \inf_{t>0}\{t : X(t) = j\}$ with hazard function

$$\lambda_j^*(t) = \lim_{\Delta t \rightarrow 0} \frac{P(T_j \leq t + \Delta t \mid T_j > t)}{\Delta t},$$

the cause j *sub-distribution hazard*. Fine and Gray¹⁸ proposed a model for the sub-distribution hazard in the form (1), i.e. with an unspecified baseline function and with covariates entering multiplicatively via a linear predictor. Similar models were studied by Lin³⁴ for expected *medical costs* in $[0, t]$ and by Lin et al.³⁵ and by Ghosh and Lin²⁰ for the expected number of events in $[0, t]$ in a recurrent events situation (without or with a terminal event). In all these cases, the parameter of interest was assumed to have the form

$$\mu(t; \mathbf{z}) = \mu_0(t) \exp(\boldsymbol{\beta}^T \mathbf{z})$$

(with the baseline function $\mu_0(t)$ unspecified except from being non-decreasing). In these cases, the Cox-type model does not specify the entire probability distribution of the data under study and, hence, no likelihood is available. Instead, inference is based on estimating

equations that are generalizations of the Cox score (5), now requiring inverse probability of censoring weighting to account for the incomplete observation. Derivation of asymptotic properties of the resulting estimators now needs the use of empirical process theory rather than martingale results.

These examples illustrate that the multiplicative ‘Cox structure’ is the default choice in survival analysis. Also in other special cases, such as random effects (‘frailty’) models (e.g., Clayton and Cuzick¹², Duchateau and Janssen¹⁵) and covariate measurement error models (e.g., Carroll et al.¹¹), Cox type models are typically studied even though the mathematical properties of the resulting models may not be quite as attractive as when looking, e.g. at additive hazards models.

In the more general field of theoretical statistics (outside survival analysis), the Cox paper has also had a remarkable influence. Thus, entirely new areas of research on partial likelihood (e.g., Gill²¹, Slud^{43,44}, Wong⁵⁰) and semi-parametric inference (e.g., Begun et al.⁷, Bickel et al.⁸, Tsiatis⁴⁷) have emerged.

Studying the many citations to Cox¹³, it appears that the vast majority comes from the medical world. This reflects that in clinical studies dealing with time-to-event problems, the Cox model (maybe in connection with the Kaplan–Meier estimator) has developed into *the* method of choice. This is likely related to the fact that in the 1970s, when the model was proposed, many clinical trials in cancer and other chronic diseases were launched worldwide where methods for survival data analysis were needed, and also that standard software packages soon made the Cox model generally available. And this is in spite of the fact that other models, such as accelerated failure time models, may provide parameters with a more intuitive interpretation than that of hazard ratios. We will return to this in the next section.

6 Some Criticism Raised Against the Cox Model

In this section, we will first focus on the two-sample situation corresponding to a single binary covariate $z \in \{0, 1\}$. This could be the situation in a randomized clinical trial with a time-to-event outcome and with $z = 1$ representing active treatment and $z = 0$ control as exemplified in Sect. 4. Analyzing the data using the Cox model provides a nice and apparently simple one-number

summary of the treatment effect, namely the hazard ratio $\exp(\beta)$. Nevertheless, several points of criticism against the model have been raised.

6.1 Hazard Ratio Interpretation-1

The hazard ratio does not have a simple clinical interpretation and it is some times confused with a ‘risk ratio’. However, in the two-sample model, the risk ratio at time t resulting from a Cox model in the absence of competing risks is

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

and only in a ‘low risk’ situation (i.e., $\exp(-\Lambda(t)) \approx 1 - \Lambda(t)$) is $RR \approx \exp(\beta)$. Thus, the ‘risk ratio’ interpretation does not hold in general. Also, in the presence of competing risks, the cause-1 risk ratio is the ratio between cumulative incidences $P(T \leq t, D = 1 | z)$ that also depend on the cause-specific hazards for the causes competing with cause 1. Thus, in this case, the risk ratio interpretation of $\exp(\beta)$ can be completely wrong, see e.g. Andersen et al.³ for an example of this.

6.2 Hazard Ratio Interpretation-2

The hazard ratio has also been criticized for not having a *causal* interpretation (e.g., Hernan²⁴, Martinussen et al.³⁹). To discuss this problem, a definition of causality is needed and we will define causality as in Martinussen et al.³⁹ following the tradition of ‘potential outcomes’ (e.g., Hernan and Robins²⁵). Thus, we let T_i^0 be the potential survival time that would be realized if subject i , possibly contrary to the fact, was given treatment 0 and define T_i^1 similarly. A causal contrast then compares the distributions of T^0 and T^1 , i.e. addressing what would happen if all subjects were treated with control compared to what would happen if all subjects were given the active treatment. Examples include $E(T^0) - E(T^1)$ or $P(T^0 \leq t)/P(T^1 \leq t)$. The hazard ratio at time t can, for a two-sample Cox model for the potential outcomes, be written as

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

and it is estimable if treatment allocation is randomized as in the PBC3 study, Sect. 4. It *does*, therefore possesses a causal interpretation, albeit not a very attractive one. On the other hand, the standard interpretation of the hazard ratio is

$$\exp(\beta) = \frac{\lim_{\Delta t \rightarrow 0} P(T^1 \leq t + \Delta t | T^1 > t) / \Delta t}{\lim_{\Delta t \rightarrow 0} P(T^0 \leq t + \Delta t | T^0 > t) / \Delta t}$$

and it is seen to contrast the two sub-populations $\{T^0 > t\}$ and $\{T^1 > t\}$ that are different when treatment does have an impact on survival. This shows that, for a time-varying hazard ratio, e.g. with $\lambda(t | z = 1) / \lambda(t | z = 0) < 1$ for $t < t_0$ and $\lambda(t | z = 1) / \lambda(t | z = 0) = 1$ for $t \geq t_0$, it would be incorrect to infer that ‘treatment works for $t < t_0$ ’ but ‘treatment is inefficient for $t \geq t_0$ ’.

Leaving the two-sample situation and looking, more generally, at a multiple Cox regression model including covariates $\mathbf{z} = (z_1, \dots, z_p)$, some further points of criticism have been raised.

6.3 Accelerated Failure Time Models

As mentioned in the introduction, a very natural approach to regression modeling in survival analysis would be to extend the classical linear regression model

$$\log(T) = \beta^T \mathbf{z} + \varepsilon$$

to allow for censored data. This model is known as the *accelerated failure time model* (AFT) as it provides parameters with an interpretation as acceleration factors. Thus, if subject 1 has $z = 1$ and subject 2 has $z = 0$ then $T_1 \approx T_2 \exp(\beta)$ and, thereby, the coefficient β has a nice and simple interpretation. This ‘physical’ interpretation was, in fact, mentioned already by Cox¹³ who, on the other hand, discussed the proportional hazards model as ‘convenient, flexible and yet entirely empirical’. Inference in the AFT model can be based on Jacod’s likelihood (Sect. 2) if a parametric specification of the distribution of the i.i.d. error terms $(\varepsilon_i, i = 1, \dots, n)$ is given. However, as mentioned in the Introduction, this would entail specifying the shape of the tail of the distribution for which little information is available based on censored data and, therefore, also semi-parametric inference in the AFT model, i.e. without specifying the distribution of the error terms, has been discussed (e.g., Buckley and James¹⁰).

6.4 Over-Simplicity

It has been argued (e.g., van Houwelingen and Putter⁴⁸) that proportional hazards is the exception rather than the rule and that more flexible hazard models should be applied (e.g., Martinussen et al.³⁸, Schemper⁴²). Also, Stensrud and Hernan⁴⁵ in a *JAMA* letter made the claim that ‘in practice, a constant hazard ratio does not occur for most medical applications’.

6.5 Non-collapsibility

Consider a situation with two covariates: a binary treatment indicator z_1 as above and another variable z_2 and consider the two Cox models $\lambda_0(t) \exp(\beta_1 z_1)$ and $\tilde{\lambda}_0(t) \exp(\tilde{\beta}_1 z_2 + \tilde{\beta}_2 z_1)$. Then, as noted, e.g. by Gail et al.¹⁹, if $\beta_2 \neq 0$, i.e. when z_2 is associated with survival, then $\beta_1 \neq \tilde{\beta}_1$ even in a randomized study where z_1 and z_2 are independent. This is known as *non-collapsibility* and has the consequence that when quoting a coefficient from a Cox model, it should always be emphasized which other covariates that the model included. In the PBC3 trial (Sect. 4), it was seen that the treatment effects with or without adjustment for two strong prognostic variables were very different (though, in that example, treatment was not quite independent of (z_2, z_3) because of the unsuccessful randomization).

6.6 Inferiority Compared to Machine Learning for Prediction Purposes

A number of studies (e.g., Kim et al.³¹, Leger et al.³³) have compared the performance of predictions based on multiple Cox regression models with that of different machine learning techniques, such as ‘deep learning’, and have found that the latter generally outperform the Cox model. In fact, machine learning methods for survival analysis, including methods for high-dimensional covariates, have been and continue to be a very active area of research. Important contributions to this field include Fang et al.¹⁷, Goeman²², Kvamme et al.³², Witten and Tibshirani⁴⁹, Hao et al.²³.

7 Discussion

As summarized above, the current situation is that the Cox¹³ proportional hazards regression model is an extremely popular and influential statistical model for survival data analysis and it provides a one-number summary of the way in which a covariate is associated with the survival time outcome—the *hazard ratio*. Nevertheless, as discussed in Sect. 6, the model has been criticized in a number of ways.

However, difficulties in interpreting the hazard ratio (be it clinically or causally) may be circumvented by reporting predicted *absolute risk* curves, e.g. in the form of survival curves as shown in the example of Sect. 4. Similarly, if a proportional hazards model is too simple to adequately fit a data set at hand then flexible extensions of the model exist (e.g., van Houwelingen and Putter⁴⁸s) from which, once more, absolute

risk curves are estimable. Also, the problem with non-collapsibility may be approached by, as an alternative to presenting hazard ratios, estimating absolute risk curves, e.g., using the *g*-formula (again exemplified in the PBC3 example). Finally, the fact that alternative methods, such as machine learning techniques, may provide superior predictions compared to a Cox model does not seem to have had the consequence that the Cox model is being abandoned. Rather, the model will likely be included in ensemble methods used for survival prediction together with more recently developed techniques in this area.

In conclusion, we believe that the Cox model will remain an important statistical tool in survival analysis, perhaps however, with a shift of focus from exclusive use of hazard ratios to more emphasis on using the model to provide absolute risk estimates, likely in combination with machine learning techniques.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Declarations

Conflict of interest

The author declares to have no conflicts of interest,

Received: 18 August 2021 Accepted: 24 December 2021
Published online: 23 January 2022

References

- Andersen PK, Borgan Ø (1985) Counting process models for life history data: a review (with discussion). *Scand J Statist* 12:97–158
- Andersen PK, Borgan Ø, Gill RD, Keiding N (1993) *Statistical models based on counting processes*. Springer, New York
- Andersen PK, Geskus RB, de Witte T, Putter H (2012) Competing risks in epidemiology: possibilities and pitfalls. *Int J Epidemiol* 41:861–70
- Andersen PK, Gill RD (1982) Cox's regression model for counting processes: a large sample study. *Ann Statist* 10:1100–1120
- Andersen PK, Skovgaard LT (2006) *Regression with linear predictors*. Springer-Verlag, New York
- Bailey KR (1983) The asymptotic joint distribution of regression and survival parameter estimates in the Cox regression model. *Ann Statist* 11:39–58
- Begun JM, Hall WJ, Huang W-M, Wellner JA (1983) Information and asymptotic efficiency in parametric-nonparametric models. *Ann Statist* 11:432–452
- Bickel PJ, Klaassen CA, Ritov Y, Wellner JA (1998) *Efficient and adaptive inference in semiparametric models*. Springer, New York
- Breslow NE (1974) Covariance analysis of censored survival data. *Biometrics* 30:89–99
- Buckley JD, James IR (1979) Linear regression with censored data. *Biometrika* 66:429–436
- Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu CM (2006) *Measurement error in nonlinear models*, 2nd edn. Chapman and Hall/CRC, Boca Raton
- Clayton DG, Cuzick J (1985) Multivariate generalizations of the proportional hazards model (with discussion). *J R Statist Soc A* 148:82–117
- Cox DR (1972) Regression models and life-tables (with discussion). *J Roy Statist Soc B* 34:187–220
- Cox DR (1975) Partial likelihood. *Biometrika* 62:269–276
- Duchateau L, Janssen P (2008) *The frailty model*. Springer, New York
- Efron B (1977) The efficiency of Cox's likelihood function for censored data. *J Amer Statist Assoc* 72:557–565
- Fang EX, Ning Y, Liu H (2017) Testing and confidence intervals for high dimensional proportional hazards models. *J R Statist Soc B* 79:1415–1437
- Fine JP, Gray RJ (1999) A proportional hazards model for the subdistribution of a competing risk. *J Amer Statist Assoc* 94:496–509
- Gail MH, Wieand S, Piantadosi S (1984) Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika* 71:431–444
- Ghosh D, Lin DY (2002) Marginal regression models for recurrent and terminal events. *Statistica Sinica* 12:663–688
- Gill RD (1992) Marginal partial likelihood. *Scand J Statist* 19:133–137
- Goeman JJ (2010) L1 penalized estimation in the Cox proportional hazards model. *Biom J* 52:70–84
- Hao L, Kim J, Kwon S, Ha ID (2021) Deep learning-based survival analysis for high-dimensional survival data. *Mathematics* 9:1244
- Hernan MA (2010) The hazards of hazard ratios. *Epidemiology* 21:13–15
- Hernan MA, Robins JM (2020) *Causal inference: what if*. Chapman and Hall/CRC, Boca Raton
- Jacobsen M (1984) Maximum likelihood estimation in the multiplicative intensity model: a survey. *Internat Statist Rev* 52:193–207
- Johansen S (1983) An extension of Cox's regression model. *Internat Statist Rev* 51:258–262
- Kalbfleisch JD, Prentice RL (1973) Marginal likelihoods based on Cox's regression and life model. *Biometrika* 60:267–278

29. Kalbfleisch JD, Prentice RL (1980) The statistical analysis of failure time data (2nd edn 2002). Wiley, New York
30. Kaplan EL, Meier P (1958) Non-parametric estimation from incomplete observations. *J Amer Statist Assoc* 53(457–481):562–563
31. Kim WJ, Sung JM, Sung D, Chae M, An SK, Namkoong K, Lee E, Chang H (2019) Cox proportional hazard regression versus a deep learning algorithm in the prediction of dementia: an analysis based on periodic health examination. *JMIR Med Inform* 7(3):e13139
32. Kvamme H, Borgan Ø, Scheel I (2019) Time-to-event prediction with neural networks and Cox regression. *J Mach Learn Res* 20:1–30
33. Leger S, Zwanenburg A, Pilz K, Lohau F, Linge A, Zöphel K, Kotzerke J, Schreiber A, Tinhofer I, Budach V, Sak A, S M, Balermphas P, Rödel C, Ganswindt U, Belka C, Pigorsch S, Combs SE, Mönnich D, Zips D, Krause M, Baumann M, Troost EGC, Löck S, Richter C (2017) A comparative study of machine learning methods for time-to-event survival data for radiomics risk modelling. *Nat Sci Rep* 7:13206
34. Lin DY (2000) Proportional means regression for censored medical costs. *Biometrics* 56:775–778
35. Lin DY, Wei LJ, Yang I, Ying Z (2000) Semiparametric regression for the mean and rate functions of recurrent events. *J R Statist Soc Ser B* 62:711–730
36. Lin DY, Wei LJ, Ying Z (1993) Checking the Cox model with cumulative sums of martingale-based residuals. *Biometrika* 80:557–572
37. Lombard M, Portmann B, Neuberger J, Williams R, Tygstrup N, Ranek L, Ring-Larsen H, Rodes J, Navasa M, Trepo C, Pape G, Schou G, Badsberg JH, Andersen PK (1993) Cyclosporin A treatment in primary biliary cirrhosis: results of a long-term placebo controlled trial. *Gastroenterol* 104:519–526
38. Martinussen T, Scheike TH, Skovgaard IM (2002) Efficient estimation of fixed and time-varying covariate effects in multiplicative intensity models. *Scand J Statist* 28:57–74
39. Martinussen T, Vansteelandt S, Andersen PK (2020) Subtleties in the interpretation of hazard contrasts. *Lifetime Data Anal* 26:833–855
40. Oakes D (1977) The asymptotic information in censored survival data. *Biometrika* 59:472–474
41. Prentice RL, Kalbfleisch JD, Peterson AV, Flournoy N, Farewell VT, Breslow N (1978) The analysis of failure time data in the presence of competing risks. *Biometrics* 34:541–554
42. Schemper M (1992) Cox analysis of survival data with non-proportional hazard functions. *J R Statist Soc Ser D (The Statistician)* 41:455–465
43. Slud EV (1986) Inefficiency of inferences with the partial likelihood. *Commun Statist Theory Methods* 15:3333–3351
44. Slud EV (1992) Partial likelihood for continuous-time stochastic processes. *Scand J Statist* 19:97–109
45. Stensrud MJ, Hernan MA (2020) Why test for proportional hazards? *JAMA* 323:1401–1402
46. Tsiatis AA (1981) A large sample study of Cox's regression model. *Ann Statist* 9:93–108
47. Tsiatis AA (2006) Semiparametric theory and missing data. Springer, New York
48. van Houwelingen HC, Putter H (2012) Dynamic prediction in clinical survival analysis. Chapman and Hall/CRC, Boca Raton
49. Witten DW, Tibshirani R (2010) Survival analysis with high-dimensional covariates. *Statist Methods Med Res* 19:29–51
50. Wong WH (1986) Theory of partial likelihood. *Ann Statist* 14:88–123



Per Kragh Andersen was born 1952 and holds candidate and PhD degrees in mathematical statistics from University of Copenhagen, Denmark, as well as a DrMedSci degree from same place. He has since 1978 been affiliated with Section of Biostatistics,

University of Copenhagen (former Statistical Research Unit) from 1998 as full professor of biostatistics. He is author or co-author of ~400 publications, including 4 books and 121 methodological articles. He was elected as Member of International Statistical Institute 1990 and received 'The Stata Journal Editor's price' in 2013.